Finite Element Method

Jonathan Ma
johnma@udel.edu

May 8, 2025

Contents

Preface

The goal of using Finite Elements (FE) discretization is to find or approximate solutions of PDEs. The FE method works well as it uses a variational formulation, which decreases the need for regularity for the solutions. We get more flexibility in approximating solutions of PDEs, and we can deal with complicated domains, singular solutions, and various types of boundary conditions.

The main ingredients of the FEM are approximation spaces, in most cases, we use piecewise polynomials with respect to a partition or mesh of the domain of the PDE. We need basic functional analysis and PDE theory for proving existence, uniqueness and approximation properties. Properties of linear algebra and calculus, and programming implementation skills are useful. Sometimes we need to use numerical experiments to confirm or discover theory or conjectures.

Model Problem

To understand the basic ideas of the FEM, we will use a model problem which will show the main ideas and aspects we will encounter along the way. Given the function $f : [0, 1] \to \mathbb{R}$, consider the problem of finding the real function u on the interval $[0, 1]$ such that

$$\begin{cases} -u''(x) = f(x), x \in (0, 1), \text{ and } u(0) = u(1) = 0. \end{cases} \tag{$*$}$$

This function corresponds to the displacement of an elastic cord u given a force $f = f(x)$. To find u, we will now use the FEM. Assume that $f \in L^2([0, 1]) \equiv \{f \mid \int_0^1 f^2\, dx < \infty\}$. We will study the variational (weak) formulation of this problem. If v is a smooth function such that $v(0) = v(1) = 0$, consider multiplying $-u''(x) = f(x)$ by v, and integrating:

$$\begin{aligned} -u''(x)v(x) &= f(x)v(x) \\ \int_0^1 -u''(x)v(x)\, dx &= \int_0^1 f(x)v(x)\, dx \\ \int_0^1 u'(x)v'(x)\, dx &= \int_0^1 f(x)v(x)\, dx. \end{aligned} \tag{$**$}$$

Before proceeding, we note the following.

1. Our problem formulation $(*)$ implies the weak formulation $(**)$.
2. In the weak formulation $(**)$, we are only dealing with the first derivatives in u, v, decreasing the amount of regularity on u at the expense of integration and variational formulation for any v.
3. If u is smooth, (u'' exists and is continuous), then $(**)$ implies $(*)$, which means that we can recover pointwise information about u'' from the weak formulation $(**)$.

Our new problem reads: find u such that $(**)$ is satisfied for all v that are smooth, with $v(1) = v(0) = 0$. For us to continue, and this reformulation to make sense, we need some more conditions on v. In other words, what is the right space for v to live in? We define $V = \{v \in C[0, 1] \mid v(0) = v(1) = 0, v' \in L^2([0, 1])\}$. In this space, we will work with a special type of derivative, called the weak derivative, which will be introduced later. For now, assume v' is the standard derivative, defined almost everywhere on $[0, 1]$.

Define a bilinear form

$$a(u,v) = \int_0^1 u'(x)v'(x)\,dx$$

on V. Now we claim V is a Hilbert space. The inner product on $L^2([0,1])$ is

$$(u,v) = \int_0^1 u(x)v(x)\,dx \quad \forall u, v \in L^2([0,1]),$$

and the corresponding norm is

$$\|u\| = \|u\|_{L^2([0,1])} = (u,u)^{1/2} = \left(\int_0^1 u(x)^2\,dx\right)^{1/2}, \forall u \in L^2(0,1).$$

With this new notation, $(**)$ becomes

$$\Big\{\text{Find } u \in V \text{ such that } a(u,v) = (f,v), \forall v \in V.$$

We will reiterate that this formulation does not include u'', even though the original problem has a second derivative.

Discrete Variational Formulation / Approximation

Let us approximate the solution of $(**)$ using the finite elemenet approach. The process is usually called finite element discretization. The discretization steps are as follows.

1. Divide $[0,1]$ into N equal parts, where $h = \frac{1}{N}$, and $x_k = hk, k = 0, 1, \ldots, N$.
2. Define $V_h \subset V$, where $V_h = \{v_h \in V \mid v_h \text{ is linear on each } [x_k, x_{k+1}]\}$. I.e. piecewise linear continuous functions. We note that v_h is uniquely determined by its values at $x_1, x_2, \ldots, x_{N-1}$ (only the interior nodes) since $v_h(x_0) = v_h(x_N) = 0$.
3. We introduce $\{\phi_j\}_{j=1}^{N-1} \in V_h$ such that $\phi_i(x_j) = \delta_{ij}$, where δ is the Kronecker delta function. We claim that $\{\phi_j\}_{j=1}^{N-1}$ is a basis for V_h. To this end, let $v_h \in V_h$ be fixed, and define $\alpha_j = v_h(x_j)$. Define $\tilde{v_h}(x) = \sum_{i=1}^{N-1} \alpha_i \phi_i(x)$. One can check that

   $$\tilde{v}_h(x_j) = \sum_{i=1}^{N-1} \alpha_i \phi_i(x_j) = \alpha_j = v_h(x_j).$$

   For linear independence, if $\sum_{i=1}^{N-1} \alpha_i \phi_i(x) = 0$, take $x = x_j$, $j = 1, \ldots, N-1$, which will imply that $\alpha_j = 0, \forall j = 1, \ldots, N-1$. Thus, the ϕ_i's are linearly independent, and we get that $\{\phi_j\}_{j=1}^{N-1}$ is a basis for V_h.
4. The vector $\alpha = (\alpha_1, \ldots, \alpha_{N-1})^T$ is a unique computational representation of the function $u_h = \sum_{i=1}^{N-1} \alpha_i \phi_i$.

With these considerations, we can now formulate the discrete variational formulation of $(**)$, which reads

$$\Big\{\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = (f, v_h), \forall v_h \in V_h.$$

To find u_h, we look for the representation $u_h = \sum_{i=1}^{N-1} \alpha_i \phi_i$ and test with nodal or basis functions. This leads to a linear system of equations. Using standard linear algebra results, one can prove that the system has unique solution $\alpha = (\alpha_1, \ldots, \alpha_{N-1})^T$. The two important questions we must then answer are...

1. How do we find the algebraic system for finding the vector $\alpha = (\alpha_1, \ldots, \alpha_{N-1})^T$ that represents the solution u_h?

2. How close is u_h from u? In what norm should we measure the error $u - u_h$?

To answer our first question, we consider $u_h = \sum_{i=1}^{N-1} \alpha_i \phi_i$, with each α_i to be determined. Take v_h to be $\phi_j, j = 1, 2, \ldots, N-1$. Then

$$\begin{aligned} a(u_h, v_h) &= (f, v_h) \\ a\left(\sum_{i=1}^{N-1} \alpha_i \phi_i, \phi_j\right) &= (f, \phi_j) \\ \implies \sum_{i=1}^{N-1} \alpha_i a(\phi_i, \phi_j) &= (f, \phi_j) \\ \implies A\alpha &= b, \end{aligned} \tag{***}$$

where

$$A = (A_{ij})_{i,j=1}^{N-1}, A_{ij} = \int_0^1 \phi_i' \phi_j', b = (b_j)_{j=1}^{N-1}, b_j = (f, \phi_j) = \int_0^1 f(x)\phi_j(x)\,dx.$$

We can now solve (***) for our answer. We can state couple of observations now about (***).

1. The matrix A is $(N-1) \times (N-1)$.

2. If $|i - j| > 1$, then $A_{ij} = \int_0^1 \phi_i' \phi_j' = 0$. This allows us to have a sparse matrix when we solve for this system (better computation time).

3. One can show that

$$A = \frac{1}{h} \begin{bmatrix} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & 0 & & \\ 0 & -1 & 2 & -1 & 0 & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix},$$

   and see that A is tridiagonal, symmetric, and invertible with a condition number $\kappa(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \approx \frac{c_0}{h^2}$. This condition number can cause problems if h gets very small, so we must be careful with preconditioning.

For our second question, we first define the energy norm

$$\|U\|_V = a(u,u)^{1/2} = \left(\int_0^1 (u')^2\,dx\right)^{1/2}.$$

We will now try to estimate $\|u - u_h\|_V$ by finding a suitable upper bound. While doing this, it is essential to note that $V_h \subset V$ (i.e. conforming). We will set $v = v_h$ to get

$$a(u, u_h) = (f, v_h), \quad \forall v_h \in V_h \subset V$$

and

$$a(u_h, v_j) = (f, v_h), \quad \forall v_h \in V_h \subset V.$$

This implies $a(u-u_h, v_j) = 0$, for all $v_h \in V_h$. Thus, the error is orthogonal to any vector in V_h. We call u_h the Ritz-Galerkin projection of u onto V_h. Then

$$\|u-u_h\|_V = \inf_{v_h \in V_h} \|u-u_h\|_V \leq \|u-v_h\|_V ,$$

for any $v_h \in V_h$. If we denote $I_h(u)$ as the linear interpolant of u then we know that $I_h(u) \in V_h$. Thus, we can actually find a bound of the error by finding the bound on the interpolant, which can be shown to be

$$\|u-u_h\|_V \leq \|u-I_h u\| \leq Ch \left\|u''\right\|_{L^2([0,1])} .$$

Therefore, if $\|u''\|_{L^2([0,1])} < \infty$, this inequality implies convergence of at least order h.

1 Sobolev Spaces

We begin our formal discussion with Sobolev spaces. Sobolev spaces admit weak differentiation, which means that functions which are not generally differentiable can still be analyzed.

1.1 Weak Derivatives

Definition 1.1. Let $\Omega \subset \mathbb{R}^d$, where $d = 1, 2, 3$, and the multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$, where $\alpha_i \in \mathbb{Z}_{\geq 0}$. Define $|\alpha| = \alpha_1 + \ldots + \alpha_d$. If $\phi : \Omega \to \mathbb{R}$ is a scalar function, we define

$$D^\alpha \phi := \frac{\partial^{|\alpha|}\phi}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} = \partial^{\alpha_1} \cdots \partial^{\alpha_d} \phi.$$

Example 1.2. Let $d = 2$, and let $\alpha = (1, 1)$. Then $|\alpha| = 2$, and

$$D^\alpha \phi = \frac{\partial^2 \phi}{\partial x_1 \partial x_2}.$$

If $\alpha = (2, 0)$,

$$D^\alpha \phi = \frac{\partial^2 \phi}{\partial x_1^2}.$$

Definition 1.3. For any continuous function φ on Ω, we define the support of φ by

$$\operatorname{supp}(\varphi) = \overline{\{x \in \Omega \mid \varphi(x) \neq 0\}}.$$

Since we've taken the closure in $\mathbb{R}^d$, $\operatorname{supp}(\varphi)$ might contain points which are not in Ω.

Definition 1.4. We can define some familiar spaces as follows:

1. $C^\infty(\Omega) := \{\phi : \Omega \to \mathbb{R} \mid D^\alpha \phi \text{ exists for any } \alpha\}$,
2. $\mathscr{D}(\Omega) = C_0^\infty := \{\phi \in C^\infty(\Omega) \mid \operatorname{support}(\phi) \subset \Omega\}$,
3. $L^1_{\text{loc}}(\Omega) := \left\{f : \Omega \to \mathbb{R} \text{ measurable} \mid \int_K |f| < \infty, \forall K \subset \Omega, K \text{ compact}\right\}$.

We now note that if Ω is bounded, then $L^2(\Omega) \subset L^1(\Omega) \subset L_{\text{loc}}(\Omega)$. In addition, for any smooth function $f \in C^\infty(\Omega)$, and $\alpha = (\alpha_1, \ldots, \alpha_d)$, using integration by parts, we have that $D^\alpha f$ satisfies

$$\int_\Omega (D^\alpha f)\phi = (-1)^{|\alpha|} \int_\Omega f(D^\alpha \phi), \quad \forall \phi \in \mathscr{D}(\Omega). \tag{1}$$

Definition 1.5. If $f \in L^1_{\text{loc}}(\Omega)$, we say that f has a weak derivative of order α (or α-weak derivative) if there exists a $g \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega g\phi = (-1)^{|\alpha|} \int_\Omega f D^\alpha \phi, \quad \forall \phi \in \mathscr{D}(\Omega). \tag{2}$$

In this case, we define $D^\alpha_w f := g$.

Theorem 1.6. *If f is smooth, then $D^\alpha_w f = D^\alpha f$.*

The α-weak derivative of f is unique as a function or class $g \in L^1_{\text{loc}}(\Omega)$, and might not be defined at every point of the domain of the function f. Furthermore, by comparing (1), (2), we note that viewing the α-weak derivative and the standard α-derivative as (integration type) linear functionals on $\mathscr{D}(\Omega)$, the action of the two functionals is the same. Thus, the α-weak derivative is designed to the same job as the standard derivative, when f is smooth enough, when integrating against functions in $\mathscr{D}(\Omega)$. The advantage here is that we can define weak derivatives for a much larger class of functions that do not have derivatives in the classical sense.

Example 1.7. Let $d = 1$, $\Omega = (0, 1)$. Define

$$f(x) = \begin{cases} x & x \in (0, 1/2] \\ 1 - x & x \in [1, 2, 1), \end{cases}$$

and

$$g(x) = \begin{cases} 1 & x \in (0, 1/2) \\ -1 & x \in (1/2, 1). \end{cases}$$

Claim. The weak derivative $D^1_w f = g(x)$, and f does not have a weak derivative of second order (D^2_w does not exist).

Proof. We need to find $g \in L^1_{\text{loc}}(\Omega)(0, 1)$ such that

$$\int_0^1 g\phi \, dx = (-1)^1 \int_0^1 f(x)\phi'(x) \, dx, \forall \phi \in \mathscr{D}((0, 1)).$$

We note that since $\phi(0) = \phi(1) = 0$, integrating by parts gives us

$$\begin{aligned} -\int_0^1 f(x)\phi(x) \, dx &= -\left(\int_0^{1/2} f(x)\phi'(x) \, dx + \int_{1/2}^1 f(x)\phi'(x) \, dx \right) \\ &= -\left(f\phi \Big|_0^{1/2} - \int_0^{1/2} f'(x)\phi(x) \, dx + f\phi \Big|_{1/2}^1 - \int_{1/2}^1 f'(x)\phi(x) \, dx \right) \\ &= \int_0^{1/2} \phi(x) \, dx - \int_{1/2}^1 \phi(x) \, dx \\ &= \int_0^1 g(x)\phi(x) \, dx, \end{aligned}$$

where $g(x) = \begin{cases} 1, & x \in (0, 1/2) \\ -1, & x \in (1/2, 1). \end{cases}$. Since $g \in L^1_{\text{loc}}(0, 1)$, we have that g is our first order weak derivative of f, or $D^1_w f = g$. Note here that we do not have to specify a value for g at $x = 1/2$,

since $\{1/2\}$ is a set of measure zero, and whatever value we assign to g is going to lead to the same class of representation of g.

For $\alpha = 2$, we now search for a function $h = D_w^2 f$, where $h \in L^1_{\text{loc}}(0,1)$, such that

$$\int_0^1 h\phi = (-1)^2 \int_0^1 f\phi''\,dx = \int_0^1 f\phi''(x)\,dx, \forall \phi \in \mathscr{D}((0,1)).$$

First, we note that

$$\begin{aligned}
\int_0^1 f\phi''(x)\,dx &= \int_0^{1/2} f\phi''(x)\,dx + \int_{1/2}^1 f\phi''(x)\,dx \\
&= -\int_0^{1/2} f'\phi'\,dx + f\phi'\Big|_0^{1/2} - \int_{1/2}^1 f'\phi'\,dx + f\phi'\Big|_{1/2}^1 \\
&= -\left(\int_0^{1/2} f'\phi'\,dx + \int_{1/2}^1 f'\phi'\,dx\right) \\
&= \int_{1/2}^1 \phi'\,dx - \int_0^{1/2} \phi'\,dx \\
&= \phi(1) - \phi(1/2) - \phi(1/2) + \phi(0) \\
&= -2\phi(1/2).
\end{aligned}$$

Is it possible then to find $h \in L^1_{\text{loc}}(0,1)$ such that

$$\int_0^1 h\phi\,dx = -2\phi(1/2), \quad \forall \phi \in C_0^\infty((0,1))? \tag{3}$$

The answer is no, and we will prove this by contradiction. Assume that (3) holds true. In particular, choose in (3), the functions $\phi_n \in C_0^\infty((0,1))$ such that $|\phi_n(x)| \le 1$, $\phi_n(1/2) = 1$ and $\phi_n(x) = 0$, for $|x - 1/2| \ge 1/n$, for $n \ge 3$. These functions can be thought of as tighter and tighter bell curves that always have their apex at $x = 1/2$. This implies that

$$|h\phi_n| \le |h|\,\chi_{[\frac{1}{2}-\frac{1}{3},\frac{1}{2}+\frac{1}{3}]}, \text{ for } n \ge 3,$$

where χ_A is the characteristic function of the set A. Since $h \in L^1_{\text{loc}}((0,1))$ and $h\phi_n \to 0$ almost everywhere, by the Lebesgue dominated convergence theorem, we have that

$$\int_0^1 h\phi_n \to \int_0^1 0 = 0.$$

But we would also have $\int_0^1 h\phi_n = -2$, a contradiction. Thus, f does not admit a second weak derivative. □

The following results can be derived using the standard multivariable integration by parts formula.

Theorem 1.8. *If $f \in C^{|\alpha|}(\Omega)$, then f admits α-weak derivatives, and $D_w^\alpha f = D^\alpha f$.*

The next result shows that if a function is globally continuous and C^1 smooth on subdomains, then the function admits first order weak derivatives.

Theorem 1.9. *If $\Omega \subset \mathbb{R}^d$ and $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}$, with $\Omega_1 \cap \Omega_2 = \emptyset$ and $u \in C(\overline{\Omega})$, with $u|_{\Omega_j} \in C^1(\overline{\Omega_j})$, for $j = 1, 2$, then u has weak derivatives of first order on Ω, and for $i = 1, 2$,*

$$\frac{\partial_w u}{\partial x_i}\Big|_{\Omega_j} = \frac{\partial u}{\partial x_i}, \text{ on } \Omega_j, j = 1, 2. \tag{4}$$

This result extends easily to any finite number of subdomains.

Proof. Let $\Gamma := \overline{\Omega_1} \cap \overline{\Omega_2}$, and let $\mathbf{n} = (n_1, n_2)$ be the outer (to Ω_1) normal unit vector on Γ. To prove (4) for $i = 1$, it would be enough to show that

$$\int_{\Omega_1} \frac{\partial u}{\partial x_1}\phi + \int_{\Omega_2} \frac{\partial u}{\partial x_1}\phi = -\int_{\Omega} u\frac{\partial \phi}{\partial x_1}, \forall \phi \in \mathscr{D}(\Omega). \tag{5}$$

Using that $\text{supp}(\phi)$ is compact and contained in Ω, we have that $\phi = 0$ on $\partial\Omega_j \setminus \Gamma$. Thus, using the integration by parts formula on each Ω_j we have

$$\begin{aligned}\int_{\Omega} u\frac{\partial \phi}{\partial x_1} &= \int_{\Omega_1} u\frac{\partial \phi}{\partial x_1} + \int_{\Omega_2} u\frac{\partial \phi}{\partial x_1} \\ &= -\left(\int_{\Omega_1} \frac{\partial u}{\partial x_1}\phi + \int_{\Omega_2} \frac{\partial u}{\partial x_1}\phi\right) + \int_{\Gamma} u\phi n_1 + \int_{\Gamma} u\phi(-n_1),\end{aligned}$$

which gives (5). □

The next result extends a known result about characterization of constant functions using the classic gradient to the case when we use the weak gradient. For more results on weak derivative calculus, we refer to Bartels' book, *Numerical Approximation of Partial Differential Equations*, page 85.

Theorem 1.10. *If $\Omega \subset \mathbb{R}^d$ is connected and $\nabla_w u = 0$, then u is constant on Ω.*

1.2 Lp Spaces

Definition 1.11. For $1 \leq p \leq \infty$, and $\Omega \subset \mathbb{R}^d$ is open, the L^p space is defined as

$$L^p(\Omega) = \left\{ f \in L^1_{\text{loc}}(\Omega) \mid \int_{\Omega} |f|^p < \infty \right\},$$

defined with norm

$$\|f\|_{L^p(\Omega)} := \|f\|_{0,p,\Omega} = \left(\int_{\Omega} |f|^p\right)^{\frac{1}{p}}, 1 \leq p < \infty.$$

If $p = \infty$, then

$$L^{\infty}(\Omega) := \{f : \Omega \to \mathbb{R} \text{ measurable } \mid \|f\|_{L^{\infty}(\Omega)} < \infty\},$$

where

$$\|f\|_{L^{\infty}(\Omega)} \text{ess sup}_{x\in\Omega} |f(x)| = \inf_{\mu(S)=0} \sup_{x\in\Omega\setminus S} |f(x)|.$$

Theorem 1.12. *$L^p(\Omega)$ is a Banach space. When $p = 2$, $L^2(\Omega)$ is a Hilbert space, with*

$$\langle u, v\rangle_{L^2(\Omega)} = \int_{\Omega} uv\, dx.$$

Theorem 1.13 (Hölder's inequality)**.** *Suppose* $1 \le p, q \le \infty$*, and* $\frac{1}{p} + \frac{1}{q} = 1$*, then if* $f \in L^p(\Omega)$*, and* $g \in L^q(\Omega)$*, then* $fg \in L^1(\Omega)$*, and*

$$\|fg\|_{L^1(\Omega} \le \|f\|_{L^p(\Omega)} \|g\|_{L^p(\Omega)}.$$

The following will be on our homework.

Theorem 1.14. *If* $1 \le p_1 < p_2 < \infty$*, and* Ω *bounded, then*

$$L^{p_2}(\Omega) \subset L^{p_1}(\Omega).$$

Definition 1.15 (Sobolev spaces)**.** For a bounded open domain Ω, define

$$W_p^m(\Omega) := \{u \in L^p(\Omega) \mid D_w^\alpha u \in L^p(\Omega), \forall |\alpha| \le m\},$$

with norm

$$\|u\|_{W_p^m(\Omega)}^p = \|u\|_{m,p,\Omega}^p = \sum_{|\alpha| \le m} \|D_w^\alpha u\|_{L^p(\Omega)}^p.$$

The seminorm on $W_k^m(\Omega)$ is

$$|u|_{k,p,\Omega}^p = \sum_{|\alpha| = m} \|D_w^\alpha u\|_{L^p(\Omega)}^p.$$

When $p = \infty$, the Sobolev space $W_\infty^m(\Omega)$ has norm

$$\|u\|_{W_\infty^m(\Omega)} = \max_{|\alpha| \le k} \sup_\Omega |\partial^\alpha f|.$$

We write $H^m(\Omega) := W_2^m(\Omega)$.

Here are some examples.

Example 1.16. For $d = 2$ (as in $\mathbb{R}^d$), $p = 2$, $m = 1$,

$$\begin{aligned} H^1(\Omega) = W_1^2(\Omega) &= \{u \in L^2(\Omega) \mid D_w^{(1,0)} u, D_w^{(0,1)} u \in L^2(\Omega)\} \\ &= \left\{ u \in L^2(\Omega) \,\middle|\, \left(\frac{\partial_w u}{\partial x_1}\right), \left(\frac{\partial_w u}{\partial x_2}\right) \in L^2(\Omega) \right\} \\ &= \left\{ u \in L^2(\Omega) \middle| \int_\Omega u^2 < \infty \text{ and } \int_\Omega |\nabla_w u|^2 < \infty \right\}, \end{aligned}$$

where $|\nabla_w u|^2 = \left(\frac{\partial_w u}{\partial x_1}\right)^2 + \left(\frac{\partial_w u}{\partial x_2}\right)^2$.

Example 1.17. Let $d = 2, p = 2$. Then we have the following:

$$\begin{aligned} |u|_{H^2(\Omega)}^2 = |u|_{W_2^2(\Omega)}^2 &= \int_\Omega \left|\frac{\partial^2 u}{\partial x^2}\right|^2 + \int_\Omega \left|\frac{\partial^2 u}{\partial^2 y}\right|^2 + \left|\frac{\partial^2 u}{\partial x \partial y}\right|^2, \\ |u|_{H^1(\Omega)}^2 &= \int_\Omega \left|\frac{\partial^2 u}{\partial x^2}\right|^2 + \int_\Omega \left|\frac{\partial^2 u}{\partial y^2}\right|^2, \\ \|u\|_{H^1(\Omega)} &= \left(\int_\Omega |u|^2 + \int_\Omega \left|\frac{\partial u}{\partial x}\right|^2 + \left|\frac{\partial u}{\partial y}\right|^2 \right)^{1/2}, \\ \|u\|_{H^1(\Omega)} &= \left(\int_\Omega |\nabla u|^2 \, dx \right)^{1/2}. \end{aligned}$$

Theorem 1.18. *When $p = 2$, we have $H^m(\Omega) = W_2^m(\Omega)$ is a Hilbert space, with norm induced by the inner product*

$$(u, v)_{H^m(\Omega)} = \sum_{|\alpha| \leq m} (D_w^\alpha u, D_w^\alpha v).$$

Theorem 1.19. *The space $W_p^m(\Omega)$ is Banach.*

Theorem 1.20. *The space $C^\infty(\Omega) \cap W_p^m(\Omega)$ is dense in $W_p^m(\Omega)$, which means that one can approximate any function in $W_p^m(\Omega)$ with smooth functions in $C^\infty(\Omega)$, where we note that $C^\infty(\overline{\Omega}) \subset C^\infty(\Omega)$.*

Definition 1.21. A domain $\Omega \subset \mathbb{R}^d$ is Lipschitz, or has Lipschitz boundary if the domain's boundary is sufficiently regular in the sense that locally it is the graph of a Lipschitz continuous function.

Theorem 1.22. *If Ω is a Lipschitz domain, then $C^\infty(\overline{\Omega})$ is dense in $W_p^m(\Omega)$.*

1.3 Sobolev Embeddings and Inequalities

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, and k a positive integer, and $1 \leq p < \infty$.

Theorem 1.23. *If either one of these two cases hold:*

1. *$k \geq d$, and $p = 1$, or*
2. *$k > \frac{d}{p}$, and $p > 1$,*

then $W_p^k(\Omega) \subset L^\infty(\Omega)$.

This is equivalent to saying

$$\|u\|_{L^\infty(\Omega)} \leq C \|u\|_{W_p^k(\Omega)}, \quad \forall u \in W_p^k(\Omega).$$

Moreover, for any $u \in W_p^k(\Omega)$, there exists a function $v \in C(\Omega)$ such that $u = v$ a.e. So under the assumptions of Theorem 1.23, $W_p^k(\Omega)$ can be viewed as a subspace of $C(\Omega)$, and functions in $W_p^k(\Omega)$ are bounded and have point values.

Example 1.24. For $d = 1$, and $p = 2$, we would need $k > \frac{1}{2}$ to have $H^k(\Omega) \subset C(\Omega)$, so $H^1(\Omega) \subset C(\Omega)$.

Example 1.25. For $d = 2$, $p = 2$, we have that $k > 1$. This means that $H^1(\Omega) \not\subset L^\infty(\Omega)$.

Example 1.26. For $d = 3$, and $p = 2$, we need $k > \frac{3}{2}$, which implies that $H^2(\Omega) \subset C(\Omega) \subset L^\infty(\Omega)$.

The following is on homework 1.

Example 1.27. Let $\Omega = \{x \in \mathbb{R}^2 \mid |x| < \frac{1}{2}\} = B(0, 1/2)$. Define $u(r, \theta) = \log \log \frac{1}{r}$. Then we can show that $u \in H^1(\Omega)$, but $u \notin C^0(\Omega)$. The latter statement is clear because as $r \to 0$, then $u \to \infty$. Fir the first statement, we would need to prove that $\int_\Omega u^2 < \infty$, and $\int_\Omega u_x^2 + u_y^2 < \infty$. If we recall that $u_x^2 + u_y^2 = u_r^2$, we can then show that $\int_\Omega u_r^2 < \infty$, and thus, $u \in H^1(\Omega)$.

1.4 Trace Spaces

Assume that $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, and assume that $C(\overline{\Omega})$ is dense in $H^1(\Omega)$. For $u \in C(\overline{\Omega})$, define the trace $\gamma u := u|_{\partial\Omega} = u(x)$ for all $x \in \partial\Omega$. If $\partial\Omega$ is not defined for u on some function, then we can use the trace to define the function on the boundary in the following way. For any nice domain, we can show that

$$\int_{\partial\Omega} u^2 < C \int_\Omega u^2 \left(\int_\Omega u^2 + \int_\Omega |\nabla u|^2 \right), \quad \forall u \in C(\overline{\Omega}), \tag{*}$$

which is a way to bound the integral around the boundary by integrals that are inside the domain. Since $C(\overline{\Omega})$ is dense in $H^1(\Omega)$, then (*) implies that $u \in L^2(\partial\Omega)$ and

$$\|u\|_{L^2(\partial\Omega)} \leq \sqrt{C}\, \|u\|_{L^2(\Omega)}\, \|u\|_{H^1(\Omega)}\,. \tag{**}$$

Now for any $u \in H^1(\Omega)$ take $(u_j) \subset C^1(\overline{\Omega})$ such that $u_j \to u$ in $H^1(\Omega)$. Define $u|_{\partial\Omega} = \lim_{j\to\infty} u_j|_{\partial\Omega}$. Then we can prove that u is well defined and $u \in L^2(\partial\Omega)$ satisfies (**). Thus we can define u on the boundary even if the function initially is not well defined there. A more general version of the trace inequality is as follows

Theorem 1.28. *Assume $\Omega \subset \mathbb{R}^d$ is Lipschitz and $1 \leq p < \infty$. Then for any $v \in W_p^1(\Omega)$, $\gamma v = v|_{\partial\Omega}$ is well defined, $\gamma v \in L^p(\partial\Omega)$, and there exists $C > 0$ such that*

$$\|v\|^p_{L^p(\partial\Omega)} \leq C\, \|v\|^{1-1/p}_{L^p(\Omega)}\, \|v\|^{1/p}_{W_p^1(\Omega)}\,,$$

for all $v \in W_p^1(\Omega)$.

Note that for $p = 2$, we obtain (**).

Definition 1.29. Define $H_0^m(\Omega)$ to be the completion of $C_0^\infty(\Omega)$ in $(H^m(\Omega), \|\cdot\|_{H^m(\Omega)})$. One can show

$$H_0^1(\Omega) := \{u \in H^1(\Omega) \mid \gamma u = 0 \text{ in } L^2(\Omega)\},$$

and

$$H_0^2(\Omega) := \{u \in H^1(\Omega) \mid \gamma u = 0 \text{ and } \nabla u \cdot n|_{\partial\Omega} = 0 \text{ in } L^2(\Omega)\}.$$

We note that $H_0^1(\Omega) \subset H^1(\Omega)$ (it is a closed subspace) and that $H_0^1(\Omega)$ takes the norm and inner product from $H^1(\Omega)$.

Theorem 1.30. *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded and is a nice domain. Then there exists a constant $C = C(\Omega)$ such that*

$$\int_\Omega u^2\, dx < C^2 \int_\Omega |\nabla u|^2\, dx, \text{ or } \|u\|_{L^2(\Omega)} < C\, |u|_{H^1(\Omega)}\,.$$

Proof. We first assume that $u \in C_0^1(D)$. Since D is bounded, it can be enclosed in a square $\Gamma := \{|x_i| \leq a, i = 1, 2\}$. We continue by assuming u to be identically 0 outside of D. Then for any $(x_1, x_2) \in \Gamma$, by Cauchy-Schwarz, we get

$$|u(x)|^2 = \left| \int_{-a}^{x_1} u_{x_1}(\xi, x_2)\, d\xi_1 \right|^2 \leq (x_1 + a) \int_{-a}^{x_1} |u_{x_1}|^2\, d\xi \leq 2a \int_{-a}^{x_1} |u_{x_1}|^2\, d\xi.$$

Therefore,

$$\int_{-a}^{a} |u(x)|^2 \, dx_1 \leq 4a^2 \int_{-a}^{a} |u_{x_1}|^2 \, d\xi.$$

Now integrating with respect to x_2 from $-a$ to a, we get

$$\iint_D |u(x)|^2 \, dx \leq 4a^2 \iint_D |u_{x_1}|^2 \, dx \leq 4a^2 \iint_D |\nabla u|^2 \, dx.$$

The theorem now follows from the fact that $C_0^1(D)$ is dense in $H_0^1(D)$. □

We note some consequences of the previous theorem.

1. Note that $\|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2$. Then $\|\cdot\|_{H^1(\Omega)}$ and $|\cdot|_{H^1(\Omega)}$ are equivalent $H_0^1(\Omega)$. To see this, we have for all $u \in H_0^1(\Omega)$,

$$|u|_{H^1(\Omega)}^2 \leq \|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + \|\nabla\|_{L^2(\Omega)}^2 \overset{P.I.}{\leq} (1+C) \|u\|_{L^2(\Omega)}^2 \leq (1+C) |\nabla u|_{H^1(\Omega)}^2.$$

2. Any function $f \in L^2(\Omega)$ can be viewed as a continuous bounded functional on $H_0^1(\Omega)$. More precisely, let $u \in L^2(\Omega)$. Then define $G_u : H_0^1(\Omega) \to \mathbb{R}$ by

$$G_u(v) = \int_\Omega u(x)v(x) \, dx, \quad \forall v \in H_0^1(\Omega).$$

To show that this is continuous, we show that it is bounded (this is clearly linear). To this end, for all $v \in H_0^1(\Omega)$,

$$\begin{aligned} |G_u(v)| &= \left| \int_\Omega u(x)v(x) \, dx \right| \\ &\leq \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq C(\Omega) \|u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq C(\Omega) \|v\|_{H^1(\Omega)}. \end{aligned}$$

Therefore, $G_u \in (H_0^1(\Omega))^*$ be definition in $H^1(\Omega)$. Therefore,

$$H_0^1(\Omega) \subset L^2(\Omega) \subset H^1(\Omega).$$

1.5 Calculus in Sobolev Spaces

Assume that $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. We define the unit normal vector n on $\partial\Omega$ (assume it is well defined almost everywhere). If u is smooth, then $\frac{\partial u}{\partial n} = \nabla u \cdot n = \sum \frac{\partial u}{\partial x_i} n_i$, where $n = (n_1, \ldots, n_d)$, and $\|n\| = 1$. If $u, v \in C^1(\overline{\Omega})$, then we have the integration by parts formula:

$$\int_\Omega u \frac{\partial v}{\partial x_i} \, dx = -\int_\Omega \frac{\partial u}{\partial x_i} v \, dx + \int_\Gamma (uv) n_i \, ds,$$

where $\Gamma = \partial\Omega$. The formula can then extend to functions in $H^1(\Omega)$. Even though there are problems at the boundaries here, we can use the trace to define u, v there so that the last part of the equation above becomes

$$\int_\Gamma (uv) n_i \, dx \Longrightarrow \int_\Gamma \gamma(uv) n_i,$$

where $\gamma(uv)$ is the trace of uv on Γ.

Next, if $v = (v_1, \ldots, v_d)$, where $v_i \in C^1(\Omega)$, then we define the divergence of v to be

$$\operatorname{div} v = \frac{\partial v_1}{\partial x_1} + \ldots + \frac{\partial v_d}{\partial x_d}.$$

If $v_i \in H^1(\Omega)$, then we think about $\frac{\partial v_i}{\partial x_i}$ in the weak sense, i.e. $\frac{\partial_w v_i}{x_i}$. If $\phi \in C^1(\Omega)$, then we define the gradient as

$$\nabla \phi = \left(\frac{\partial \phi}{\partial x_1}, \ldots, \frac{\partial \phi}{\partial x_d} \right).$$

We now introduce the Green's formulas. Assume that $v \in (C^1(\overline{\Omega}))^d$, $\phi \in C^1(\overline{\Omega})$. Then as a consequence of the integration by parts formula, we have the 1st Green's formula:

$$\int_\Omega v \cdot \nabla \phi \, dx + \int_\Omega \operatorname{div} v \phi \, dx = \int_{\partial\Omega} (\phi v) \cdot n \, ds.$$

We note here that when $\phi = 1$, we obtain the Divergence theorem,

$$\int_\Omega \operatorname{div} v \, dx = \int_{\partial\Omega} v \cdot n \, ds.$$

This holds for all $v \in (H^1(\Omega))^d$ and $\phi \in H^1(\Omega)$.

If $u, v \in C^2(\overline{\Omega})$, then we obtain Green's 2nd formula, by replacing $v = \nabla u$, and $\phi = v$:

$$\int_\Omega \nabla u \cdot \nabla u \, dx + \int_\Omega \Delta u \cdot v \, dx = \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, ds.$$

We can rewrite Green's 2nd formula using L^2 inner product notation:

$$(\nabla u, \nabla v) + (\Delta u, v) = \left\langle \frac{\partial u}{\partial n}, v \right\rangle_{\partial\Omega},$$

where $\langle \cdot, \cdot \rangle$ notation is in general used to denote the integration on the boundary of the domain when the functions to be integrated are smooth enough. Green's 2nd formula can be extended to functions $u, v \in H^1(\Omega)$ such that $\Delta u \in L^2(\Omega)$. In order to make sense of the RHS of Green's 2nd formula, we first note that for any $v \in H^1(\Omega)$, we have $\gamma(v) = v|_{\partial\Omega}$ belongs to a closed subspace of $L^2(\partial\Omega)$, which is defined by

$$H^{1/2}(\partial\Omega) := \{\gamma(v) = v|_{\partial\Omega} \mid v \in H^1(\Omega) \subset L^2(\partial\Omega)\},$$

with the norm

$$\|v\|_{H^{1/2}(\partial\Omega)} := \inf\{\|w\|_{H^1(\Omega)} \mid w \in H^1(\Omega), \gamma(w) = v\}.$$

Next, for a fixed $u \in H^1(\Omega)$ with $\Delta u \in L^2(\Omega)$, we can view $\frac{\partial u}{\partial n}$ as a bounded functional on $H^{1/2}(\Omega)$. Indeed, for any $v \in H^{1/2}(\Gamma)$, let $w \in H^1(\Omega)$ be such that $\gamma(w) = v$. Define

$$G_u(v) := (\nabla u, \nabla w) + (\Delta u, w), \quad \forall v \in H^{1/2}(\Gamma),$$

where $\Gamma := \partial\Omega$. One can show that $G_u(\cdot)$ is well defined, linear, and

$$|G_u(v)| \leq c(u) \|v\|_{H^{1/2}(\Omega)}, \quad \forall v \in H^{1/2}(\Gamma),$$

where $c(u) = (\|\nabla u\|^2 + \|\Delta u\|^2)^{1/2}$. Thus, $G_u(v)$, which agrees with $\int_\Gamma \frac{\partial u}{\partial n} v \, ds$ for u, v smooth, belongs to $(H^{1/2}(\Gamma))^*$, which by definition or tradition is denoted by $H^{-1/2}(\Gamma)$. Consequently, in this case, $\left\langle \frac{\partial u}{\partial n}, v \right\rangle_{\partial\Omega}$ is the duality between $H^{-1/2}(\Gamma)$, and $H^{1/2}(\Gamma)$, where $\frac{\partial u}{\partial n}$ is the functional $G_u(\cdot)$ acting on $v \in H^{1/2}(\Gamma)$.

2 Variational Methods on Elliptic Problems

2.1 Lax-Milgram Lemma

We will present the most important lemmas in the study of finite elements.

Lemma 2.1 (Lax-Milgram)**.** *Assume that $a(\cdot,\cdot)$ is a bilinear form on a separable Hilbert space, $(H, \|\cdot\|)$, where $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$, that is continuous and coercive:*

(C1). $\exists M > 0$, such that $|a(u,v)| \le M \|u\| \|v\|$, for all $u, v \in H$,

(C2). $\exists m > 0$ such that $a(u,u) \ge m \|u\|^2$, for all $u \in H$.

Assume further that $F : H \to \mathbb{R}$ is a bounded linear functional on H such that there exists $C > 0$ such that $|F(v)| \le c\|v\|$, for all $v \in H$. Then the problem

$$\Big\{ \text{Find } u \in H, a(u,v) = F(v), \forall v \in H$$

has a unique solution, and the solution has the property

$$\frac{1}{M} \|F\|_{H^*} \le \|u\| \le \frac{1}{m} \|F\|_{H^*}, \tag{*}$$

where $\|F\|_{H^*} = \sup_{v\neq 0} \frac{F(v)}{\|v\|}$.

Proof. We only prove (*) holds. Assume that $a(u,v) = F(v)$. Then

$$\|F\|_{H^*} = \sup_{v\neq 0} \frac{F(v)}{\|v\|} = \sup_{v\neq 0} \frac{a(u,v)}{\|v\|} \le \sup_{v\neq 0} \frac{M\|u\|\|v\|}{\|v\|} = M\|u\|,$$

and

$$\|F\|_{H^*} = \sup_{v\neq 0} \frac{F(v)}{\|v\|} \ge \frac{a(u,u)}{\|u\|} \ge \frac{m\|u\|^2}{\|u\|} = m\|u\|,$$

by continuity and coercivity. □

Example 2.2. Consider the following problem. Find u such that

$$\begin{cases} -\Delta u = f, \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ f \in L^2(\Omega). \end{cases}$$

Show that this problem has a unique solution.

Proof. First, taking $v \in H_0^1(\Omega)$, and multiplying it by our equations above, and integrating over Ω, we have for all $v \in H_0^1(\Omega)$,

$$\begin{aligned} \int_\Omega -\Delta u \cdot v &= \int_\Omega fv \\ \int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, ds &= \int_\Omega fv \\ (\nabla u, \nabla v) &= (f, v), \quad \forall v \in H_0^1(\Omega), \end{aligned}$$

where we applied Green's identities, and noted that $v = 0$ on $\partial\Omega$. We now define $H = H_0^1(\Omega)$, $F(v) = \int_\Omega fv \in (H_0^1(\Omega))^* = H^*$, and $a(u,v) = (\nabla u, \nabla v)$. We need to check that a is coercive and continuous. For continuity, we use the Cauchy-Schwarz inequality (twice), and for simplicity we assume that we are in 2D, yielding

$$\begin{aligned}
|a(u,v)| &= |(\nabla u, \nabla v)| \\
&= \int_\Omega \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}\,dx \\
&\le \left(\int_\Omega \left(\frac{\partial u}{\partial x_1}\right)^2\right)^{1/2}\left(\int_\Omega \left(\frac{\partial v}{\partial x_1}\right)^2\right)^{1/2} + \left(\int_\Omega \left(\frac{\partial u}{\partial x_2}\right)^2\right)^{1/2}\left(\int_\Omega \left(\frac{\partial v}{\partial x_2}\right)^2\right)^{1/2} \\
&\le \sqrt{\int_\Omega \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2}\sqrt{\int_\Omega \left(\frac{\partial v}{\partial x_1}\right)^2 + \left(\frac{\partial v}{\partial x_2}\right)^2} \\
&= \|\nabla u\|\,\|\nabla v\| \le \|u\|_{H^1(\Omega)}\,\|v\|_{H^1(\Omega)}\,.
\end{aligned}$$

Therefore, we have continuity. For coercivity, we can see the following:

$$\begin{aligned}
a(u,v) = \|\nabla u\|^2_{L^2(\Omega)} &\ge C(\Omega)\,\|u\|^2_{L^2(\Omega)}\,, \\
2\,\|\nabla u\|^2_{L^2(\Omega)} &\ge C(\Omega)\,\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}\,,
\end{aligned}$$

therefore

$$a(u,v) \ge \frac{1}{2}\min\{C(\Omega),1\}\left(\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}\right) = \frac{1}{2}\min\{C(\Omega),1\}\,\|u\|^2_{H^1(\Omega)}\,.$$

Therefore we have coercivity. □

2.2 Ritz-Galerkin Method

Assume that the hypothesis of the Lax-Milgram theorem, with V as a Hilbert space holds. That is

1. $a : V \times V \to \mathbb{R}$, a bilinear form with coercivity and continuity,
2. $F : V \to \mathbb{R}$ is a bounded linear functional,
3. $V_h \subset V$ is a finite dimensional subspace.

Then Lax-Milgram tells us that the problem

$$\left\{\text{Find } u \in V \text{ such that } a(u,v) = F(v), \quad \forall v \in V, \right. \tag{*}$$

has a unique solution. We would like the discrete variational formulation of this problem, which is the following:

$$\left\{\text{Find } u_h \in V_h \text{ such that } a(u_h,v_h) = F(v_h), \quad \forall v_h \in V. \right. \tag{**}$$

We can apply Lax-Milgram on V_h since it satisfies all the conditions on $V_h \subset V$, and thus we have a unique solution to (*). We call u_h the Ritz-Galerkin approximation of u. We now have to answer two questions. Firstly, how do we compute u_h? Secondly, how close is u_h to u? To answer the first question, the basic idea is to use bases, since we are in a finite dimensional space V_h. We can then try writing

$$u_h = \sum_{i=1}^{n} \alpha_i \phi_i,$$

where $\{\phi_1, \ldots, \phi_n\}$ is a basis for V_h, and $\alpha_i \in \mathbb{R}$. Plugging this representation into our discrete variational formulation (**) will lead us to an $N \times N$ linear system in $(\alpha_1, \ldots, \alpha_n)$. For the second question, the proximity of u_h to u can be computed by Cea's theorem.

Theorem 2.3 (Cea's theorem)**.** *Assume $a : V \times V \to \mathbb{R}$ and $F \in V^*$ satisfying the Lax-Milgram assumptions and hypotheses. Also assume that $V_h \subset V$ is a subspace. If u is the solution of (*), and u_h is the solution of (**), then*

$$\|u - u_h\|_V \leq \frac{M}{m} \inf_{v_h \in V_h} \|u - v_h\| .$$

Proof. Using the coercivity of a, we have that

$$m \|u - u_h\|^2 \leq a(u - u_h, u - u_h).$$

Since u is the solution of (*), we get that $a(u, v_h) = F(v_h)$, for all $v_h \in V_h$, and from (**), we get that $a(u_h, v_h) = F(v_h)$, for all $v_h \in V_h$. By subtracting these two equations, we get that

$$a(u - u_h, v_h) = 0, \forall v_h \in V_h. \tag{***}$$

Now consider

$$\begin{aligned} m \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h + v_h - u_h) \\ &= a(u - u_h, u - v_h) + \underbrace{a(u - u_h, v_h - u_h)}_{=0} \\ &= a(u - u_h, u - v_h) \\ &\leq M \|u - u_h\| \|u - v_h\| . \end{aligned}$$

Dividing both sides by $\|u - u_h\| m$, we get

$$\|u - u_h\| \leq \frac{M}{m} \|u - v_h\| , \forall v_h \in V_h \Longrightarrow \|u - u_h\| \leq \frac{M}{m} \inf_{v_h \in V_h} \|u - v_h\| .$$

□

Before moving on, consider (***). If a is a symmetric bilinear form, a is then an inner product, which due to coercivity and boundedness, produces and equivalent norm on V. Using (***), we can conclude that u_h is the projection of u onto V_h, with respect to $a(\cdot, \cdot)$. In this case, u_h is the Galerkin projection.

3 Finite Element Spaces

Definition 3.1. A subdivision of a bounded domain $\Omega \subset \mathbb{R}^d$ is a finite collection of open sets $\{T_i\}$ such that

1. $T_i \cap T_j = \emptyset$, for all $i \neq j$,
2. $\bigcup_i \overline{T_i} = \overline{\Omega}$,
3. $\operatorname{Int} T_i \neq \emptyset$.

Definition 3.2. A triangulation $\mathscr{T}_h$ of Ω, where Ω is a polynomial domain in $\mathbb{R}^2$ or a polyhedral domain in $\mathbb{R}^3$ (i.e. $\partial\Omega$ is a union of polyhedral faces) is a subdivision of Ω consisting of "triangles" (element domains) with the property that no vertex of any triangle lies in the interior of an edge or a face of another triangle.

Triangles are not actually triangles. But they can be. Given a triangulation $\mathscr{T}_h$ of Ω, a finite element space associated with $\mathscr{T}_h$ is a piecewise polynomial space with respect to $\mathscr{T}_h$ with specific finite element conditions.

Definition 3.3. If $T \in \mathscr{T}_h$, the triplet $(T, P_T, \mathcal{N}_T)$ is a finite element if

1. P_T is a finite dimensional space of polynomials on T,
2. $\mathcal{N}_T = \{\mathcal{N}_1, \ldots, \mathcal{N}_k\}$ is a set of linear functionals of P_T that are unisolvent on P_T i.e. if $u \in P_T$ and $\mathcal{N}_i(u) = \gamma_i$, for $i = 1, 2, \ldots, k$, then $u \in P_T$ satisfying $\mathcal{N}_i(u) = \gamma_i$ is unique. In other words, $\mathcal{N}_T$ is a basis on the dual of P_T.

In this case, $N_1, \ldots, N_k$, where $k = \dim P_T$ are called degrees of freedom of T.

Example 3.4. Let $T = (0, 1)$, $P_T = \{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\}$. We see that $\dim(P_T) = 3$, therefore we will need $\mathcal{N}_T = \{\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3\}$. Take for instance

$$\mathcal{N}_1(u) = u(0), \mathcal{N}_2(u) = u(1/2), \mathcal{N}_3(u) = u(1), u \in P_T.$$

These are all bounded linear functionals. If $u = ax^2 + bx + c$, we can now check unisolvence by considering

$$\begin{aligned}
\mathcal{N}_1(\delta_1) = \delta_1 &\implies c = \delta_1 \\
\mathcal{N}_2(\delta_2) = \delta_2 &\implies \frac{a}{4} + \frac{b}{2} + c = \delta_2 \\
\mathcal{N}_3(\delta_3) = \delta_3 &\implies a + b + c = \delta_3.
\end{aligned}$$

This can be solved via a matrix representation:

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1/4 & 1/2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \implies \det A \neq 0.$$

We see our system has a unique solution, therefore, $\mathcal{N}_T$ is unisolvent. Note that the set of degrees of freedom is not unique for a fixed P_T. For example, we can take $\hat{N_T} = \{\mathcal{N}_1, \hat{\mathcal{N}_2}, \mathcal{N}_3\}$, where $\hat{\mathcal{N}_2} = \int_0^1 u(x)\,dx$. We can show again that this is a set of degrees of freedom.

Definition 3.5. Assume we have a finite element $(T, P_T, \mathcal{N}_T)$. A basis $\{\phi_1, \ldots, \phi_k\}$ for P_T is called a dual to $\mathcal{N}_T$ (or nodal basis, or shape functions) if

$$\mathcal{N}_i(\phi_j) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

One advantage of finding a nodal basis is that any $u \in P_T$ can be written

$$u = \sum_{i=1}^{k} \mathcal{N}_i(u)\phi_i,$$

since if

$$v = \sum_{i=1}^{k} \mathcal{N}_i(u)\phi_i,$$

then

$$\mathcal{N}_j(v) = \sum_{i=1}^{k} \mathcal{N}_i(u)\mathcal{N}_j(\phi_i) = \mathcal{N}_j(u), \forall j = 1, 2, \ldots, k.$$

Therefore $v = u$ since $\mathcal{N}_T$ is unisolvent.

Example 3.6. As a continuation of Example 3.4, take $\mathcal{N}_T = \{\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3\}$. We now look for ϕ_1 such that $\mathcal{N}_1(\phi_1) = 1$, $\mathcal{N}_2(\phi_1) = 0$, $\mathcal{N}_3(\phi_1) = 0$. Conducting this process, we can find $\phi_1(x) = 2(x - 1/2)(x - 1)$. Similarly, we can find $\phi_2 = (4/3)(1 - x)(1 + x)$, and $\phi_3(x) = 2x(x - 1/2)$. Therefore $\{\phi_1, \phi_2, \phi_3\}$ is a nodal basis.

3.1 Global Degrees of Freedom and Global Nodal Functions

Definition 3.7. Assume $\mathscr{T}_h = \bigcup_i T_i$ is a triangulation of Ω, where $T = T_i$ corresponds to $(T, P_T, \mathcal{N}_T)$, where $\mathcal{N}_T = \{\mathcal{N}_1, \ldots, \mathcal{N}_k\}$ and $P_T = \text{span}\{\phi_1, \ldots, \phi_k\}$ is a nodal basis, i.e. $\mathcal{N}_i(\phi_j) = \delta_{ij}$. We will define the global DoF and global nodal basis with the following conditions.

1. If a local DoF is asociated with a vertex or edge that is shared by 2 or more triangles, then this leads to only one DoF.

2. A global nodal function for a node or an edge is obtained by "combining" the local nodal functions that are adjacent to the node or edge. Extend the functions by 0 on the rest of the triangle.

3. For a local DoF associated with evaluation or integration at interior (to T_i) nodes (or nodes adjacent to only one triangle), we declare it a global DoF and the corresponding nodal function is just the extension by 0 of the local nodal function.

4. We use a global counting of the DoF and the nodal functions based on conditions 1-3.

Example 3.8. How do we use these conditions? Let $\Omega = (0, 2) = (0, 1) \cup (1, 2)$. First, let $T = (0, 1)$, and then consider DoF evaluation at $0, 1/2$, and 1. Then we can find that the nodal functions are

$$\phi_1 = 2\left(x - \frac{1}{2}\right)(x - 1), \phi_2 = 4x(1 - x), \phi_3 = 2x\left(x - \frac{1}{2}\right).$$

Now, for $T = (1, 2)$, we will use DoF evaluation at $1, 3/2, 2$, and our nodal basis is

$$\phi_3 = 2\left(x - \frac{3}{2}\right)(x - 2), \phi_4 = 4(2 - x)(x - 1), \phi_5 = 2\left(x - \frac{3}{2}\right)(x - 1).$$

To find our global DoF, we now define $u : (0, 2) \to \mathbb{R}$, and define the five linear functionals

$$\mathcal{N}_1(u) = u(0), \mathcal{N}_2(u) = u(1/2), \mathcal{N}_3(u) = u(1), \mathcal{N}_4(u) = u(3/2), \mathcal{N}_5(u) = u(2).$$

Then using these functionals, we can then define our global functions by extending the local bases to 0 in the triangles that are not using the local nodal basis. As well, note the overlap of the vertex

at $x = 1$:

$$\phi_1(x) = \begin{cases} 2\left(x - \frac{1}{2}\right)(x-1), & x \in (0,1) \\ 0, & x \in (1,2), \end{cases}$$

$$\phi_2(x) = \begin{cases} 4x(1-x), & x \in (0,1) \\ 0, & x \in (1,2), \end{cases}$$

$$\phi_3(x) = \begin{cases} 2x(x - \frac{1}{2}), & x \in (0,1) \\ 2(x - \frac{3}{2})(x-2), & x \in (1,2), \end{cases}$$

$$\phi_4(x) = \begin{cases} 0, & x \in (0,1) \\ 4(2-x)(x-1), & x \in (1,2), \end{cases}$$

$$\phi_5(x) = \begin{cases} 0, & x \in (0,1) \\ 2(x - \frac{3}{2})(x-1), & x \in (1,2). \end{cases}$$

Note that by constructing these global nodal functions, we still have that $\mathcal{N}_i(\phi_j) = \delta_{ij}$, for all $i, j = 1, \ldots, 5$ and each of these is locally supported, which allows for better implementation. If we had taken $\mathscr{T}_h = (0,1) \cup (1,2)$, with $\Omega = (0,2)$, and defined $V_h = \text{span}(\phi_1, \ldots, \phi_5) \subset H^1((0,2))$, then we can also define $V_h = \{v \in H^1((0,2)) \mid v|_T \in P^2 \text{ for any } T\}$[1].

3.2 Linear Elements in 2D, 3D

Definition 3.9. Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain, and $\mathscr{T}_h$ a triangulation of Ω, with $\mathscr{T}_h = \bigcup_i T_i$. Then $(T = T_i, P_T, \mathcal{N}_T)$, the *linear element* is defined as follows:

$$T = [z_1, z_2, z_3], P_T = \{ax + by + c \mid a, b, c \in \mathbb{R}\} = P^1 = \text{span}(1, x, y), \mathcal{N}_T = \{\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3\},$$

where $\mathcal{N}_i(v) = v(z_i)$, for $i = 1, 2, 3$.

Note that $\mathcal{N}$ needs to be unisolvent for this to make sense, so to check this, we need to show that $\mathcal{N}_i(v) = 0$ for all i for some $v \in P_T$ implies that $v = 0$. To this end, assume that $z_i = (x_i, y_i), i = 1, 2, 3$. This means that if $v = ax + by + c$, we have

$$\begin{array}{l} ax_1 + by_1 + c = 0 \\ ax_2 + by_2 + c = 0 \\ ax_3 + by_3 + c = 0 \end{array} \Longrightarrow \underbrace{\begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}}_{A} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We can see that $|A| = (z_2 - z_1) \times (z_3 - z_2) = 2 \times \text{Area}(T) := 2\,|T| \neq 0$. Therefore this is unique and thus $v = 0$. Therefore $\mathcal{N}_T$ is unisolvent. To find a dual nodal basis, we need to investigate affine or barycentric coordinates. Let $T = [a_1, a_2, a_3] \subset \mathbb{R}^2$, (or $T = [a_1, a_2, a_3, a_4] \subset \mathbb{R}^3$). In the 2D case, define

$$T = \left\{ \overline{x} = \sum_{j=1}^{n} \lambda_j a_j \mid 0 \leq \lambda_j \leq 1, \sum_{j=1}^{3} \lambda_j = 1 \right\}.$$

[1] We note this is a bad way to define it if the ϕ's are not known, so proceed with caution here.

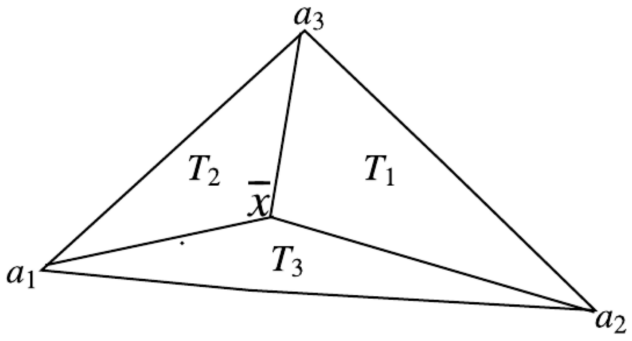

Figure 1: Barycentric coordinates.

We call $(\lambda_1, \lambda_2, \lambda_3)$ the barycentric coordinates of $\overline{x} = [x, y]^T$. They are unique if we assume that $\lambda_1 + \lambda_2 + \lambda_3 = 1$. In fact we can show that for all $\overline{x} \in T$, there exists a unique $(\lambda_1, \lambda_2, \lambda_3)$ such that $\overline{x}$ can be represented by the barycentric coordinates. However, how can we find these $(\lambda_1, \lambda_2, \lambda_3)$ as functions of $\overline{x} = [x, y]^T$? If we assume that $a_i = [x_i, y_i]^T$, for $i = 1, 2, 3$, we know that

$$\begin{gathered} \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ \lambda_1 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \lambda_2 \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + \lambda_3 \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \end{gathered} \Longrightarrow \underbrace{\begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}}_{A} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ y \end{bmatrix}.$$

We can see that $\det A = 2\,|T| \neq 0$. Therefore we have unique coordinates for each $\overline{x}$. By Cramer's rule, we can see that the solution to the problem is

$$\lambda_1 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x & x_2 & x_3 \\ y & y_2 & y_3 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix}} = \frac{\text{Area}(T_1)}{\text{Area}(T)} = \frac{|T_1|}{|T|}, \quad \lambda_2 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x & x_3 \\ y_1 & y & y_3 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix}} = \frac{\text{Area}(T_2)}{\text{Area}(T)} = \frac{|T_2|}{|T|},$$

$$\lambda_3 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x \\ y_1 & y_2 & y \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix}} = \frac{\text{Area}(T_3)}{\text{Area}(T)} = \frac{|T_3|}{|T|},$$

where T_1, T_2, and T_3 are the triangles. These are now linear functions in x, y and are the nodal basis, since $\mathcal{N}_i(\lambda_i) = \delta_{ij}$. So $\lambda_1, \lambda_2, \lambda_3$ form a nodal basis for $(T, P_T, \mathcal{N}_T)$, where $\mathcal{N}_T$ is the evaluation at a_1, a_2, a_3. Note that the barycenter of a triangle is the point with coordinates $(1/3, 1/3, 1/3)$. Recalling that $\overline{x} = [x, y]^T$, we can then define any value in T as

$$P(\overline{x}) = \alpha_1 \lambda_1(\overline{x}) + \alpha_2 \lambda_2(\overline{x}) + \alpha_3 \lambda_3(\overline{x}),$$

and see that $P(a_i) = \alpha_i$. Therefore, we have a nodal basis:

$$P(\overline{x}) = \sum_i P(a_i) \lambda_i(\overline{x}),$$

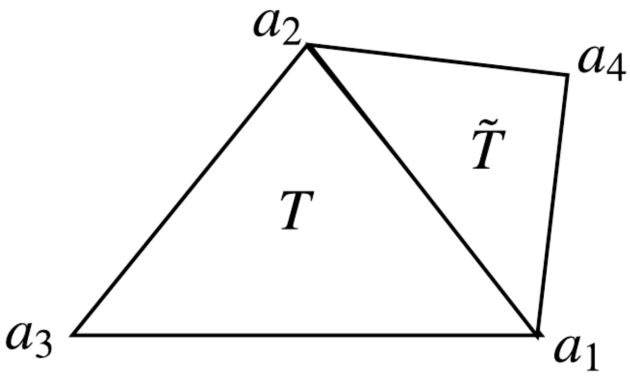

Figure 2: Triangle T.

where the λ_i's are linear functions of x, y. Next, let $\overline{x} \in [a_1, a_2]$. Then we can write $\overline{x} = (1-\theta)a_1 + \theta a_2$, for some $\theta \in [0,1]$. As well, we can see that

$$\begin{aligned} P(\overline{x}) = P((1-\theta)a_1 + \theta a_2) &= \sum_{i=1}^{3} P(a_i)\lambda_i((1-\theta)a_1 + \theta a_2) \\ &= (1-\theta)\sum_{i=1}^{3} P(a_i)\lambda_i(a_1) + \theta \sum_{i=1}^{3} P(a_i)\lambda_i(a_2) \\ &= (1-\theta)P(a_1) + \theta P(a_2) \end{aligned}$$

for any $\theta \in [0,1]$. Thus the value of $\overline{x}$ for P on the line $[a_1, a_2]$ is only dependent on the values of the endpoints of the line. Now consider the triangle $T = [a_1, a_2, a_4]$: see Fig. 2. Assume we have a nodal basis of $\{\widehat{\lambda_1}, \widehat{\lambda_2}, \widehat{\lambda_3}\}$. If $q \in P^1[a_1, a_2, a_4]$, then we have that

$$q(\overline{x}) = q(a_1)\widehat{\lambda_1}(\overline{x}) + q(a_2)\widehat{\lambda_2}(\overline{x}) + q(a_4)\widehat{\lambda_4}(\overline{x}).$$

Then we can see that if $P(a_1) = q(a_1)$ and $\overline{x} = (1-\theta)a_1 + \theta a_2$, and we have

$$q(\overline{x}) = (1-\theta)q(a_1) + \theta q(a_2) = P(\overline{x}).$$

This implies that if triangles share an edge, we need that all of the vertices agree, i.e., we must demand continuity on shared vertices and edges.

3.2.1 C^0-P^1 Elements

We now turn to search for the global nodal functions. Consider the space C^0-P^1, the space of continuous piecewise linear functions. Let $\mathscr{T}_h$ be a triangulation of a polygonal domain Ω, and denote the vertices of $\mathscr{T}_h$ by $z_1, \ldots, z_n$. Define the global space

$$V_h = \left\{ v \in C^0(\Omega) \mid v_h|_T \in P_T = P_1, \forall T \in \mathscr{T}_h \right\}.$$

Then the global degrees of freedom are the evaluations at z_i, i.e., $\mathcal{N}_T = \{N_1, \ldots, N_n\}$, where $N_i(u) = u(z_i)$, for $i = 1, 2, \ldots, n$. Then we have a global basis of $\phi_1, \ldots, \phi_n$, where $\phi_i(z_j) = \delta_{ij}$, and u_i is linear on any $T \in \mathscr{T}_h$. Note that if $v_h \in V_h$, we have that $v_h = \sum_{i=1}^{n} \alpha_i \phi_i$, $v_h(z_j) = \alpha_j$, thus $v_h = \sum_{i=1}^{n} v_h(z_i)\phi_i$. Therefore, we can see that any $v_h \in V_h$ is uniquely determined by $v_h(z_i)_{i=1}^{n} \in \mathbb{R}^n$.

3.2.2 C^0-P^2 Elements

Let $(T, P_T, \mathcal{N}_T)$ be a finite element, where $T = [z_1, z_2, z_3]$, and

$$P_T = P^2 = \left\{ax^2 + bxy + cy^2 + dx + ey + f \mid a, b, c, d, e, f \in \mathbb{R}\right\}$$

and $\mathcal{N}_T = \{N_1, N_2, N_3, N_{12}, N_{13}, N_{23}\}$, where $N_i(u) = u(z_i)$, and $N_{ij}(u) = z_{ij}$, $i = 1, 2, 3$, $1 \leq i < j \leq 3$, where z_{ij} is the midpoint of $[z_i, z_j]$. For our nodal basis $\{\phi_1, \phi_2, \phi_3, \phi_{12}, \phi_{13}, \phi_{23}\}$, we now need

$$\phi_i(z_j) = \delta_{ij}, \phi_i(z_{jk}) = 0, \phi_{ij}(z_{kl}) = 1, \text{ if } ij = kl, \phi_{ij}(z_{kl}) = 0 \text{ if } ij \neq kl, \phi_{ij}(z_k) = 0.$$

One can then find that

$$\begin{aligned}
\phi_1 &= \lambda_1(2\lambda_1 - 1), \quad \phi_{12} = 4\lambda_1\lambda_2 \\
\phi_2 &= \lambda_2(2\lambda_2 - 1), \quad \phi_{23} = 4\lambda_2\lambda_3 \\
\phi_3 &= \lambda_2(2\lambda_3 - 1), \quad \phi_{13} = 4\lambda_1\lambda_3.
\end{aligned}$$

Note that the support for a midpoint nodal function is at most 2 triangles. However, a node that is defined on the vertex has support on how many triangles share the node. Defining

$$V_h = \{v \in C^0(\Omega) \mid v_h|_T \in P_T = P^2, \forall T \in \mathscr{T}_h\},$$

we can find the global DoF by taking the evaluations at the nodes and the evaluations at the midpoints of all edges. One crucial question we must ask ourselves now is the following. Are the spaces C^0-P^1 and C^0-P^2 subspaces of $H^1(\Omega)$, for a given triangulation $\mathscr{T}_h$? The answer is yes, and is a corollary of the following theorem.

Theorem 3.10. *Define $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. Assume that the triangulation $\mathscr{T}_h = \{T_i\}_{i \in I}$ is a subdivision of Ω. If $v : \Omega \to \mathbb{R}$ satisfies*

1. *$v|_{T_i} \in H^1(T_i)$ for all $T_i \in \mathscr{T}_h$, and*
2. *for any common faces $F_{ij} = \overline{T_i} \cap \overline{T_j}$, we have that $(v|_{T_i})|_{F_{ij}} = (v|_{T_j})|_{F_{ij}}$ in the trace sense,*

then $v \in H^1(\Omega)$.

Proof. Let $w_j = D_j v = \frac{\partial v}{\partial x_j}$, $j = 1, 2, \ldots, d$, defined on each $T_i \in \mathscr{T}_h$, not global but in each element. We need to prove that

$$\int_\Omega w_j \phi = -\int_\Omega v D_j \phi, \quad \forall \phi \in \mathscr{D}(\Omega).$$

To accomplish this, we use Green's formula, as $v \in H^1(T_i)$,

$$\begin{aligned}
\int_\Omega w_j \phi &= \sum_{T_i \in \mathscr{T}_h} \int_{T_i} w_j \phi \\
&= -\sum_{T_i \in \mathscr{T}_h} \left(\int_{T_i} v D_j \phi - \int_{\partial T_i} v \phi n^{T_i} \right),
\end{aligned}$$

where n^{T_i} is the normal vector of the triangle T_i. Since $\phi = 0$ on $\partial\Omega$, we see that $\int_E v\phi n^E = 0$, where $E \subset \partial\Omega$. We just need to consider the edges that are shared between two triangles. If T_i, T_k share a face F_{ik}, then we know that $n^{T_i} = -n^{T_k} := n_{ik}$. Then on the faces,

$$\sum_{T_i \in \mathscr{T}_h} \int_{\partial T_i} v\phi n^{T_i} = \sum_{F_{ik}} \int_{F_{ik}} \phi \left(v|_{T_i} - v|_{T_k} \right) n_{ik} = 0,$$

by our second hypothesis. This proves that C^0-P^1 and C^0-P^2 are conforming spaces, i.e., subspaces of $H^1(\Omega)$. □

3.3 Finite Element Model Problem

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. Given functions f, u_D, and g, along with a constant $c \geq 0$, we seek to find u such that

$$\begin{aligned} -\Delta u + cu &= f, \quad \text{in } \Omega, \\ u &= u_D, \quad \text{on } \Gamma_D, \\ \frac{\partial u}{\partial n} &= g, \quad \text{on } \Gamma_N, \end{aligned} \tag{*}$$

where $\partial\Omega = \Gamma_D \cup \Gamma_N$. We define $H^1_D = \left\{ v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0 \right\}$. Taking $v \in H^1_D$, and multiplying (*) by v, we get

$$\begin{aligned} (-\Delta u + cu)v &= fv \\ \implies \int_\Omega (-\Delta u + cu)v &= \int_\Omega fv \\ \implies \int_\Omega -\Delta uv + c\int_\Omega uv &= \int_\Omega fv \\ \implies \int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} \frac{\partial u}{\partial n} v + c\int_\Omega uv &= \int_\Omega fv \qquad \text{(Green's second formula)} \\ \int_\Omega \nabla u \cdot \nabla v - \overbrace{\int_{\Gamma_D} \frac{\partial u}{\partial n} v}^{=0} - \int_{\Gamma_N} \overbrace{\frac{\partial u}{\partial n}}^{=g} v + c\int_\Omega uv &= \int_\Omega fv \\ \int_\Omega \nabla u \cdot \nabla v - \int_{\Gamma_N} gv + c\int_\Omega uv &= \int_\Omega fv \\ \implies \int_\Omega (\nabla u \cdot \nabla v + cuv)\, dx &= \int_\Omega fv\, dx + \int_{\Gamma_N} gv\, ds, \quad \forall v \in H^1_D. \end{aligned} \tag{1}$$

We call (1) our variational formulation of (*). The Dirichlet condition $u = u_D$ on Γ_D imposes $v = 0$ on Γ_D, because $\frac{\partial u}{\partial n}|_{\Gamma_D}$ is not known. Therefore $u = u_D$ on Γ_D is called an essential boundary condition, and $V = H^1_D$ is called the test space. The Neumann boundary condition $\frac{\partial u}{\partial n} = g$ on Γ_N is called a natural boundary condition, since it does not imply any restriction on the test space. We reformulate (1) to the following problem. Find $u \in H^1(\Omega)$ such that

$$\begin{aligned} u &= u_D \text{ on } \Gamma_D \\ a(u, v) &= (f, v) + \langle g, v \rangle_{\Gamma_N}, \quad \forall v \in V = H^1_D(\Omega). \end{aligned} \tag{2}$$

We cannot use Lax-Milgram here, since u, v are in different spaces. We can however prove that if $f \in L^2(\Omega), g \in L^2(\Gamma_N)$, (or $H^{-1/2}(\Gamma_N)$), and $u \in H^{1/2}_\Gamma(\Omega) = \{u|_{\Gamma_D} \mid u \in H^1(\Omega)\}$, then (2) has a unique solution.

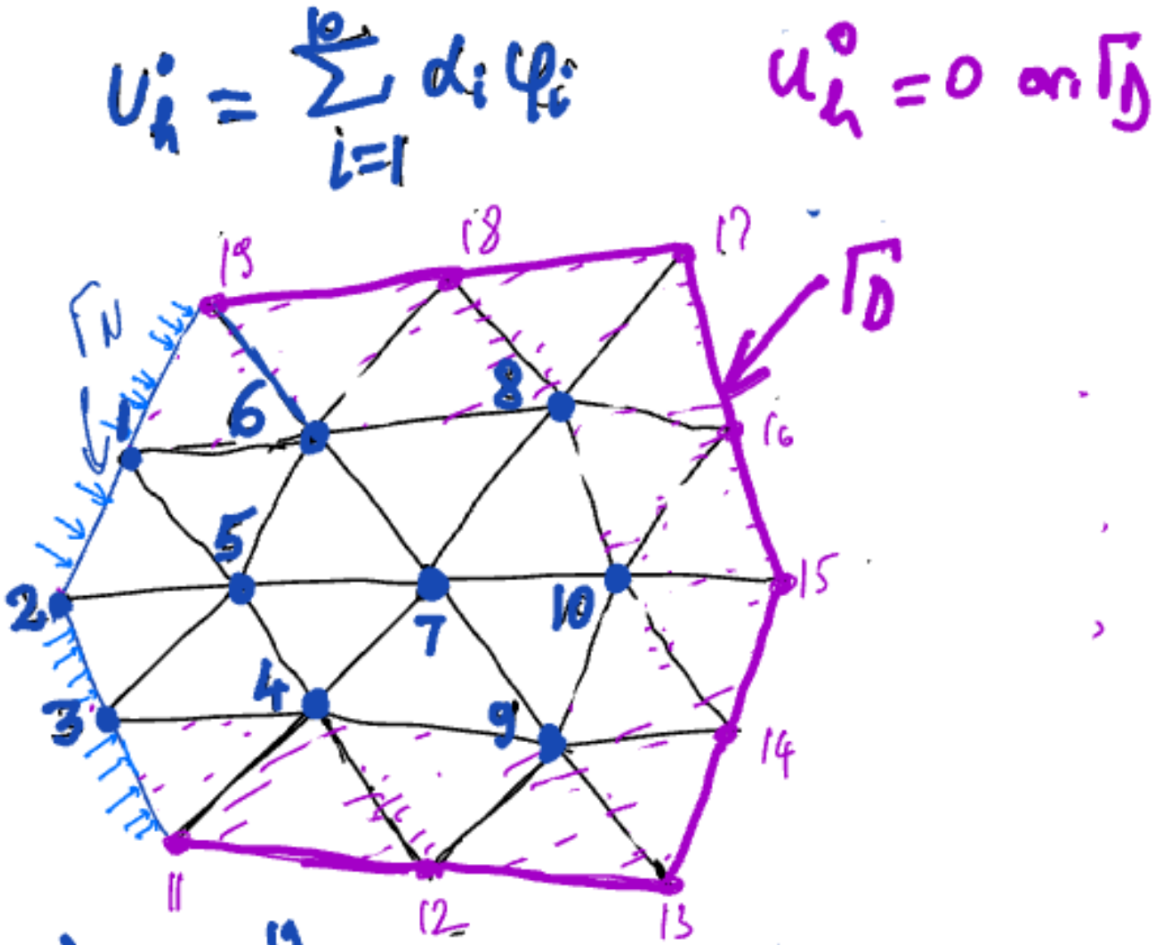

Figure 3: One such triangulation.

To apply Lax-Milgram here, we will assume that u_D is defined on the entire Ω, where $u_D \in H^1(\Omega)$, an extension. Then we write $u = u_0 + u_D$, where $u_0 = u - u_D \in H^1_D(\Omega)$. Then (2) becomes the problem, Find $u_0 \in H^1_D(\Omega)$ such that

$$a(u_0, v) = (f, g) + \langle g, v \rangle_{\Gamma_N} - a(u_D, v), \quad \forall v \in H^1_D(\Omega).$$

Thus v, u_0 are in the same space, and we can apply Lax-Milgram.

3.4 Discretization of C^0-P^1 Galerkin Finite Element

Let $\mathscr{T}_h$ be a triangulation of Ω. Assume that $V_h = \text{span}\{\phi_1, \ldots, \phi_n\}$. Let N be the number of vertices in $\mathscr{T}_h$. For simplication, Let the ϕ_i's be the nodal basis, where $\phi_i(z_j) = \delta_{ij}$, $i, j = 1, 2, \ldots, N$. Since $V_h \subset H^1(\Omega)$, we now look for $u_h = \sum_{i=1}^N \alpha_i \phi_i$, where $\alpha_i = u_h(z_i), i = 1, \ldots, N$. If we define $V_{h,\Gamma_D} = V_h \cap H^1_D(\Omega) = \text{span}\{\phi_1, \ldots, \phi_n\}$, then the discrete variational formulation becomes, Find $U_h \in V_h$ such that

$$\begin{aligned} u_h(z_k) &= u_D(z_k), k = n+1, n+2, \ldots, N, \\ a(u_h, \phi_i) &= (f, \phi_i) + \langle g, \phi_i \rangle_{\Gamma_N}, i = 1, 2, \ldots, n. \end{aligned} \tag{*}$$

We can then see that u_h can be written as

$$u_h = \underbrace{\sum_{i=1}^{n} \alpha_i \phi_i}_{u_h^0} + \underbrace{\sum_{k=n+1}^{N} u_D(z_k)\phi_k}_{u_h^D}.$$

Computing this splitting, and using this as we did in the previous continuous case, we can the prove that we have a unique solution to (*). TO this end, we can then write our final formulation of the problem: Find $u_h^0 = \sum_{j=1}^{n} \alpha_j \phi_h$, such that

$$\begin{aligned} a(u_h^0, \phi_i) &= (f, \phi_i) + \langle g, \phi_i \rangle_{\Gamma_n} - a(u_h^D, \phi_i), \quad i = 1, 2, \ldots, n \\ \sum_{j=1}^{n} \alpha_j a(\phi_j, \phi_i) &= b_i + t_i - a(u_h^D, \phi_i), \quad i = 1, 2, \ldots, n, \end{aligned}$$

where $b = (b_i = (f, \phi_i))_{i=1}^n$ is called the load vector, $t = (t_i = \int_{\Gamma_N} g\phi_i \, ds)_{i=1}^n$ is called the traction vector, the nodes not on Γ_D, $z_1, \ldots, z_n$ are called free nodes, and the nodes on Γ_D, $z_{n+1}, z_{n+2}, \ldots, z_N$ are called the Dirichlet nodes.

Therefore, we have the following matrix system:

$$\begin{aligned} A = [a(\phi_i, \phi_h)]_{i,j} &= \underbrace{[(\nabla\phi_i, \nabla\phi_j)]_{i,j}}_{S} + \underbrace{c[\phi_i, \phi_j]_{i,j}}_{M}, \\ A^{\text{free}}\alpha &= b(\text{free}) + t(\text{free}) - (A(\text{all}, \text{Dir}) \cdot \overline{u_h^D})(\text{free}), \end{aligned} \tag{1}$$

where

$$\overline{u_h^D} = \begin{bmatrix} u_D(z_{n+1}) \\ \vdots \\ u_D(z_N) \end{bmatrix} = u_D(\text{Dir}),$$

and where $b(\text{free})$ implies just the elements with the free nodes, and $A(\text{all}, \text{Dir})$ means that we take all of the rows and only the columns corresponding to Dirichlet nodes.

3.5 Implementation of the Galerkin Method

To implement (1), we need to build S, M, both $N \times N$, and b, t, both $N \times 1$, from $\mathscr{T}_h$, where N is the number of nodes. We will vectorize this system since this allows us to reduce the complexity of the system from N^2 to either N or $N \log N$. For assembly, we will use no loops and capitalize on the sparse structure of the matrices. Consider a triangle K with vertices z_1, z_2, z_3. The steps are the following.

1. Build the local information, that is, for every triangle K¡ we must construct
   (a) local mass matrices: $M = ((\lambda_i, \lambda_j))_{i,j}^3$, where $(\lambda_i, \lambda_j) = \int_K \lambda_i \lambda_j$,
   (b) local stiffness matrices: $S = ((\nabla\lambda_i, \nabla\lambda_j))_{i,j=1}^3$, where $(\nabla\lambda_i, \nabla\lambda_j) = \int_K \nabla\lambda_i \cdot \nabla\lambda_j$,
   (c) local load vector: $b = \left(\int_K f\lambda_i\right)_{i=1}^3$,
   (d) local traction vector: $t = \int_{\Gamma_N} g\lambda_i$.
2. Assemble the local information by using the MATLAB commands and functions `sparse, accumarray, reshape, bsx fun, kron`.

We will now discuss how to build each of the 4 pieces of local information that we need in step 1.

3.5.1 Local Mass Information

Let K be a generic triangle, with $K = [z_1, z_2, z_3]$. Let $\phi_i|_K = \lambda_i$ be the local nodal function. We now need to compute $\int_K \lambda_i \lambda_j \, d\mathbf{x}$, where $\mathbf{x} = (x, y)^t$. There are several ways to compute this, but we will see that the easiest way is to use the *Holland-Bell formula:*

$$\int_K \lambda_1^m \lambda_2^n \lambda_3^p = \frac{m!n!p!}{(m+n+p+2)} \Delta K,$$

where

$$\Delta K = \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = 2 \cdot \text{Area}(K) = 2\,|K|\,.$$

We can see that with this formula,

$$(\lambda_i, \lambda_i) = \int_K \lambda_i^2 = \frac{2!0!0!}{(2+0+0+2)!} \Delta K = \frac{1}{12} \Delta K,$$

and

$$(\lambda_i, \lambda_j) = \int_K \lambda_i \lambda_j = \frac{1!1!0!}{(1+1+0+2)!} \Delta K = \frac{1}{24} \Delta K, \quad i \neq j$$

We can see then that our mass matrix for K is

$$M_K = \frac{\Delta K}{24} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

which depends only on ΔK or the area of K.

3.5.2 Reference Triangles and Affine Transformations to the Physical Element

Let K be a generic triangle $[z_1, z_2, z_3]$. If we would like to integrate on K, it may be very complicated given the fact that Gaussian quadratures or other numerical integration techniques are defined for a much simpler domain. Thus we define the reference triangle, $\hat{K}$ to be the triangle with coordinates $(0,0), (1,0), (0,1)$ in a coordinate system (ξ, η). We would like to find a linear transformation, $T_K : K \to \hat{K}$ so that we can integrate on $\hat{K}$, rather than K, but still admit the same answer, and allow us to use quadrature techniques. The following theorem gives us this linear transformation..

Theorem 3.11. *Let K be a triangle with vertices $[z_1, z_2, z_3]$, where $z_i = (x_i, y_i)$. Then there exists a unique linear transformation $T_K : K \to \hat{K}$ such that $T(0,0) = z_1, T(1,0) = z_2$, and $T(0,1) = z_3$, and it is given by*

$$T_K(\xi, \eta) = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \mathbf{x}.$$

This transformation preserves the ratio of barycentric coordinates.

Now defining our linear transformation as $T_K \hat{x} = B_K \hat{x} + z_1$, and if $F : K \to \mathbb{R}$, we can see that

$$\int_K F(\mathbf{x}) \, d\mathbf{x} = \int_{\hat{K}} F(T_K(\hat{x})) \cdot \det(B_K) \, d\hat{x},$$

where $\det(B_K) = \Delta K = 2\,|K|$. Thus this linear transformation allows us to have that

$$\int_K F(\mathbf{x}) \, d\mathbf{x} = 2 \int_{\hat{K}} F(T_k(\hat{x}))\,|K| \; d\hat{x}.$$

Thus, for all triangles K, we will need to store B_K since it admits many important properties.

3.5.3 Local Stiffness Matrix

Let $K = [z_1, z_2, z_3]$ be our generic triangle, and we will now need $S_K = ((\nabla\lambda_i, \nabla\lambda_j))_{i,j=1}^3$. Recall that

$$\lambda_1 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x & x_2 & x_3 \\ y & y_2 & y_3 \end{vmatrix}}{\Delta K} = \frac{1}{\Delta K}\left(\begin{vmatrix} x_2 & x_3 \\ y_2 & y_3 \end{vmatrix} - x\begin{vmatrix} 1 & 1 \\ y_2 & y_3 \end{vmatrix} + y\begin{vmatrix} 1 & 1 \\ x_2 & x_3 \end{vmatrix}\right).$$

Therefore, if we were to calculate $\frac{\partial\lambda_1}{\partial x}$ and $\frac{\partial\lambda_1}{\partial y}$, we would find that they are both constants, since λ_1 is linear in y and x. Therefore, we can see that

$$(\nabla\lambda_1, \nabla\lambda_1) = \int_K \left(\frac{\partial\lambda_1}{\partial x}\frac{\partial\lambda_1}{\partial x} + \frac{\partial\lambda_1}{\partial y}\frac{\partial\lambda_1}{\partial y}\right) = \left(\frac{\partial\lambda_1}{\partial x}\frac{\partial\lambda_1}{\partial x} + \frac{\partial\lambda_1}{\partial y}\frac{\partial\lambda_1}{\partial y}\right)|K|.$$

We can then find readily that

$$\begin{aligned}
\frac{\partial\lambda_1}{\partial x} &= \frac{1}{\Delta K}(y_2 - y_3) \\
\frac{\partial\lambda_2}{\partial x} &= \frac{1}{\Delta K}(y_3 - y_1) \\
\frac{\partial\lambda_3}{\partial x} &= \frac{1}{\Delta K}(y_1 - y_2) \\
\frac{\partial\lambda_1}{\partial y} &= \frac{1}{\Delta K}(x_3 - x_2) \\
\frac{\partial\lambda_2}{\partial y} &= \frac{1}{\Delta K}(x_1 - x_3) \\
\frac{\partial\lambda_3}{\partial y} &= \frac{1}{\Delta K}(x_2 - x_1).
\end{aligned}$$

3.5.4 Mesh Structure

To start coding this, we will create a mesh structure in MATLAB, called `T`. Define `nT` to be the number of triangles and `nC` to be the number of nodes. Then to create `T`, we need to have 4 inputs:

1. `T.coordinates`, the x, y coordinates of all the nodes in the mesh, which is $\texttt{nC} \cdot 2$.
2. `T.elements`, the elements of the mesh in therms of the nodes, which is $\texttt{nT} \cdot 3$.
3. `T.dirichlet`, the Dirichlet boundary edges in terms of the nodes.
4. `T.neumann`, the Neumann boundary edges in terms of the nodes.

There are a lot of outputs that are necessary for this triangulation. Here are the functions that we will need to output:

Table 1: Outputs Needed for Triangulation

Output	Explanation
`T.baryc`	the (x, y) coordinates of the barycenter of each element
`T.detB`	the determinant of the matrix B_K
`T.g11`	$\frac{\partial \lambda_1}{\partial x}$ of size $\texttt{nT} \cdot 1$
`T.g12`	$\frac{\partial \lambda_1}{\partial y}$ of size $\texttt{nT} \cdot 1$
`T.g21`	$\frac{\partial \lambda_2}{\partial x}$ of size $\texttt{nT} \cdot 1$
`T.g22`	$\frac{\partial \lambda_2}{\partial y}$ of size $\texttt{nT} \cdot 1$
`T.g31`	$\frac{\partial \lambda_3}{\partial x}$ of size $\texttt{nT} \cdot 1$
`T.g32`	$\frac{\partial \lambda_3}{\partial y}$ of size $\texttt{nT} \cdot 1$
`T.gx`	the partial derivatives with respect to x of nodal functions g_{11}, g_{21}, g_{31}
`T.gy`	the partial derivatives with respect to y of nodal functions g_{12}, g_{22}, g_{32}
`T.midptNeu`	coordinates of the midpoint of Neumann edges
`T.ledgesNeu`	lengths of the Neumann edges

3.5.5 How to Implement the Code

The first important function we need to recall is `sparse`. Let A be an $N \times N$ matrix. To store this matrix in MATLAB, we would need N^2 storage for this, but if A is a sparse matrix, then the number of nonzero entries is approximately kN, where $k \ll N$. It would benefit us to only use kN blocks of memory for this. For example, consider the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 9 \\ 0 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix}.$$

To save this with the `sparse` function, MATLAB saves the indices of the nonzero entries, and the corresponding entries. Letting the row index $i \in I$, and the column indices $j \in J$, and the values are $a_{ij} \to a$, then we have

$$I = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 2 \end{bmatrix}, J = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix}, a = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 9 \end{bmatrix}.$$

We get that `sparse(I,J, a, 4, 3)` gives us the required code to make A, as this implies that $(1, 1) \to 1, (2, 2) \to 3, (4, 2) \to 5, (2, 3) \to 9$. If we type `full(A)`, this will give us back the full matrix in its original representation as above. The optional (4,3) on the end of the function allows us to change the size of the matrix and include more zeros if necessary. For example, then we can have

$$\texttt{sparse(I,J,a,5,5)} \to \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \end{bmatrix}.$$

3.5.6 Assembling the Global Mass Matrix

To construct the load vector, we will need to calculate

$$(f, \phi_i) = \int_\Omega f\phi_i = \sum_{K \text{ adj. to } z_i} \int_K f\phi_i = \sum_{K \text{ adj. to } z_i} \int_K f\lambda_i.$$

So, for each K, we need to compute $\int_K f\lambda_i$, for $i = 1, 2, 3$, with $K = [z_1, z_2, z_3]$. To do this, we will utilize Gaussian quadrature using the barycenter of K, $b_K = \frac{z_1+z_2+z_3}{3}$. We can then approximate this by

$$\int_K F(x)\,dx \approx F(b_k)\,|K| = Q_1(F).$$

This approximation is exact for any linear functions $F(x, y) = ax + by + c$. Therefore, we get that

$$\int_K f\lambda_i = f(z_{123})\lambda(z_{123}) = \frac{1}{3}f(z_{123}) = \int_K f\lambda_2 = \int_K f\lambda_3.$$

To calculate the traction vector $t_i = \int_{\Gamma_N} g\phi_i\,ds$, for all ϕ_i, we will use the midpoint rule

$$\int G(x)\,ds \approx G(m_e)\,|e|\,,$$

where m_e is the midpoint of the edge, and $|e|$ is the length of the edge. Therefore, we need the coordinates of the m_e and the length of e for all $e \subset \Gamma_N$. Then we can see that

$$t_i = \sum_{e\subset\Gamma_i \text{ adj. to } i} \int_e g\phi_i\,ds = \sum_{e\subset\Gamma_N \text{ adj. to } i} g(m_e)\phi(m_i)\,|e| = \frac{1}{2}g(m_e)\,|e|\,.$$

We can now build this matrix by using this approximation.

4 Let's Skip Ahead Abit

To future me, don't skip classes.

4.1 Stokes and Navier Stokes Equations

The Navier-Stokes equations model flow motion, where the liquid is homogeneous, incompressible, and Newtonian, with steady flow. They are as follows.

$$\begin{cases} -\Delta u = \mathrm{Re}(u\cdot\nabla)u + \nabla p = f \\ \nabla\cdot u = 0, \text{ in } \Omega, \\ u = 0, \text{ on } \partial\Omega, \end{cases}$$

where Ω is bounded in $\mathbb{R}^2$, or $\mathbb{R}^3$, $u = \vec{u}$ is the velocity field of the fluid, Re is the Reynolds number, p is the pressure, $f = \vec{f}$ are external forces, $\nu = \frac{1}{\mathrm{Re}}$ is kinematic viscosity, $\Delta u = [\Delta u_1, \Delta u_2, \Delta u_3]^T$, and $(u\cdot\Delta)u = \sum_{i=1}^3 u_i \frac{\partial u}{\partial x_i}$.

If the Reynolds number is small, we are assuming a very viscous flow, like molasses, which we can study as Stokes system:

$$\begin{cases} -\Delta u + \nabla p = f \\ \nabla\cdot u = 0, \text{ in } \Omega \\ u = 0, \text{ on } \partial\Omega. \end{cases}$$

To make p unique, we must choose p such that $\int_\Omega p = 0$, so we are searching $\{p \in L^2(\Omega) \mid \int_\Omega p = 0\}$. We are searching for $u \in (H_0^1(\Omega))^d$. We may replace $\nabla \cdot u = 0$ with $\nabla \cdot u = g$, but we necessarily need

$$\int_\Omega g = \int_\Omega \nabla \cdot u = \int_{\partial\Omega} u \cdot n \, ds = 0.$$

Thus we require $\int_\Omega g = 0$.

Assume $d = 2$. Then we can write our Stokes system explicitly as

$$\begin{cases} -\Delta u_1 + \frac{\partial p}{\partial x_1} = f_1 \text{ in } \Omega \\ -\Delta u_2 + \frac{\partial p}{\partial x_2} = f_2 \text{ in } \Omega \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} = g = 0 \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega. \end{cases}$$

We now would like to study the variational formulation of this problem. We will take three test functions $v_1, v_2 \in H_0^1(\Omega)$, and $q \in L_0^2(\Omega)$. Taking the first equation, multiplying by v_1, and integrating, we get the following, by applying Green's formulas:

$$\begin{aligned} \int_\Omega -\Delta u_1 v_1 + \int_\Omega \frac{\partial p}{\partial x_1} v_1 &= \int_\Omega f_1 v_1 \\ \int_\Omega \nabla u_1 \nabla v_1 - \int_{\partial\Omega} \frac{\partial u_1}{\partial n} v_1 \, ds - \int_\Omega p \frac{\partial v_1}{\partial x_1} + \int_{\partial\Omega} p v_1 n_1 &= \int_\Omega f_1 v_1 \\ \int_\Omega \nabla u_1 \nabla v_1 - \int_\Omega p \frac{\partial v_1}{\partial x_1} &= \int_\Omega f_1 v_1. \end{aligned}$$

Similarly,

$$\begin{aligned} \int_\Omega \nabla u_2 \nabla v_2 - \int_\Omega p \frac{\partial v_2}{\partial x_2} &= \int_\Omega f_2 v_2 \\ -\int_\Omega \operatorname{div} u \cdot q &= -\int_\Omega g q. \end{aligned}$$

By adding the equations together, we get

$$\begin{aligned} \int_\Omega \nabla u_1 \cdot \nabla v_1 + \nabla u_2 \cdot v_2 - \int_\Omega p \operatorname{div} v &= \int_\Omega f \cdot v \\ -\int_\Omega \operatorname{div} u \cdot q &= -\int_\Omega g q. \end{aligned}$$

This leads to the mixed formulation / saddle point problem: Find $(u, p) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega)$ such that

$$\begin{aligned} a(u, v) + b(v, p) &= \langle f, v \rangle, \quad \forall v \in (H_0^1)^2 \\ b(u, q) &= \langle g, q \rangle, \quad \forall q \in L_0^2(\Omega), \end{aligned}$$

where $a(u, v) = (\nabla u, \nabla v), b(u, p) = -\int_\Omega (\operatorname{div} v) p$.

4.2 Laplace Equation as a Mixed Problem

Given $f \in L^2(\Omega), c = c(x) > 0$, a smooth function on $\Omega \subset \mathbb{R}^d, d = 2, 3$, consider the problem: Find u such that

$$\begin{cases} -\operatorname{div}(c\nabla u) = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega. \end{cases}$$

Note that if $c = 1$, we are left with $-\Delta = f$. Denote the flux as $\sigma := c\nabla u$. We now look for $u \in H_0^1(\Omega)$. Then we need to find σ such that $-\operatorname{div}(\sigma) = f \in L^2(\Omega)$. The natural space for σ is

$$H(\operatorname{div}, \Omega) = \{u \in (L^2(\Omega))^d \mid \operatorname{div}(u) =\in L^2(\Omega)\},$$

where the norm is

$$\|z\|^2_{H(\operatorname{div},\Omega)} = \|z\|^2_{L^2(\Omega)} + \|\operatorname{div}(z)\|^2_{L^2(\Omega)}.$$

Therefore, our new system is then

$$\begin{cases} \nabla\sigma = -f \text{ in } \Omega \\ c^{-1}\sigma - \nabla u = 0 \text{ on } \partial\Omega. \end{cases}$$

Letting $z \in H(\operatorname{div}, \Omega)$, and $v \in H_0^1(\Omega)$, multiplying and integrating the first equation, we get

$$\int_\Omega c^{-1}\sigma \cdot z - \int_\Omega \nabla u z = 0, \quad \forall z \in H(\operatorname{div}, \Omega).$$

For the second formula, we first note that Green's formula admits

$$\int_\Omega \operatorname{div}(\sigma) v \, dx = -\int_\Omega \sigma \cdot \nabla v \, dx + \underbrace{\int_{\partial\Omega} (\sigma) \cdot v \, ds}_{=0} = -\int_\Omega \sigma \cdot \nabla v \, dx.$$

Therefore, the second equation gives us

$$-\int_\Omega \sigma \cdot \nabla v \, dx = -\int_\Omega dv \, dx, \forall v \in H_0^1(\Omega).$$

Defining $a(\sigma, z) = \int_\Omega c^{-1}\sigma \cdot z \, dx$ and $b(\sigma, v) = -(\sigma, \nabla v)$, for all $(\sigma, v) \in H(\operatorname{div}, \Omega) \times H_0^1(\Omega)$. We can now reformulate the problem: Find $(\sigma, u) \in H(\operatorname{div}, \Omega) \times H_0^1(\Omega)$ such that

$$\begin{aligned} a(\sigma, z) + b(z, u) &= 0, \quad \forall z \in H(\operatorname{div}, \Omega) \\ b(\sigma, v) &= \langle f, g \rangle, \quad \forall v \in H_0^1(\Omega). \end{aligned}$$

This is a similar system to our Stokes system, implying that this problem can probably be written in some general form.

4.3 Mixed or Saddle Point Problems

The abstract general form for a mixed method or saddle point problem on $\Omega \subset \mathbb{R}^d$ is as follows. Let V, Q be Hilbert spaces, with bilinear forms $a(\cdot,\cdot) : V \times V \to \mathbb{R}, b(\cdot,\cdot) : V \times Q \to \mathbb{R}$. Denote $a_0(\cdot,\cdot)$ as the inner product on V, and $(\cdot,\cdot)$ the inner product on Q. We solve the problem: Find $(u, p) \in V \times Q$ such that

$$\begin{cases} a(u, v) + b(v, p) = \langle f, v \rangle, \quad \forall v \in V \\ b(u, q) = \langle g, q \rangle, \quad \forall q \in Q. \end{cases}$$

We now define $B : V \to Q$, where $(Bv, q) = b(v, q), \forall v \in V, q \in Q, B^* : Q \to V^*$, where $\langle B^* q, v\rangle = b(v, q)$, $\forall v \in V, q \in Q, A_0 : V \to V^*$, where $\langle A_0 u, v\rangle = a_0(u, v)$, for all $u, v \in V$. We define $\|u\|_V = |u|$ and $\|p\|_Q = \|p\|$ for simplicity. As well, define $V_0 = \{v \in V \mid b(v, q) = 0, \forall q \in Q\} = \{v \in V \mid Bv = 0\} = \ker(B)$.

Example 4.1. For Stokes problem, $V = H_0^1(\Omega), Q = L_0^2(\Omega)$. As well, since $b(v, q) == \int_\Omega \operatorname{div} v \cdot q \, dx$. Then we get

$$a_0(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \quad (p, q) = \int_\Omega pq \, dx.$$

Solving for B, we now see

$$(Bv, q) = b(v, q) = -\int_\Omega \operatorname{div} v \cdot q = -(\operatorname{div} v, q), \forall q \in L_0^2(\Omega).$$

Therefore, $Bv = -\operatorname{div} v$. As well, using the definition of the weak derivative, we get

$$\langle B^* q, v\rangle = b(v, q) = -\int_\Omega \operatorname{div} v \cdot q = \int_\Omega \nabla q \cdot v = (\nabla q, v), \quad \forall v \in H_0^1(\Omega).$$

Therefore, $B^* q = \nabla q$. For the space of V_0, we get

$$V_0 = \{v \in (H_0^1(\Omega))^2 \mid Bv = 0\} = \{v \in (H_0^1(\Omega))^2 \mid \operatorname{div} = 0\},$$

and thus, V_0 is the space of divergence free functions. Lastly, A_0 can be defined as

$$\langle A_0 u, v\rangle = \langle \nabla u, \nabla v\rangle = -(\Delta u, v) \Longrightarrow A_0 u = -\Delta u.$$

Why is this called the Saddle point problem? Consider the Lagrangian

$$\mathcal{L}(v, q) = \frac{1}{2} a(v, v) - \langle f, v\rangle + b(v, q) - \langle g, q\rangle .$$

Then our solution to Stokes problem can be thought of as

$$(u, p) = \inf_{q \in Q} \sup_{v \in V} L(v, q).$$

This is achieved when $v = u, q = p$, and this is called the saddle point property, and the solution must be obtained at the maximum over V and minimum over Q.

4.4 Brezzi's Theorem

We will end off the finite element notes with this theorem, and an application to Navier Stokes.

Theorem 4.2. *Assume the following.*

1. *a, b are continuous bilinear forms.*
2. *$a(u, v) \le M_0 |u| |v|, \forall u, v \in V$, and $b(u, q) \le M_1 |u| \|q\|, \forall u \in V, q \in Q$.*
3. *$a(\cdot, \cdot)$ is coercive on V_0, also known as* kernel ellipticity*, i.e., $a(u, u) \ge m_0 |u|^2$, for all $u \in V_0$.*

4. The LBB (Ladyshenskaja-Babuska-Brezzi) condition holds: there exists an $m > 0$ such that

$$\sup_{v \in V} \frac{b(v,p)}{|v|} \geq m \|p\|, \forall p \in Q,$$

which can be rewritten as the inf-sup condition,

$$\inf_{p \in P} \sup_{v \in V} \frac{b(v,p)}{\|p\| \, |v|} \geq m.$$

Then the mixed variation / saddle point problem has a unique solution $(u,p) \in V \times Q$, and

$$|u|_V + \|p\|_Q \leq C(\|f\|_{V^*} + \|g\|_{Q^*}),$$

where C is a constant of m, m_0, M, M_0.

We claim that the Stokes system satisfies Brezzi's theorem.

4.5 Navier Stokes Equations

Consider the Navier Stokes equations:

$$\begin{aligned} -\Delta u + \frac{1}{\nu}(u \cdot \nabla)u + \nabla p &= f, \text{ in } \Omega, \\ \nabla \cdot u &= 0, \text{ in } \Omega, \\ u &= 0, \text{ on } \partial\Omega. \end{aligned}$$

Assume that the Navier Stokes problem has a solution, which is a Millennium problem, currently. The technique to find a good approximation to the solution of the Navier Stokes problem is as follows. We solve the corresponding Stokes problem to get $(u^{(0)}, p^{(0)})$, and use Picard Iteration to solve for $(u^{(k+1)}, p^{(k+1)})$, by solving

$$\begin{aligned} -\Delta u^{(k+1)} + \frac{1}{\nu}(u^{(k)} \cdot \nabla)u^{(k+1)} + \nabla p^{(k+1)} &= f, \text{ in } \Omega \\ \nabla \cdot u^{(k+1)} &= 0, \text{ in } \Omega, \\ u^{(k+1)} &= 0, \text{ on } \partial\Omega. \end{aligned}$$

This is called the Oseen problem, which has a general form which we have good theory for.

The general Oseen problem is, given $w \in (H_0^1(\Omega))^d$, and $\operatorname{div} w = 0, V = (H_0^1(\Omega))^d, Q = L_0^2(\Omega)$, find $(u,p) \in V \times Q$ such that

$$\begin{aligned} -\Delta u + \frac{1}{\nu}(w \cdot \nabla)u + \nabla p &= f, \text{ in } \Omega \\ \nabla \cdot u &= 0, \text{ in } \Omega \\ u &= 0, \text{ on } \partial\Omega. \end{aligned}$$

To construct the variational formulation for this problem, multiplying $v \in V$, and $q \in Q$ by the corresponding equations and integrating, we get two problems

$$\begin{aligned} \int_\Omega \nabla u \cdot \nabla v + \frac{1}{\nu}\int_\Omega (w \cdot \nabla u) \cdot v - \int_\Omega p \operatorname{div} v &= \int_\Omega f v, \forall v \in V \\ -\int_\Omega \operatorname{div} u \cdot q &= 0, \forall q \in Q. \end{aligned}$$

We claim that this problem is linear, rather than nonlinear. To see this, note that

$$c(w,u,v) = \frac{1}{\nu}\int_\Omega (w\cdot\nabla u)\cdot v = \left(w\cdot\begin{bmatrix}\frac{\partial u}{\partial x_1}\\ \frac{\partial u}{\partial x_2}\end{bmatrix}, \begin{bmatrix}v_1\\ v_2\end{bmatrix}\right) = \sum_{i,j=1}^{d}\int_\Omega w_j\frac{\partial u_i}{\partial x_j}v_i.$$

Therefore, if we fix w, we see that this is a bilinear form, and thus, this is linear. We do note as well that c is antisymmetric in the 2nd and 3rd components, i.e., $c(w,u,v) = -c(w,v,u)$. Defining $a(u,v) = a_0(u,v) + c(w,u,v)$, and $b(v,p) = -\int_\Omega \operatorname{div} v\cdot p$, where $Bv = -\operatorname{div} v$, and $V_0 = \{v\in V \mid \operatorname{div} v = 0\}$, Oseen's problem can be written as our saddle point problem.

We can try to apply Brezzi's theorem. I will omit the proof that Oseens problem satisfies Brezzis theorem.

5 Multigrid Preconditioning

Consider

$$-u'' = f, u(0) = u(1) = 0.$$

The variational form of this is

$$a(u,v) := \int_0^1 u'v'\,dx = (f,v), \forall v\in H_0^1(\Omega).$$

Setting $V_h = \operatorname{span}\{\phi_1,\dots,\phi_N\}$ to be the C^0-P^1 elements for $[0,1]$, we consider $u_h = \sum_{i=1}^N u_i\phi_i$, and testing with $v=\phi_j$, we can obtain the system $Au=b$, A is tridiagonal with -1s on the super diagonals and 2 on the diagonal, $b=(b_i)_{i=1}^N$, $B_i = h^2 f(x_i)$. We know the eigenvalues and eigenvectors corresponding to A are $\lambda_k = 4\sin^2(\theta_k/2), \xi_k = \sqrt{2}\begin{bmatrix}\sin(\theta_k) & \sin(2\theta_k) & \dots & \sin(N\theta_k)\end{bmatrix}^T$, $\theta_k = \frac{k\pi}{N+1}$. We can extend ξ_k to ξ^k, where ξ_k is 0 on all other nodes other than x_k. This forms a basis for V_h. We call all eigenvectors corresponding to k such that $\frac{N+1}{2} < k \le N$, high frequency, corresponding to $\theta_k\in[\pi/2,\pi]$. For simple iterative methods, the high frequency modes are damped significantly, while low frequency modes take a while to fully damp.

Because of this phenomenon, it takes a long time for the computation to fully run. Consider the Richardson iteration scheme for solving $Au=b$:

$$u^{k+1} = u^k + \omega(b-Au^k), \quad \omega\in\left(0,\frac{2}{\lambda_{\max}(A)}\right).$$

Adding u to both sides and subtracting, we can see that

$$u-u^{k+1} = (I-\omega A)(u-u^k) \Longrightarrow e_{k+1} = (I-\omega A)e_k \Longrightarrow e_{k+1} = (I-\omega A)^{k+1}e_0,$$

where e_n is the nth error vector $u-u^n$. Noting that $\{\xi^1,\dots,\xi^N\}$ is a basis for $\mathbb{R}^N$, we have

$$A\xi^k = \lambda_k\xi^k \Longrightarrow (I-\omega A)^m\xi^k = (1-\omega\lambda_k)^m\xi^k.$$

Expanding $e_0 = \sum_{k=1}^N \alpha_k\xi^k$, we can see that

$$e_m = \sum_{k=1}^N \alpha_k(E-\omega A)^m\xi^k = \sum_{k=1}^N \alpha_k(1-\omega\lambda_k)^m\xi^k.$$

We denote that $\rho_k := 1-\omega\lambda_k$ as the damping coefficient of the kth mode, and thus ξ^k is damped by ρ_k^m. As an example, assume that $\omega = 1/4$. Then one can calculate that $\rho_1 = \cos^2\left(\frac{\pi}{2}\frac{1}{N+1}\right) \approx 1-ch^2$, and $\rho_N = \cos^2\left(\frac{\pi}{2}\frac{N}{N+1}\right) \approx \frac{1}{2}$.

Note that ρ_N corresponds to high frequency and ρ_1 corresponds to low frequency modes. Then we can see that it shouldn't take too many iterations for the high frequency nodes to be effectively be damped, since a decay rate of $\frac{1}{2}$ is pretty steep. But we see for ρ_1, as $h \to 0$, ρ_1 is near 1, which will take many iterations to fully damped, so it will take lots of computational power to get close to an error of 0. We can see iterative methods are effective at decaying the higher frequency modes, but we are in need of another technique to get rid of the lower frequency nodes. This leads us to the idea of the multigrid preconditioner.

5.1 V-Cycle Multigrid

Let V be a Hilbert space. Consider the problem: Find $u \in V$ such that $a(u,v) = \langle f, v\rangle$, $\forall v \in V$, where f is continuous on V and a is symmetric and coercive on V. Denote $(\cdot,\cdot)$ as another inner product on V.

Let $V_1 \subset V_2 \subset \ldots \subset V_J = V_h$ be nested subsets of finite dimensional spaces. Then the problem is now to solve: Find $u_J \in V_J$ such that $a(u_J, v) = (f_J, v), \forall v \in V_J$, where f_J is the projection of f onto V_J, i.e. $(f_J, v_h) = (f, v_h), \forall v_h \in V_h$.

We define the following operators:

$$A_j : V_j \to V_j, (A_j u, v) = a(u,v), \forall u, v \in V_i, i = 1, 2, \ldots, J,$$
$$Q_j : V_J \to v_j, (Q_j f, v) = (f, v), \forall f \in V_J, \forall v \in V_j, j = 1, 2, \ldots, J.$$