MEK4250

Exam preperation for Finite Elements in Computational Mechanics

August Femtehjell

Spring 2025

Contents

1 The finite element method **4**

1.1 Short-form answer 4

1.1.1 Ciarlet's definition 4

1.1.2 Functionals and function spaces 4

1.1.3 Reference element 5

1.1.4 Common elements in common spaces 6

1.2 Long-form answer 8

1.2.1 Ciarlet's definition of a finite element 8

1.2.2 Concept of functionals and function spaces 8

1.2.3 How DOFs define our space 9

2 Weak formulation and finite element error estimation **10**

2.1 Short-form answer 10

2.2 Long-form answer 13

2.2.1 Weak form of the Poisson equation 13

2.2.2 Lax–Milgram's theorem 14

2.2.3 Extension to convection-diffusion equation 15

2.2.4 Cea's lemma 16

2.2.5 Convergence rates 17

3 Discretization of Convection-Diffusion **18**

3.1 Short-form answer 18

3.1.1 Oscillations 19

3.2 Long-form answer 22

3.2.1 Weak form of the Convection-Diffusion equation 22

3.2.2 Well-posedness 23

3.2.3 Oscillations in Galerkin methods 24

3.2.4 Streamline diffusion/Petrov–Galerkin 25
3.2.5 Cea's lemma 26

4 Discretization of Stokes **27**
4.1 Short-form answer 27
4.2 Long-form answer 30
4.2.1 Weak form of Stokes problem 30
4.2.2 Brezzi conditions 32
4.2.3 Oscillations in the pressure 33
4.2.4 inf-sup and approximation 33
4.2.5 Stability without inf-sup 34

5 Discretization of Navier–Stokes **36**
5.1 Short-form answer 36
5.2 Long-form answer 37
5.2.1 Operator splitting 37
5.2.2 Issues with operator splitting 40
5.2.3 Algebraic splitting 40

6 Other formulations **41**
6.1 Short-form answer 41

Abstract

This document contains my preperation for the final oral exam for the course MEK4250–Finite Elements in Computational Mechanics, taught at the University of Oslo in the spring of 2025. The code for everything, as well as this document, can be found at my GitHub repository: https://github.com/augustfe/MEK4250.

Exam formalities

Six problems/topics are given for this exam. For each problem, the candidate must prepare a 20 minutes oral presentation. Try to communicate a good overview and understanding of the topic, but compose the talk so that you can demonstrate knowledge about details too. The student is expected to be able to stick to one subject for the 30 minutes for top grades. There are no aids besides a whiteboard and this document with the exam problems (experience with this type of exam and various aids tells that learning the content by heart gives by far the best delivery that demonstrates solid understanding).

We will throw a die and the number of eyes determines the topic to be presented. After your presentation, you will be given some questions, either about parts of your presentation or facts from the other topics. After each presentation, the next candidate can throw the die and thereby get about 10 minutes to collect the thoughts before presenting the assigned topic.

1 The finite element method

Explain Ciarlet's definition of a finite element. Explain the concept of functionals and function spaces. How are degrees of freedom used to ensure that the finite element spaces are part of certain function spcaes? Show that a finite element may conveniently be defined in terms of a reference element. List common elements in common spaces.

1.1 Short-form answer

1.1.1 Ciarlet's definition

Ciarlet defined a finite element as a triplet (T, V, D). Here, T is a bounded domain in $\mathbb{R}^d$, which is typically a polyhedron. This corresponds to a section of the triangulation of the domain. Next, $V = \{\psi_i\}_{i=1}^n$ is a set of linearly independent basis functions defined on T. Finally, $D = \{d_i\}_{i=1}^n$ is a set of degrees of freedom, which are linear functionals defined on V.

We typically want each domain T to be shape regular, i.e. such that they are all triangles or all quadrilaterals. This allows us to perform the computations in a more efficient manner, where we are able to reuse a lot of the computations for each element. Note that the basis functions ψ_i are only defined locally on the element T, and thus they are not defined globally.

The degrees of freedom d_i are what tie the local basis functions together, ensuring properties such as continuity of a certain order across the boundaries of the elements. With the degrees of freedom, we have the associated *nodal basis* $\{\phi_i\}_{i=1}^n$, which are defined such that

$$d_j(\phi_i) = \delta_{ij}. \tag{1.1}$$

Finite elements are typically implemented directly through this nodal basis, where the coefficients of the basis functions are defined by the degrees of freedom.

1.1.2 Functionals and function spaces

Function spaces are vector spaces, where each element is a function. Typically, we are looking at function spaces defined through various properties, such as the space of all continuous functions, or the space of all functions with a finite integral. A functional on a vector space V is then an object $L(\cdot)$, which takes in a vector $v \in V$ and returns a number. For instance, from a set of coordinates $x_1, x_2, \ldots, x_n$, we can define a functional L on a function space $V = \mathrm{span}\{\phi_i\}_{i=1}^n$ such that

$$L(v) = v(x_1) + v(x_2) + \cdots + v(x_n), \qquad v \in V \tag{1.2}$$

or a set of functionals L_j such that

$$L_j(\phi_i) = \phi_i(x_j) = \delta_{ij}, \qquad i = 1, 2, \ldots, n. \tag{1.3}$$

Typical function spaces we are interested in the finite element method are Sobolev spaces $W^{k,p}(\Omega)$, which are defined as

$$W^{k,p}(\Omega) = \left\{ u : \left(\int_\Omega \sum_{i \le k} \left| \frac{\partial^i u}{\partial x^i} \right|^p \mathrm{d}x \right)^{1/p} < \infty \right\}, \tag{1.4}$$

essentially spaces where the function and its derivatives up to order k are in $L^p(\Omega)$.

As far as I'm aware, we are really only interested in the spaces where $p = 2$, which we denote as $H^k(\Omega)$. The H is chosen after Hilbert, as these are in fact Hilbert spaces. The inner product for $H^k(\Omega)$ is defined as

$$(u, v)_{H^k(\Omega)} = (u, v)_k = \sum_{i \le k} \int_\Omega \frac{\partial^i u}{\partial x^i} \frac{\partial^i v}{\partial x^i} \, \mathrm{d}x. \tag{1.5}$$

With the functionals we defined previously, if we choose the points x_i to be along the edges of the elements, we can ensure that they are continuous across elements, which then gives us that the finite element space is a subspace of $H^1(\Omega)$. Similarly, we can choose a different set of functionals such that we ensure higher order continuity, such as $H^2(\Omega)$.

1.1.3 Reference element

In order to illustrate the benefit of a reference element, we consider simply a number of segments of the real number line, with points $x_1 < x_2 < \cdots < x_n$. Thus, each T_i is simply the segment $[x_i, x_{i+1}]$ for $i = 1, 2, \ldots, n-1$. On each segment, we have the linear Lagrange basis functions ϕ_i, which are defined such that $\phi_i(x_j) = \delta_{ij}$. This is illustrated with $n = 5$ points in Fig. 1.

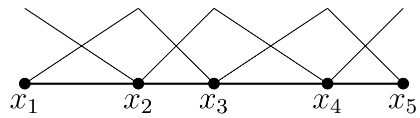

Figure 1: A number of segments of the real number line, with points $x_1 < x_2 < \cdots < x_n$, defining basis functions ϕ_i on each segment.

When solving the finite element problem, we typically need to compute matrices A and M defined by

$$A_{ij} = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, \mathrm{d}x, \qquad M_{ij} = \int_\Omega \phi_i \phi_j \, \mathrm{d}x, \tag{1.6}$$

where A is the stiffness matrix and M is the mass matrix. With the basis functions we have here, these matrices will be incredibly sparse, as for instance ϕ_i is only non-zero on the segment $[x_{i-1}, x_{i+1}]$. As such, we can beforehand say that for instance $A_{1,5}$ will be zero.

We then only need to compute the integrals for the non-zero entries, and we can additionally do this just on each segment, rather than the entire domain. We are then left with the integrals of the form

$$A_{ij} = \int_\Omega \phi_i'(x)\phi_j'(x)\,\mathrm{d}x = \int_{x_i}^{x_{i+1}} \phi_i'(x)\phi_j'(x)\,\mathrm{d}x \tag{1.7}$$

for $j = i, i+1$. As of now, we need to compute this for each i, however with a change of basis we can instead compute it just once for each segment, getting

$$A_{j,k}^{(i)} = \int_0^1 \ell_j'(X)\frac{\mathrm{d}X}{\mathrm{d}x}\ell_k'(X)\frac{\mathrm{d}X}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}X}\,\mathrm{d}X, \tag{1.8}$$

where X is the reference element, and dX/dx is the Jacobian of the transformation from the reference element to the segment. In the linear case, $\frac{dX}{dx}$ is simply the length of the segment, and we can pull it out of the integral. Constructing the total stiffness matrix then becomes a matter of computing the integral once for the reference element, and then adding the scaled contributions for each segment. A similar argument holds for higher order spaces, mapping the physical element to the reference element, however the Jacobian will be more complicated, especially in the case of curved elements.

1.1.4 Common elements in common spaces

As mentioned previously, we are mostly interested in the spaces L^2, H^1 and H^2, as well as the spaces $H(\mathrm{div})$ and $H(\mathrm{curl})$. Let's firstly describe the new spaces $H(\mathrm{div})$ and $H(\mathrm{curl})$. They are given by

$$\begin{aligned} H(\mathrm{div}) &= \left\{u \in L^2(\Omega)^d : \nabla\cdot u \in L^2(\Omega)\right\}, \\ H(\mathrm{curl}) &= \left\{u \in L^2(\Omega)^d : \nabla\times u \in L^2(\Omega)\right\}. \end{aligned} \tag{1.9}$$

We are typically working with polynomial spaces, which are C^∞ within their domain. If we allow for discontinuities across the boundaries, we are in L^2. If we enforce continuity across the boundaries, we are in H^1, and are typically using Lagrange elements. If we enforce continuity of the first derivatives as well, we are in H^2, and are typically using Hermite elements. For the spaces $H(\mathrm{div})$ and $H(\mathrm{curl})$, we typically use Raviart-Thomas elements and Nedelec elements respectively.

Raviart-Thomas elements ensure continuity of the normal component across boundaries, while jumps in the tangential component are allowed. Similarly, as

Stokes' Theorem relates curl to the tangential component, Nedelec elements ensure continuity of the tangential component across boundaries, while jumps in the normal component are allowed.

1.2 Long-form answer

1.2.1 Ciarlet's definition of a finite element

Ciarlet defines a finite element by a triplet (T, V, D), where

- T is a bounded domain in R^d, most typically a polyhedron.
- $V = \{\psi_i\}_{i=1}^n$ is a set of linearly independent basis functions on T
- $D = \{d_i\}_{i=1}^n$ is a set of (linearly independent) degrees of freedom defined in terms of linear functionals on V. (We remark for $v \in V$ we may evaluate $d_i(v)$ since d_i is a linear functional on V.)

T defines the triangulation, or cells, in our domain. The basis functions are then just defined on their cell T. The magic happens when we introduce D, dubbed the degrees of freedom, or dofs for short. They are in a sense how we "tie" the function spaces together, as they are originally just defined locally.

Most elements are implemented through a nodal basis, defined by the set of basis functions $\{\phi_i\}_{i=0}^n$ satsifying $d_j(\phi_i) = \delta_{ij}$. The simplest example is the Lagrange element, where for a basis function L_j, the dofs are defined by

$$d_i(L_j) = L_j(x_i) = \delta_{ij}. \tag{1.10}$$

Initially, the set $\{x_i\}$ consists of the nodes of each triangle.

Say we have two triangles, each with the vertices $\{x_1, x_2, x_3\}$ and $\{x_4, x_5, x_6\}$ respectively. If the triangles are next to eachother, we would perhaps then have that $x_2 = x_4$ and $x_3 = x_5$. However, we still av the set of dofs $\{d_i\}_{i=1}^6$, even though we only have four unique points. This results in discontinuous Lagrange elements, as they do not directly communicate across the boundaries. If we however say that $d_2 = d_4$ and $d_3 = d_5$, we would have continuous elements.

1.2.2 Concept of functionals and function spaces

A function space is a vector space where each element is a function. Typically, the function spaces are defined by some properties, for instance the space of all continuous functions, all functions with a finite integral, and especially in our case all functions where the function as well as the derivatives have finite integrals. This essentially forms our "main" space $H^1(\Omega)$, defined by

$$H^1(\Omega) = \left\{ f : \int_\Omega |f|^2 + |\nabla f|^2 \,\mathrm{d}x < \infty \right\}. \tag{1.11}$$

A functional is then simply something which takes in a vector, for instance from a function space, and returns a number, either real or complex (although typically real in our case).

1.2.3 How DOFs define our space

Problem description

Formulate a finite element method for the Poisson problem with a variable coefficient $\kappa : \Omega \to \mathbb{R}^{d\times d}$. Assume that κ is positive and symmetric. Show that Lax–Milgram's theorem is satisfied. Consider extensions to e.g. convection-diffusion equation and the elasticity equation. Derive *a priori* error estimates in terms of Cea's lemma for the finite element method in the energy norm. Describe how to perform an estimation of convergence rates.

2.1 Short-form answer

The Poisson problem with a variable coefficient is given by

$$\begin{aligned} -\nabla\cdot(\kappa\nabla u) &= f \qquad && \text{in } \Omega, \\ u &= g && \text{on } \partial\Omega_D, \\ \kappa\frac{\partial u}{\partial n} &= h && \text{on } \partial\Omega_N, \end{aligned} \tag{2.1}$$

where we assume that κ is positive, symmetric and bounded.

In order to get the weak form of this problem, we multiply by a test function v and integrate by parts, which yields

$$\begin{aligned} \int_\Omega -\nabla\cdot(\kappa\nabla u)\, v\,\mathrm{d}x &= \int_\Omega f\, v\,\mathrm{d}x \\ \int_\Omega \kappa\nabla u\cdot\nabla v\,\mathrm{d}x &= \int_\Omega f\, v\,\mathrm{d}x + \int_{\partial\Omega}\kappa\frac{\partial u}{\partial n}\, v\,\mathrm{d}s. \end{aligned}$$

We can split the boundary integral into two parts, one for the Dirichlet boundary and one for the Neumann boundary:

$$\begin{aligned} \int_{\partial\Omega}\kappa\frac{\partial u}{\partial n}\, v\,\mathrm{d}s &= \int_{\partial\Omega_D}\kappa\frac{\partial u}{\partial n}\, v\,\mathrm{d}s + \int_{\partial\Omega_N} h\, v\,\mathrm{d}s \\ &= \int_{\partial\Omega_D}\kappa\frac{\partial u}{\partial n}\, 0\,\mathrm{d}s + \int_{\partial\Omega_N} h\, v\,\mathrm{d}s. \end{aligned}$$

As the solution is known along the Dirichlet boundary, we can choose $v \in H^1_{0,D}(\Omega)$, which makes the boundary integral there vanish. This leaves us with the weak form of the Poisson problem, which amounts to finding $u \in H^1_{g,D}(\Omega)$ such that

$$\int_\Omega \kappa\nabla u\cdot\nabla v\,\mathrm{d}x = \int_\Omega f\, v\,\mathrm{d}x + \int_{\partial\Omega_N} h\, v\,\mathrm{d}s \qquad \forall v \in H^1_{0,D}(\Omega). \tag{2.2}$$

We then get the finite element formulation by discretizing the function space.

For Lax–Milgram to hold, we need to show that:

$$a(u, v) \leq C_1 \|u\|_V \|v\|_V, \qquad \forall u, v \in V, \tag{2.3}$$

$$a(u, u) \geq C_2 \|u\|_V^2, \qquad \forall u \in V, \tag{2.4}$$

$$L(v) \leq C_3 \|v\|_V, \qquad \forall v \in V, \tag{2.5}$$

where $a(u, v)$ is the bilinear form in the weak form, and $L(v)$ is the rhs.

We firstly have

$$a(u, v) = (\kappa \nabla u, \nabla v)_{L^2} \leq \kappa_{\max} |u|_1 |v|_1 \leq \kappa_{\max} \|u\|_1 \|v\|_1, \tag{2.6}$$

where we've used the fact that κ is bounded and the Cauchy–Schwarz inequality.

Next, we have

$$a(u, u) = (\kappa \nabla u, \nabla u)_{L^2} \geq \kappa_{\min} |u|_1^2 \geq C_2 \|u\|_1^2, \tag{2.7}$$

where we've used the fact that κ is bounded below and positive. We've also used the fact that the H^1-norm is equivalent to the H^1-semi-norm on H_0^1, applying lifting if necessary.

Finally, we have

$$\begin{aligned} L(v) &= \int_\Omega f\, v \,\mathrm{d}x + \int_{\partial\Omega_N} h\, v \,\mathrm{d}s \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega_N)} \|v\|_{L^2(\partial\Omega_N)} \\ &\leq C_3 \|v\|_V, \end{aligned}$$

assuming we can bound $\|v\|_{L^2(\partial\Omega_N)}$ by $\|v\|_1$, and that f and h are bounded.

Extending the equation to the convection-diffusion equation, we simply add the convection term

$$w \cdot \nabla u = 0 \tag{2.8}$$

to the strong form of the equation, which alters the weak form to

$$a(u, v) = \int_\Omega \kappa \nabla u \cdot \nabla v \,\mathrm{d}x + \int_\Omega w \cdot \nabla u\, v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x + \int_{\partial\Omega_N} h\, v \,\mathrm{d}s, \tag{2.9}$$

where we've made the assumption that w is bounded.

With the energy norm defined as $\|u\|_E^2 = a(u, u)$, we have

$$\begin{aligned} \|u - u_h\|_E^2 &= a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v + v - u_h) \\ &= a(u - u_h, u - v) + a(u - u_h, v - u_h) \\ &= a(u - u_h, u - v) + 0 \\ &\leq \|u - u_h\|_E \|u - v\|_E \end{aligned}$$

which gives us that

$$\|u - u_h\|_E \leq \|u - v\|_E, \tag{2.10}$$

where the choice of $v \in V$ is arbitrary. Here, we've used Galerkin orthogonality, which states that the error is orthogonal to the finite element space. Next, choosing v to be the polynomial interpolant of u of degree t, we have

$$\|u - I_h u\|_E^2 = a(u - I_h u, u - I_h u) \leq \frac{k_1}{1 + C_P}\|u - I_h u\|_1^2 \leq \frac{k_1}{1 + C_P}(Bh^t)^2\|u\|_{t+1}^2, \tag{2.11}$$

which yields the error estimate

$$\|u - u_h\|_E \leq Ch^t\|u\|_{t+1}. \tag{2.12}$$

With this, we can estimate the convergence rate with elements of a given degree, simply by changing the mesh size h, and comparing the error repeatedly.

2.2 Long-form answer

2.2.1 Weak form of the Poisson equation

The Poisson problem with a variable coefficient κ is given by

$$\begin{aligned} -\nabla\cdot(\kappa\nabla u) &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega_D, \\ \kappa\frac{\partial u}{\partial n} &= h \quad \text{on } \partial\Omega_N, \end{aligned} \tag{2.13}$$

with $\partial\Omega_D$ and $\partial\Omega_N$ disjoint parts of the boundary $\partial\Omega$. Here, $\partial\Omega_D$ denotes the Dirichlet boundary, while $\partial\Omega_N$ denotes the Neumann boundary.

Setting up the weak formulation roughly follows the following steps:

1. Multiply with a test function v and integrate over the domain Ω
2. Integrate by parts, and apply Green's lemma.
3. Apply the boundary conditions.

Multiplying with a test function v and integrating over the domain Ω gives us

$$\int_\Omega -\nabla\cdot(\kappa\nabla u)\,v\ \mathrm{d}x = \int_\Omega f\,v\ \mathrm{d}x. \tag{2.14}$$

This is however not ideal, as we are now required to have $u \in H^2(\Omega)$, which is not ideal. We therefore apply Green's lemma to the left-hand side, which gives us

$$\int_\Omega -\nabla\cdot(\kappa\nabla u)\,v\ \mathrm{d}x = \int_\Omega \kappa\nabla u\cdot\nabla v\ \mathrm{d}x - \int_{\partial\Omega}\kappa\frac{\partial u}{\partial n}\,v\ \mathrm{d}s. \tag{2.15}$$

This eases the requirements on u, as we now only require $u \in H^1(\Omega)$, while strengthening the requirements on v to $v \in H^1(\Omega)$.

We can now apply the boundary conditions. Splitting the boundary integral into two parts, we have

$$\int_{\partial\Omega}\kappa\frac{\partial u}{\partial n}\,v\ \mathrm{d}s = \int_{\partial\Omega_D}\kappa\frac{\partial u}{\partial n}\,v\ \mathrm{d}s + \int_{\partial\Omega_N}\kappa\frac{\partial u}{\partial n}\,v\ \mathrm{d}s. \tag{2.16}$$

As we have a section of Dirichlet boundary, we need not solve for u here, as we know the value of u on this section. We may therefore set $v = 0$ on $\partial\Omega_D$ by having $v \in H^1_0(\Omega)$, which gives us

$$\int_{\partial\Omega_D}\kappa\frac{\partial u}{\partial n}\,v\ \mathrm{d}s + \int_{\partial\Omega_N}\kappa\frac{\partial u}{\partial n}\,v\ \mathrm{d}s = \int_{\partial\Omega_N} h\,v\ \mathrm{d}s. \tag{2.17}$$

This gives us the weak formulation for the Poisson problem

$$\int_\Omega \kappa\nabla u\cdot\nabla v\ \mathrm{d}x = \int_\Omega f\,v\ \mathrm{d}x + \int_{\partial\Omega_N} h\,v\ \mathrm{d}s. \tag{2.18}$$

2.2.2 Lax–Milgram's theorem

Lax–Milgram's theorem states:

Theorem 1. *Let V be a Hilbert space, $a(\cdot,\cdot)$ be a bilinear form, $L(\cdot)$ be a linear form, and let the following three conditions be satisfied:*

1. *$a(u,u) \geq \alpha \|u\|_V^2$ for all $u \in V$, where $\alpha > 0$ is a constant.*
2. *$a(u,v) \leq C\|u\|_V \|v\|_V$ for all $u, v \in V$, where $C > 0$ is a constant.*
3. *$L(v) \leq D\|v\|_V$ for all $v \in V$, where $D > 0$ is a constant.*

Then, the problem of finding $u \in V$ such that

$$a(u,v) = L(v) \quad \forall v \in V \tag{2.19}$$

is well-posed in the sense that there exists a unique solution with the stability condition

$$\|u\|_V \leq \frac{C}{\alpha}\|L\|_{V^*}. \tag{2.20}$$

We can now show that Lax–Milgram's theorem is satisfied for the weak formulation of the Poisson problem. We have the bilinear form

$$a(u,v) = \int_\Omega \kappa \nabla u \cdot \nabla v \ \mathrm{d}x, \tag{2.21}$$

and the linear form

$$L(v) = \int_\Omega f\, v \ \mathrm{d}x + \int_{\partial\Omega_N} h\, v \ \mathrm{d}s. \tag{2.22}$$

We can now show that the three conditions of Lax–Milgram's theorem are satisfied.

Firstly we have that

$$\begin{aligned} a(u,u) &= \int_\Omega \kappa \nabla u \cdot \nabla u \ \mathrm{d}x = \int_\Omega (\nabla u)^T \kappa^T \nabla u \ \mathrm{d}x \\ &= \int_\Omega (\nabla u)^T \kappa \nabla u \ \mathrm{d}x \geq \int_\Omega k_0 |\nabla u|^2 \ \mathrm{d}x \\ &= k_0 |u|_1^2 \geq \alpha \|u\|_1^2 \end{aligned}$$

where we've first used the fact that κ is symmetric, and then that κ is positive, and finally that the H_0^1 semi-norm is equivalent to the H_0^1 norm.

For the second point, we firstly show that $a(\cdot,\cdot)$ defines an inner product, assuming that κ is bounded. The inequality then follows simply from the Cauchy–Schwarz inequality. In order to show that $a(\cdot,\cdot)$ is symmetric, we simply have

$$a(u,v) = \int_\Omega \kappa \nabla u \cdot \nabla v \ \mathrm{d}x = \int_\Omega (\nabla v)^T \kappa \nabla u \ \mathrm{d}x = a(v,u). \tag{2.23}$$

We can now show that $a(u, v) \leq C\|u\|_V \|v\|_V$. As κ is bounded, we have

$$\kappa\xi \cdot \xi \leq k_1|\xi|^2 \tag{2.24}$$

for all $\xi \in \mathbb{R}^d$. This gives us that

$$a(u, u) = \int_\Omega \kappa \nabla u \cdot \nabla u \, \mathrm{d}x \leq k_1 \int_\Omega |\nabla u|^2 \, \mathrm{d}x \leq k_1 |u|_1^2. \tag{2.25}$$

We then have that

$$\begin{aligned} a(u, v)^2 \leq a(u, u)a(v, v) = k_1^2 |u|_1^2 |v|_1^2 \\ a(u, v) \leq k_1 |u|_1 |v|_1 \end{aligned}$$

Additionally, for H_0^1 we have

$$\|u\|_1^2 = \|u\|_{L^2}^2 + |u|_1^2 \leq C|u|_1^2 + |u|_1^2 = (1 + C)|u|_1^2$$

by Poincaré's lemma, which gives us

$$a(u, v) \leq k_1 |u|_1 |v|_1 \leq \frac{k_1}{1 + C} \|u\|_1 \|v\|_1. \tag{2.26}$$

Lastly, we have

$$\begin{aligned} L(v) = \int_\Omega f \, v \, \mathrm{d}x + \int_{\partial\Omega_N} h \, v \, \mathrm{d}s &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega_N)} \|v\|_{L^2(\partial\Omega_N)} \\ &\leq (\|f\|_{L^2}(\Omega) + \|h\|_{L^2(\partial\Omega_N)}) \|v\|_{L^2(\Omega)} \leq D\|v\|_{L^2(\Omega)} \leq D\|v\|_1, \end{aligned}$$

showing that the third condition is satisfied as well. Here, we assume that $f \in L^2(\Omega)$ and $h \in L^2(\partial\Omega_N)$, which is a reasonable assumption.

2.2.3 Extension to convection-diffusion equation

The convection-diffusion equation is given by

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) + w \cdot \nabla u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega_D, \\ \kappa \frac{\partial u}{\partial n} &= h \quad \text{on } \partial\Omega_N, \end{aligned} \tag{2.27}$$

where w is the convection term.

To be continued

2.2.4 Cea's lemma

Cea's lemma states that if u_h is the finite element solution, and u is the exact solution, then

$$\|u - u_h\|_V \leq \frac{CB}{\alpha} h^t \|u\|_{t+1}, \tag{2.28}$$

where B is the constant derived from the polynomial approximation, h is a measure of the mesh size, and C and α are constants from Lax–Milgram's theorem.

Here, we assume that the energy norm is given by

$$\|w\|_E = a(w, w)^{1/2}, \tag{2.29}$$

and that we should find an error estimate in this norm. We then have

$$\begin{aligned}
\|u - u_h\|_E^2 &= a(u - u_h, u - u_h) \\
&= a(u - u_h, u - v + v - u_h) \qquad v \in V \\
&= a(u - u_h, u - v) + \underbrace{a(u - u_h, v - u_h)}_{0 \text{ as } v - u_h \in V} \\
&= a(u - u_h, u - v) \\
&\leq \|u - u_h\|_E \|u - v\|_E.
\end{aligned}$$

Dividing by $\|u - u_h\|_E$ gives us

$$\|u - u_h\|_E \leq \|u - v\|_E. \tag{2.30}$$

We then further have, choosing $v = I_h u$ as the polynomial approximation of order t,

$$\begin{aligned}
\|u - I_h u\|_E^2 &= a(u - I_h u, u - I_h u) \\
&\leq \frac{k_1}{1 + C} \|u - I_h u\|_1^2 \\
&\leq \frac{k_1}{1 + C} (Bh^t)^2 \|u\|_{t+1}^2.
\end{aligned}$$

This finally gives us that

$$\|u - u_h\|_E \leq \sqrt{\frac{k_1}{1 + C}} Bh^t \|u\|_{t+1}. \tag{2.31}$$

I believe this is what the question is asking for, but I am not sure.

2.2.5 Convergence rates

We can estimate the convergence rates by looking at the error estimates for varying mesh sizes h. Considering the error estimates for $h = h_1$ and $h = h_2$, we have

$$\|u - u_{h_1}\|_E \leq Ch_1^t \|u\|_{t+1} \quad \text{and} \quad \|u - u_{h_2}\|_E \leq Ch_2^t \|u\|_{t+1}.$$

This gives

$$\begin{aligned} \frac{\|u - u_{h_2}\|_E}{\|u - u_{h_1}\|_E} &\leq \frac{Ch_2^t \|u\|_{t+1}}{Ch_1^t \|u\|_{t+1}} \\ &= \left(\frac{h_2}{h_1}\right)^t. \end{aligned}$$

Taking the logarithm of both sides gives us

$$\begin{aligned} \log\left(\frac{\|u - u_{h_2}\|_E}{\|u - u_{h_1}\|_E}\right) &\leq t \log\left(\frac{h_2}{h_1}\right) \\ \frac{\log\left(\frac{\|u - u_{h_2}\|_E}{\|u - u_{h_1}\|_E}\right)}{\log\left(\frac{h_2}{h_1}\right)} &\leq t. \end{aligned}$$

If we solve for varying mesh sizes $\{h_i\}_{i=1}^n$, keeping h_{i+1}/h_i constant, we get a series of lower bounds for the convergence rate t.

3 Discretization of Convection-Diffusion

Derive a proper variational formulation of the convection-diffusion problem. Derive sufficient conditions that make the problem well-posed. Discuss why oscillations appear for standard Galerkin methods and show how SUPG methods resolve these problems. Discuss also approximation properties in light of Cea's lemma.

3.1 Short-form answer

We consider the form of the convection-diffusion equation given by

$$\begin{aligned} -\mu\Delta u + w\cdot\nabla u &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

where μ is the diffusion coefficient, w is the velocity, and f is the source term. Here, I'm going to be working with $g = 0$, in order to simplify the computations. This is a valid assumption, as we can apply lifting.

Multiplying by a test function v and integrating by parts yields on the lhs

$$\int_\Omega -\mu\Delta u\, v + w\cdot\nabla u\, v\,\mathrm{d}x = \int_\Omega \mu\nabla u\cdot\nabla v + w\cdot\nabla u\, v\,\mathrm{d}x - \int_{\partial\Omega}\mu\frac{\partial u}{\partial n} v\,\mathrm{d}s,$$

which yields the weak form

$$\int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega w\cdot\nabla u\, v\,\mathrm{d}x = \int_\Omega f\, v\,\mathrm{d}x + \int_{\partial\Omega}\mu\frac{\partial u}{\partial n} v\,\mathrm{d}s. \tag{3.2}$$

As u is known along the boundary, we don't have to solve for u here, and we can choose $v \in H_0^1(\Omega)$, which makes the boundary integral vanish. We are then left with the problem of finding $u \in H_0^1(\Omega)$ such that

$$a(u,v) = L(v) \qquad \forall v \in H_0^1(\Omega), \tag{3.3}$$

where

$$a(u,v) = \int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega (w\cdot\nabla u)\, v\,\mathrm{d}x \quad \text{and} \quad L(v) = \int_\Omega f\, v\,\mathrm{d}x. \tag{3.4}$$

In order to look closer at when Lax–Milgram holds, we split the bilinear form into two parts b and c_w, by

$$a(u,v) = b(u,v) + c_w(u,v), \tag{3.5}$$

where

$$b(u,v) = \int_\Omega \mu\nabla u\cdot\nabla v \quad \text{and} \quad c_w(u,v) = \int_\Omega (w\cdot\nabla u)\, v. \tag{3.6}$$

We then have that

$$
\begin{aligned}
c_w(u,v) &= \int_\Omega (w\cdot\nabla u)\, v\,\mathrm{d}x \\
&= -\int_\Omega (w\cdot\nabla v)\, u\,\mathrm{d}x - \int_\Omega \nabla\cdot w\, u\, v\,\mathrm{d}x + \int_\Omega u\, v\, w\cdot n\,\mathrm{d}s,
\end{aligned}
$$

where with Dirichlet conditions we can discard the last term. Then, when $\nabla\cdot w = 0$, we have that

$$
c_w(u,v) = -\int_\Omega (w\cdot\nabla v)\, u\,\mathrm{d}x = -c_w(v,u). \tag{3.7}
$$

This means that c_w is skew-symmetric, which results in $c_w(u,u) = 0$, which means that we have

$$
a(u,u) = b(u,u) \geq \mu |u|_1^2, \tag{3.8}
$$

which means that we have coercivity, as $|\cdot|_1$ is equivalent to $\|\cdot\|_1$ on $H_0^1(\Omega)$.

Next, we have

$$
\begin{aligned}
a(u,v) &= \int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega (w\cdot\nabla u)\, v\,\mathrm{d}x \\
&\leq \mu|u|_1|v|_1 + |w|_\infty |u|_1\|v\|_0 \\
&\leq (\mu + |w|_\infty C_\Omega)|u|_1|v|_1.
\end{aligned}
$$

We therefore need $f\in H_0^{-1}$, $\nabla\cdot w = 0$, and $|w|_\infty < \infty$ for Lax–Milgram to hold. This then gives us the stability estimate

$$
|u|_1 \leq \frac{\mu + C_\Omega\|w\|_\infty}{\mu}\|f\|_{-1}. \tag{3.9}
$$

If then $C_\Omega\|w\|_\infty \gg \mu$, the stability constant will be large, and the solution will be unstable.

3.1.1 Oscillations

In order to explain the oscillations, consider the 1D case, with $w = 1$. We then have the problem

$$
\begin{gathered}
-u_x - \mu u_{xx} = 0 \\
u(0) = 0, \quad u(1) = 1.
\end{gathered} \tag{3.10}
$$

Using linear first order Lagrange elements, we seek $u\in H^1_{(0,1)}$ such that

$$
\int_0^1 -u_x v + \mu u_x v_x\,\mathrm{d}x = 0 \tag{3.11}
$$

for all $v \in H^1_{(0,0)}$.

This discretization in 1D is equivalent with the finite difference scheme

$$-\frac{\mu}{h^2}\left[u_{i+1} - 2u_i + u_{i-1}\right] - \frac{w}{2h}\left[u_{i+1} - u_{i-1}\right] = 0, \tag{3.12}$$

with $u_0 = 0$ and $u_N = 1$. In the extreme case where $\mu = 0$, the oscillations appear as each u_{i+1} is coupled with u_{i-1}, but not with u_i. Thus, for odd N, we will have that all even terms u_{2i} will be equal to $u_0 = 0$, and all odd terms u_{2i+1} will be equal to $u_N = 1$. This is the oscillations we observe.

If we however replace the central scheme for the derivative with a forward of backward scheme, we get

$$\frac{w}{h}\left[u_{i+1} - u_i\right] = 0 \quad \text{or} \quad \frac{w}{h}\left[u_i - u_{i-1}\right] = 0, \tag{3.13}$$

which causes the oscillations to disappear, however the approximation is now only first order accurate. We call this the upwind scheme, as it is biased in the direction of the flow.

If we discretize u_x with a central scheme, and then add diffusion with a constant $\epsilon = h/2$, we get

$$\frac{u_{i+1} - u_{i-1}}{2h} + \frac{h}{2}\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = \frac{u_i - u_{i-1}}{h}, \tag{3.14}$$

exactly the upwinding scheme! In these cases, we are then actually solving the problem

$$-(\mu + \epsilon)u_{xx} + wu_x = f. \tag{3.15}$$

The choice of ϵ was crucial in order to stabilize the scheme, however it was chosen rather arbitrarily. In order to avoid this, we turn to the Petrov–Galerkin method, where we allow the test function space to be different from the trial function space. We choose the trial function space to be as before, however in the test space we utilize the basis functions

$$L_j = N_j + \beta h(w \cdot \nabla N_j). \tag{3.16}$$

This leaves us with the matrix A_{ij} defined by

$$\begin{aligned} A_{ij} &= a(N_i, L_j) = a(N_i, N_j + \beta h(w \cdot \nabla N_j)) \\ &= \underbrace{\mu \nabla N_i \cdot \nabla N_j + (w \cdot \nabla N_i)N_j \,\mathrm{d}x}_{\text{standard Galerkin}} \\ &\quad + \underbrace{\beta h \int_\Omega \mu \nabla N_i \cdot \nabla(w \cdot \nabla N_j)\,\mathrm{d}x}_{=0 \text{ for linear elements}} + \underbrace{\beta h \int_\Omega (w \cdot \nabla N_i)(w \cdot \nabla N_j)\,\mathrm{d}x}_{\text{Artificial diffusion in } w\text{-direction}} . \end{aligned}$$

The rhs also changes, as

$$L(L_j) = \int_\Omega f\, L_j \,\mathrm{d}x = \int_\Omega f(N_j + \beta h(w \cdot \nabla N_j))\,\mathrm{d}x, \tag{3.17}$$

which means that we are adding diffusion in a consistent way to the rhs as well.

3.2 Long-form answer

3.2.1 Weak form of the Convection-Diffusion equation

Here, we consider the convection-diffusion equation given by

$$\begin{aligned} -\mu \Delta u + w \cdot \nabla u &= f, \quad \text{in } \Omega, \\ u &= g, \quad \text{on } \partial\Omega, \end{aligned} \tag{3.18}$$

assuming Dirichlet conditions on the whole boundary.

In order to derive the weak form, we follow the steps:

1. Multiply the equation with a test function v and integrate.
2. Integrate by parts, and apply Gauss–Green's lemma.
3. Apply the boundary conditions.

The first step gives us

$$\int_\Omega -\mu \Delta u \, v + w \cdot \nabla u \, v \; \mathrm{d}x = \int_\Omega f \, v \; \mathrm{d}x. \tag{3.19}$$

Then, we use integration by parts on the first term in order to ease the requirement of $u \in H^2(\Omega)$ to $u \in H^1(\Omega)$, while strengthening the requirement on v from $v \in L_2(\Omega)$ to $v \in H^1(\Omega)$. This gives us

$$\int_\Omega \mu \nabla u \cdot \nabla v + w \cdot \nabla u \, v \; \mathrm{d}x = \int_\Omega f \, v \; \mathrm{d}x + \int_{\partial\Omega} \mu \frac{\partial u}{\partial n} v \; \mathrm{d}s. \tag{3.20}$$

Next, we consider the boundary term. As the solution is known on the boundary, we need to solve for u on the boundary, and can therefore choose $v \in H_0^1(\Omega)$, such that

$$\int_{\partial\Omega} \mu \frac{\partial u}{\partial n} v \; \mathrm{d}s = \int_{\partial\Omega} \mu \frac{\partial u}{\partial n} 0 \; \mathrm{d}s = 0, \tag{3.21}$$

effectively removing the boundary term from our formulation.

We can now write the weak form of the convection-diffusion equation as

$$\int_\Omega \mu \nabla u \cdot \nabla v + w \cdot \nabla u \, v \; \mathrm{d}x = \int_\Omega f \, v \; \mathrm{d}x. \tag{3.22}$$

The bilinear form is then given by

$$a(u, v) = \int_\Omega \mu \nabla u \cdot \nabla v + w \cdot \nabla u \, v \; \mathrm{d}x, \tag{3.23}$$

and the linear form is given by

$$L(v) = \int_\Omega f \, v \; \mathrm{d}x. \tag{3.24}$$

The weak form of the problem is then, find $u \in V$ such that

$$a(u, v) = L(v), \quad \forall v \in V. \tag{3.25}$$

3.2.2 Well-posedness

For well-posedness, we rely on the Lax–Milgram theorem, meaning we have to find sufficient conditions such that:

1. $a(u,u) \geq \alpha \|u\|_V^2$ for some $\alpha > 0$ and all $u \in V$.
2. $a(u,v) \leq \beta \|u\|_V \|v\|_V$ for some $\beta > 0$ and all $u, v \in V$.
3. $L(v) \leq D\|v\|_V$ for some $D > 0$ and all $v \in V$.

Here, we'll work through the conditions in reverse order, adding conditions as we go. For the third condition, we simply have by Cauchy–Schwarz

$$L(v) = \int_\Omega f\, v \,\mathrm{d}x \leq \|f\|_{L^2}\|v\|_{L^2} \leq \|f\|_{L^2}\|v\|_1, \tag{3.26}$$

showing that we require that $f \in L^2(\Omega)$ in order to satisfy the third condition. For the second condition, we apply lifting to u, such that we can use Poincaré's inequality. We then have

$$\begin{aligned}
a(u,v) &= \int_\Omega \mu \nabla u \cdot \nabla v + w \cdot \nabla u\, v \,\mathrm{d}x \\
&\leq \mu |u|_1 |v|_1 + \|w\|_{L^\infty} |u|_1 \|v\|_{L^2} \\
&\leq \mu |u|_1 |v|_1 + C\|w\|_{L^\infty} |u|_1 |v|_1 \\
&\leq \left(\mu + C\|w\|_{L^\infty}\right) |u|_1 |v|_1 .
\end{aligned}$$

As we've applied lifting to u, we can assume that $u \in H_0^1(\Omega)$, such that $|u|_1$ is an equivalent norm to $\|u\|_1$.

Finally, for the first condition, we write

$$a(u,v) = b(u,v) + c_w(u,v), \tag{3.27}$$

where $b(u,v) = \int_\Omega \mu \nabla u \cdot \nabla v$ and $c_w(u,v) = \int_\Omega w \cdot \nabla u\, v$. For b, we already have

$$b(u,u) = \int_\Omega \mu (\nabla u)^2 \,\mathrm{d}x = \mu |u|_1^2 \geq \frac{\mu}{1+C}\|u\|_2^2, \tag{3.28}$$

as

$$\|u\|_2^2 = |u|_1^2 + \|u\|_{L^2}^2 \leq (1+C)|u|_1^2. \tag{3.29}$$

c_w is a bit more involved, however we start with integration by parts in order to get

$$\begin{aligned}
c_w(u,v) &= \int_\Omega w \cdot \nabla u\, v \,\mathrm{d}x \\
&= -\int_\Omega w \cdot \nabla v\, u \,\mathrm{d}x - \int_\Omega \nabla \cdot w\, u\, v \,\mathrm{d}x + \int_{\partial\Omega} w \cdot n\, u\, v \,\mathrm{d}s.
\end{aligned}$$

The boundary term vanishes as we've applied lifting, and if we assume that $\nabla \cdot w = 0$ such that we have incompressibility, we are left with

$$c_w(u, v) = \int_\Omega w \cdot \nabla u \, v \, \mathrm{d}x = -\int_\Omega w \cdot \nabla v \, u \, \mathrm{d}x = -c_w(v, u). \tag{3.30}$$

c_w is then skew-symmetric, such that we have

$$c_w(u, u) = -c_w(u, u) = 0. \tag{3.31}$$

The convection-diffusion equation is then well-posed if we assume that w is bounded and incompressible, such that $\nabla \cdot w = 0$.

3.2.3 Oscillations in Galerkin methods

In order to illustrate the oscillations, we consider a simplified scenario in one dimension, where we set $w = -1$. We then have the equation

$$\begin{aligned} -\mu u_{xx} - u_x &= 0, \\ u(0) &= 0, \\ u(1) &= 1. \end{aligned} \tag{3.32}$$

The variational problem is then, find $u \in H_0^1(0, 1)$ such that

$$\int_0^1 \mu u_x v_x - u_x v \, \mathrm{d}x = 0, \quad \forall v \in H_0^1(0, 1). \tag{3.33}$$

Using first order Lagrange elements, we have that the discretization is equivalent to the central finite difference scheme

$$-\frac{\mu}{h^2}\left[u_{i+1} - 2u_i + u_{i-1}\right] - \frac{w}{2h}\left[u_{i+1} - u_{i-1}\right] = 0, \quad i = 1, \ldots, N-1, \tag{3.34}$$

which for $\mu = 0$ reduces to $u_{i+1} = u_{i-1} = \ldots$ and $u_{i+2} = u_i = u_{i-2} = \ldots$. This is the cause of the oscillations, as if N is odd, then we'll have that all terms of the form $u_{2i} = 0$, while $u_{2i+1} = 1$, determined by the boundary conditions.

In order to get rid of these oscillations, we can apply upwinding, which amounts to using the approximations

$$\begin{aligned} \frac{du}{dx}(x_i) &= \frac{1}{h}\left[u_{i+1} - u_i\right] \quad \text{if } w < 0, \\ \frac{du}{dx}(x_i) &= \frac{1}{h}\left[u_i - u_{i-1}\right] \quad \text{if } w > 0. \end{aligned} \tag{3.35}$$

The oscillations will then dissapear, however we are now using a first order scheme, rather than the second order scheme we had with the central finite difference scheme. One way to look at this is by noting that

$$\frac{u_i - u_{i-1}}{h} = \frac{u_{i+1} - u_{i-1}}{2h} + \frac{h}{2} \frac{-u_{i+1} + 2u_i - u_{i-1}}{h^2}, \tag{3.36}$$

as this shows that we are adding a diffusion term with coefficient $\varepsilon = \frac{h}{2}$ to the equation.

This shows that we are then actually solving the problem

$$-(\mu + \varepsilon)u_{xx} - u_x = f, \tag{3.37}$$

as opposed to the original problem.

3.2.4 Streamline diffusion/Petrov–Galerkin

Streamline diffusion/Petrov–Galerkin (SUPG) methods are a more general and ordered way of dealing with the oscillations we saw in the previous section. We then add the diffusion in a consistent way, such that we aren't changing the solution as $h \to 0$.

The Petrov–Galerkin method is maybe unsurprisingly very similar to the standard Galerkin method, given by: Find $u_h \in V_{h,g}$ such that

$$a(u_h, v_h) = L(v_h), \quad \forall v_h \in W_{h,0}, \tag{3.38}$$

where the difference is that the test space is now different from the trial space.

In matrix form, the Galerkin formulation yields

$$A_{ij} = a(N_i, N_j) = \int_\Omega \mu \nabla N_i \cdot \nabla N_j + w \cdot \nabla N_i \, N_j \; \mathrm{d}x, \tag{3.39}$$

while the Petrov–Galerkin formulation yields

$$A_{ij} = a(N_i, L_j) = \int_\Omega \mu \nabla N_i \cdot \nabla L_j + w \cdot \nabla N_i \, L_j \; \mathrm{d}x, \tag{3.40}$$

where L_j is the test function. Choosing the functions L_j carefully is the key to adding diffusion in a consistent way.

We let $L_j = N_j + \beta h(w \cdot \nabla N_j)$. This gives us

$$\begin{aligned}
A_{ij} &= a(N_i, N_j + \beta h(w \cdot \nabla N_j)) \\
&= \int_\Omega \mu \nabla N_i \cdot (N_j + \beta h w \cdot \nabla N_j)\ \mathrm{d}x + \int_\Omega w \cdot \nabla N_i (N_j + \beta h(w \cdot \nabla N_j))\ \mathrm{d}x \\
&= \underbrace{\int_\Omega \mu \nabla N_i \cdot N_j\ \mathrm{d}x + \int_\Omega w \cdot \nabla N_i\, N_j\ \mathrm{d}x}_{\text{Standard Galerkin term}} \\
&\quad + \underbrace{\beta h \int_\Omega \mu \nabla N_i \cdot \nabla(w \cdot \nabla N_j)\ \mathrm{d}x}_{\text{Vanishes for linear elements}} + \underbrace{\beta h \int_\Omega (w \cdot \nabla N_i)(w \cdot \nabla N_j)\ \mathrm{d}x}_{\text{Artificial diffusion in } w \text{ direction}}.
\end{aligned}$$

The right hand side also changes, denoting the linear form now as $b(L_j)$, such that we have

$$b(L_j) = \int_\Omega f L_j\ \mathrm{d}x = \int_\Omega f(N_j + \beta h(w \cdot \nabla N_j))\ \mathrm{d}x. \tag{3.41}$$

We are in other words changing both sides of the equation, such that the artifical diffusion is consistent.

3.2.5 Cea's lemma

Cea's lemma states that, given the conditions for Lax–Milgram are satisfied, we have

$$\|u - u_h\|_V \leq \frac{CB}{\alpha} h^t \|u\|_{t+1}. \tag{3.42}$$

where B comes the polynomial approximation properties, and α and C are the constants from the Lax–Milgram theorem. For convection-dominated problems, $\frac{C}{\alpha}$ is large, which causes poor approximation on coarse grids.

In order to get improved error estimates for the SUPG method, we utilize an alternative norm, given by

$$\|u\|_{sd} = (h\|w \cdot \nabla u\|^2 + \mu|\nabla u|^2)^{1/2}. \tag{3.43}$$

Given that the conditions for Lax–Milgram still hold, solving the SUPG problem on a finite element space of order 1 gives us

$$\|u - u_h\|_{sd} \leq C h^{3/2} \|u\|_2. \tag{3.44}$$

The norm $\|u\|_{sd}$ is called the *streamline diffusion* norm.

The proof for this is very involved, and I'd rather not to too much into detail about it. One thing to note however is that this error bound is independent of the convection velocity, which is a big improvement over the standard Galerkin method.

4 Discretization of Stokes

Derive a proper variational formulation of the Stokes problem. Discuss the four Brezzi conditions that are needed for a well-posed continuous problem. Explain why oscillations might appear in the pressure for some discretization techniques. Present expected approximation properties for mixed elements that satisfy the inf-sup condition, and discuss a few examples like e.g. Taylor–Hood, Mini, and Crouzeix–Raviart. Discuss also how one might cirumvent the inf-sup condition by stabilization.

4.1 Short-form answer

Stokes problem is given by

$$\begin{aligned} -\Delta u + \nabla p &= f && \text{in } \Omega, \\ \nabla \cdot u &= 0 && \text{in } \Omega, \\ u &= g && \text{on } \partial\Omega_D, \\ \frac{\partial u}{\partial n} - p\, n &= h && \text{on } \partial\Omega_N. \end{aligned} \tag{4.1}$$

Here, $u\colon \Omega \to \mathbb{R}^n$ is the fluid field, while $p\colon \Omega \to \mathbb{R}$ is the pressure field. The presence of both the Dirichlet and Neumann boundary conditions leads to a well-posed problem, as long as neither is empty. If we don't have a Dirichlet boundary, then the velocity field is only determined up to a constant, while if the Neumann boundary is empty, the pressure field is only determined up to a constant.

We find the weak form of the Stokes problem by multiplying the first equation by a test function $v \in V$ and integrating. This yields

$$\begin{aligned} \int_\Omega (-\Delta u + \nabla p) \cdot v \,\mathrm{d}x &= \int_\Omega f \cdot v \,\mathrm{d}x, \\ \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Omega p \nabla \cdot v \,\mathrm{d}x &= \int_\Omega f \cdot v \,\mathrm{d}x + \int_{\partial\Omega} \left(\frac{\partial u}{\partial n} - p\, n \right) \cdot v \,\mathrm{d}s. \end{aligned}$$

Now applying the boundary conditions, we can split the boundary integral into two parts, one for the Dirichlet boundary and one for the Neumann boundary:

$$\int_{\partial\Omega} \left(\frac{\partial u}{\partial n} - p\, n \right) \cdot v \,\mathrm{d}s = \int_{\partial\Omega_D} \left(\frac{\partial u}{\partial n} - p\, n \right) \cdot v \,\mathrm{d}s + \int_{\partial\Omega_N} \left(\frac{\partial u}{\partial n} - p\, n \right) \cdot v \,\mathrm{d}s. \tag{4.2}$$

As the solution is known on $\partial\Omega_D$, we can choose $v \in H^1_{0,D}(\Omega)$, such that the first term vanishes. For the second term, we can simply insert the Neumann condition, which finally gives us

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Omega p \nabla \cdot v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x + \int_{\partial\Omega_N} h \cdot v \,\mathrm{d}s. \tag{4.3}$$

For the second equation, we multiply by a test function $q \in Q$ and integrate, which gives us

$$\int_{\Omega} q \nabla \cdot u \, \mathrm{d}x = 0. \tag{4.4}$$

The weak form of the problem is then to find $(u, p) \in H^1_{g,D}(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} a(u, v) + b(p, v) &= f(v) \quad \forall v \in H^1_{0,D}(\Omega), \\ b(q, u) &= 0 \qquad\ \ \forall q \in L^2(\Omega). \end{aligned} \tag{4.5}$$

Note that we have here switched $p := -p$, in order to get symmetry.

For saddle-point problems of the form described, the four Brezzi conditions we need to satisfy is then

1. Continuity of a: For all $u_h, v_h \in V_h$, we have

$$a(u_h, v_h) \leq C_1 \|u_h\|_{V_h} \|v_h\|_{V_h}. \tag{4.6}$$

2. Coercivity of a: There exists a constant $C_2 > 0$ such that

$$a(u_h, u_h) \geq C_2 \|u_h\|^2_{V_h} \tag{4.7}$$

   for all $u_h \in Z_h$, where

$$Z_h = \{u_h \in V_h \mid b(u_h, q_h) = 0 \quad \forall q_h \in Q_h\}.$$

3. Continuity of b: For all $q_h \in Q_h$ and $u_h \in V_h$, we have

$$b(q_h, u_h) \leq C_3 \|q_h\|_{Q_h} \|u_h\|_{V_h}. \tag{4.8}$$

4. "Coercivity" of b: There exists a constant $C_4 > 0$ such that

$$\sup_{u_h \in V_h} \frac{b(q_h, u_h)}{\|u_h\|_{V_h}} \geq C_4 \|u_h\|_1 \quad \forall q_h \in Q_h. \tag{4.9}$$

For Stokes problem, the first three conditions aren't too difficult to check. The last one however is a bit tricky, and typically requires specially designed finite elements in order to satisfy it.

Oscillations appear when the last condition is not satisfied. This happens because there exists some $q_h \in Q_h$ such that $b(q_h, u_h) = 0$ for all $u_h \in V_h$, while $q_h \neq 0$. This means that if we have some solution pressure $p_h \in Q_h$, we can simply add as much of q_h as we want. These spurious pressure modes typically take the form of oscillations.

For mixed elements, we get the expected approximation of

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq Ch^k \|u\|_{k+1} + Dh^{t+1} \|p\|_{t+1}, \tag{4.10}$$

where k is the polynomial degree of the velocity space and t is the polynomial degree of the pressure space.

One popular choice of element is the Taylor–Hood element, which is quadratic in velocity and linear in pressure, while being continuous across element boundaries. In other words, it's a P_2-P_1 element. This gives the expected approximation order of

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq h^2 \left(C\|u\|_3 + D\|p\|_2 \right). \tag{4.11}$$

Another popular choice is the mini element, which is linear in both velocity and pressure, however it has an additional bubble function in the velocity. This bubble function is designed such that it is zero along the element edges, which on the unit triangle means that the basis functions are

$$\{1, x, y, xy(1 - x - y)\}, \tag{4.12}$$

where the last function is the bubble function. This gives the expected approximation order

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq h \left(C\|u\|_2 + Dh\|p\|_2 \right). \tag{4.13}$$

Finally we have the Crouzeix–Raviart element, which is linear in velocity and constant in pressure. Additionally, the velocity is only continuous at the midpoint of each edge, as illustrated in Fig. 2. With this element, we get the expected approximation order

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq h \left(C\|u\|_2 + D\|p\|_1 \right). \tag{4.14}$$

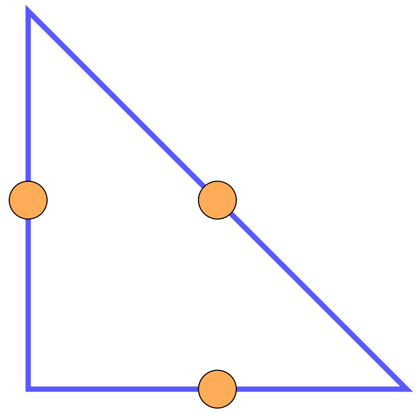

Figure 2: Crouzeix–Raviart element with DOFs marked in orange.

4.2 Long-form answer

4.2.1 Weak form of Stokes problem

The Stokes problem describes the flow of a slowly moving viscous incompressible Newtonian fluid. Let $u : \Omega \to \mathbb{R}^n$ be the fluid field, and $p : \Omega \to \mathbb{R}$ be the fluid pressure. Stokes problem can then be written as

$$\begin{aligned} -\Delta u + \nabla p &= f \quad \text{in } \Omega, \\ \nabla \cdot u &= 0 \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega_D, \\ \frac{\partial u}{\partial n} - p\, n &= h \quad \text{on } \partial\Omega_N. \end{aligned} \tag{4.15}$$

Here, f is the body forcce, $\partial\Omega_D$ and $\partial\Omega_N$ are the Dirichlet and Neumann boundaries, respectively. Additionally, g is the prescribed fluid velocity on the Dirichlet boundary, and h is the surface force or stress on the Neumann boundary.

The presence of the Dirichlet and Neumann boundary conditions lead to a well-posed problem, so long as neither are empty. If the Dirichlet condition if empty, the velocity is only determined up to a constant. If the Neumann condition is empty, the pressure is only determined up to a constant.

We being by setting up the weak form of the Stokes problem. Setting up the weak form amounts to the following steps:

1. Multiply by test functions and integrate of the domain Ω.

2. Integration by parts, and apply Gauss–Green's lemma.

3. Apply the boundary conditions.

This leads us to initially have, multiplying the first equation by a test function v and the second equation by a test function q,

$$\begin{aligned}\int_\Omega (-\Delta u + \nabla p)\cdot v \,\mathrm{d}x &= \int_\Omega f\cdot v \,\mathrm{d}x \\ \int_\Omega (\nabla\cdot u) q \,\mathrm{d}x &= 0.\end{aligned} \tag{4.16}$$

This is however not ideal, as we would currently be requiring that $u \in H^2_{g,D}$ and $p \in H^1$. We therefore apply integration by parts, which yields

$$\int_\Omega (-\Delta u + \nabla p)\cdot v \,\mathrm{d}x = \int_\Omega \nabla u : \nabla v + p(\nabla\cdot v)\,\mathrm{d}x + \int_{\partial\Omega} \left(\frac{\partial u}{\partial n} - pn\right)\cdot v \,\mathrm{d}s. \tag{4.17}$$

We then consider the boundary conditions, which yields

$$\begin{aligned}\int_{\partial\Omega} \left(\frac{\partial u}{\partial n} - pn\right)\cdot v \,\mathrm{d}s &= \int_{\partial\Omega_D} \left(\frac{\partial u}{\partial n} - pn\right)\cdot v \,\mathrm{d}s + \int_{\partial\Omega_N} \left(\frac{\partial u}{\partial n} - pn\right)\cdot v \,\mathrm{d}s \\ &= \underbrace{\int_{\partial\Omega_D} \left(\frac{\partial u}{\partial n} - pn\right)\cdot 0 \,\mathrm{d}s}_{\text{As we choose } v\in H^1_{0,D}} + \int_{\partial\Omega_N} h\cdot v \,\mathrm{d}s \\ &= \int_{\partial\Omega_N} h\cdot v \,\mathrm{d}s.\end{aligned}$$

Defining

$$\begin{aligned} a(u,v) &= \int_\Omega \nabla u : \nabla v \,\mathrm{d}x \\ b(v,p) &= \int_\Omega p(\nabla\cdot v)\,\mathrm{d}x \\ L(v) &= \int_\Omega f\cdot v \,\mathrm{d}x + \int_{\partial\Omega_N} h\cdot v \,\mathrm{d}x \end{aligned} \tag{4.18}$$

we can rewrite the weak form of Stokes problem succinctly as: Find $(u,p) \in V \times Q$ such that for all $(v,q) \in \hat{V} \times \hat{Q}$

$$\begin{aligned} a(u,v) - b(v,p) &= L(v), \\ b(u,q) &= 0. \end{aligned} \tag{4.19}$$

The finite element formulation follows directly from this: Find $u_h \in V_{g,h}$ and $p_h \in Q_h$ such that

$$\begin{aligned} a(u_h,v_h) + b(p_h,v_h) &= L(v_h) \quad \forall v_h \in V_{0,h}, \\ b(q_h,u_h) &= 0 \qquad\quad \forall q_h \in Q_h. \end{aligned} \tag{4.20}$$

4.2.2 Brezzi conditions

For a saddle point problem of the form Eq. (4.20) to be well-posed, we require that four conditions are satisfied.

1. Boundedness of a:
$$a(u_h, v_h) \leq C_1 \|u_h\|_{V_h} \|v_h\|_{V_h} \quad \forall u_h, v_h \in V_h. \tag{4.21}$$

2. Boundedness of b:
$$b(u_h, q_h) \leq C_2 \|u_h\|_{V_h} \|q_h\|_{Q_h} \quad \forall u_h \in V_h, q_h \in Q_h. \tag{4.22}$$

3. Coersivity of a:
$$a(u_h, u_h) \geq C_3 \|u_h\|^2_{V_h} \quad \forall u_h \in Z_h, \tag{4.23}$$
where $Z_h = \{u_h \in V_h | b(u_h, q_h) = 0 \, \forall q_h \in Q_h\}$.

4. "Coercivity" of b:
$$\sup_{u_h \in V_h} \frac{b(u_h, q_h)}{\|u_h\|_{V_h}} \geq C_4 \|q_h\|_{Q_h} > 0 \quad \forall q_h \in Q_h. \tag{4.24}$$

The first three conditions are easily verified for Stokes problem, while the last one is difficult unless the elements are designed specifically to meet this condition. As the first three a rather simple, they are included here for completeness.

Boundedness of a Let $u_h, v_h \in V_h$. Then we have by the Cauchy–Schwarz inequality

$$a(u_h, v_h) = \int_\Omega \nabla u_h : \nabla v_h \, \mathrm{d}x \leq \|\nabla u_h\|_{L^2} \|\nabla v_h\|_{L^2} \leq \|u_h\|_1 \|v_h\|_1. \tag{4.25}$$

This shows that a is bounded, with $C_1 = 1$.

Boundedness of b Let $u_h \in V_h$ and $q_h \in Q_h$. We again utilize Cauchy–Schwarz inequality in order to show the bound, as

$$b(u_h, q_h) = \int_\Omega q_h (\nabla \cdot u_h) \, \mathrm{d}x \leq \|q_h\|_{L^2} \|\nabla \cdot u_h\|_{L^2} \leq \|q_h\|_{L^2} \|u_h\|_1. \tag{4.26}$$

The tricky portion here is to bound $\|\nabla \cdot u_h\|_{L^2} \leq \|u_h\|_1$, however we can actually do it without a coefficient, according to Kent.

Coercivity of a Finally, let $u_h \in V_h$. Then, we simply have

$$a(u_h, u_h) = \int_\Omega \nabla u_h : \nabla u_h \,\mathrm{d}x = \|\nabla u_h\|_{L^2}^2 \geq C_3 \|u_h\|_1, \tag{4.27}$$

where we can find C_3 as the norms are equivalent on H_0^1.

4.2.3 Oscillations in the pressure

The simplest case where we discover oscillations in the pressure, is in Poiseuille flow. The analytical solution here is given by

$$u = (y(1-y), 0) \quad \text{and} \quad p = 1 - x. \tag{4.28}$$

We choose this problem as the solution is known, which simplifies the analysis.

We then discretize the problem in a similarly simple manor, with linear Lagrange elements for velocity and pressure. What we would see then when we solve the problem is that the velocity is well-represented, while there are wild oscillations present in the pressure. If we had chosen quadratic elements in the velocity however, both the velocity and the pressure would be captured accurately.

The reason for the appearance of the oscillations with the P_1–P_1 elements stems from the fact that they are not *inf-sup* stable. That is, they do not fulfill the fourth Brezzi condition. The P_2–P_1 elements however fulfill this, resulting in a nice solution.

4.2.4 inf-sup and approximation

In order to obtain order optimal convergence rates, we require that the inf-sup condition,

$$\inf_{p \in Q_h} \sup_{v \in V_{h,g}} \frac{(p, \nabla \cdot v)}{\|v\|_1 \|p\|_0} \geq \beta > 0, \tag{4.29}$$

is satisfied. When this is satisfied, we get convergence

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq C h^k \|u\|_{k+1} + D h^{\ell+1} \|p\|_{\ell+1}, \tag{4.30}$$

where k and ℓ are the polynomial degree of the velocity and pressure respectively. Let's see how this is reflected in the approximations of a couple elements.

The Taylor–Hood element For the Taylor–Hood element, we have $u \in P_2$ and $p \in P_1$, subject to the restriction that they are continuous across elements. This element satisfies the inf-sup condition, and we can therefore directly get the approximation property

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq C h^2 (\|u\|_3 + \|p\|_2). \tag{4.31}$$

For the generalized P_k–P_{k-1} Taylor–Hood element, we similarly get

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq Ch^k(\|u\|_{k+1} + \|p\|_k). \tag{4.32}$$

The Crouzeix–Raviart element The Crouzeix–Raviart element also satisfies the inf-sup condition, and are given by $u \in P_1$ and $p \in P_0$. Here, u is *only* continuous in the mid-point of each side, and p is discontinuous. We then get the error estimate

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq Ch(\|u\|_2 + \|p\|). \tag{4.33}$$

It is inf-sup stable with our formulation of Stokes problem, however if we replace it with the more physically correct formulation

$$-\nabla \cdot \epsilon(u) - \nabla p = f, \tag{4.34}$$

where $\epsilon(u) = \frac{1}{2}(\nabla + \nabla^T)$ is the symmetric gradient, it cannot be used. We can generalize the element to odd degrees, but not even.

The Mini element The Mini element is linear in velocity and pressure, however it contains an extra degree of freedom in the velocity which is zero on all edges, dubbed a bubble function. For instance in the 2D case, the bubble function, on the reference triangle, is given by

$$b(x, y) = xy(1 - x - y). \tag{4.35}$$

This yields the error estimate

$$\|u - u_h\|_1 + \|p - p_h\|_0 \leq C_0 h\|u\|_2 + C_1 h^2\|p\|_2. \tag{4.36}$$

Note that the convergence in the velocity is linear, meaning that adding the bubble function brought stability, but not an increase in approximation order.

4.2.5 Stability without inf-sup

When stabilizing, we typically replace a system of the form

$$\begin{aligned} Au + B^T p &= f \\ Bu &= 0, \end{aligned} \tag{4.37}$$

with a system

$$\begin{aligned} Au + B^T p &= f \\ Bu - \epsilon Dp &= \epsilon d, \end{aligned} \tag{4.38}$$

where D is a positive, but not necessarily positive definite, matrix. We obtain a nonsingular system, as multiplying the first equation by A^{-1} and factoring yields

$$\begin{aligned}
u + A^{-1}B^T p &= A^{-1} f \\
\underbrace{Bu}_{\text{Insert second equation}} \qquad &= BA^{-1}f - BA^{-1}B^T p \\
\epsilon d + \epsilon D p &= BA^{-1}f - BA^{-1}B^T p \\
(BA^{-1}B^T + \epsilon D)p &= BA^{-1}f - \epsilon d,
\end{aligned} \tag{4.39}$$

and $(BA^{-1}B^T + \epsilon D)$ is nonsingular if D is nonsingular, as both components are positive.

Solving for p in the second equation we get

$$\begin{aligned}
\epsilon D p &= Bu - \epsilon d \\
p &= (\epsilon D)^{-1}(Bu - \epsilon d),
\end{aligned} \tag{4.40}$$

and inserting this into the second equation yields

$$\begin{aligned}
Au + B^T(\epsilon D)^{-1}(Bu - \epsilon d) &= f \\
\left(A + \frac{1}{\epsilon}B^T D^{-1} B\right) u &= f + D^{-1} d.
\end{aligned} \tag{4.41}$$

Then, $\left(A + \frac{1}{\epsilon}B^T D^{-1}B\right)$ is nonsingular as A is invertible, and $B^T D^{-1} B$ is positive.

The question is then how to choose D, and we have a couple of options. The three main techiques are:

1. Pressure stabilization, by choosing $\nabla \cdot v + \epsilon \Delta p = 0$. This is motivated through the convection-diffusion equation. This sets $D = A$.

2. Penalty method, by choosing $\nabla \cdot v + \epsilon p = 0$. Typically one then chooses the Velocity-Schur complement. This sets $D = M$.

3. Artificial compressibility, by choosing $\nabla \cdot -\epsilon \frac{\partial p}{\partial t}$. This is practical as it allows for time stepping. This sets $D = \frac{1}{\Delta t} M$.

5 Discretization of Navier–Stokes

Explain the difference between operator splitting and algebraic splitting in the context of the incompressible Navier–Stokes equations. We remark that algebraic splitting is a term usually used for discretizations where the PDEs are discretized in space prior to time. Show disadvantages for operator splitting schemes associated with boundary conditions. Explain the advantage with operator splitting schemes.

5.1 Short-form answer

The incompressible Navier–Stokes equation is given by

$$\begin{aligned} \frac{\partial u}{\partial t} + u \cdot \nabla u = \frac{\nabla p}{\rho} + \nu \nabla^2 u + f, \\ \nabla \cdot u = 0. \end{aligned} \tag{5.1}$$

Operator splitting refers to discretizing the equation in time before solving the equation in space. Algebraic splitting on the other hand refers to splitting the equation in space first.

When operator splitting, we typically firstly discretize with a forward Euler scheme, or similar, which gives us an equation of the form

$$\frac{u^{n+1} - u^n}{\Delta t} + u^n \cdot \nabla u^n = \frac{\nabla p^n}{\rho} + \nu \nabla^2 u^n + f^n. \tag{5.2}$$

This has some issues, primarily the fact that we have no way of updating the pressure, as we have no expression for p^{n+1}. In addition, we have no reason to assume that $\nabla \cdot u^{n+1} = 0$.

When algebraically splitting, we first find the weak form of the momentum equation, which is given by, with $\langle \cdot, \cdot \rangle$ denoting the inner product in $L^2(\Omega)$,

$$\left\langle \rho \frac{\partial u}{\partial t}, v \right\rangle + \langle \rho u \cdot \nabla u, v \rangle - \langle p, \nabla \cdot v \rangle + \langle \nabla u, \nabla v \rangle = \langle f, v \rangle + \langle t_N, v \rangle_{\partial \Omega_N}, \tag{5.3}$$

while the weak form of the continuity equation is given by

$$\langle \nabla \cdot u, q \rangle = 0. \tag{5.4}$$

We can then rewrite this in the discretized matrix form as

$$\begin{aligned} M \frac{\partial u}{\partial t} + N(u) u + Au + Bp = f \\ B^T u = 0, \end{aligned} \tag{5.5}$$

with the corresponding matrices as in the weak form. This system however carries some issues, as it is non-linear, non-symmetric and indefinite, and there is in general no good method for solving it.

5.2 Long-form answer

5.2.1 Operator splitting

The incompressible Navier–Stokes equation is given by

$$\begin{aligned} \frac{\partial u}{\partial t} + u \cdot \nabla u &= -\frac{\nabla p}{\rho} + \nu \nabla^2 u + f, \\ \nabla \cdot u &= 0. \end{aligned} \tag{5.6}$$

Operator splitting refers in this case to discretizing in time prior to space. If we discretize simply with a forward Euler scheme, we end up with

$$\frac{u^{n+1} - u^n}{\Delta t} + u^n \cdot \nabla u^n = -\frac{\nabla p^n}{\rho} + \nu \nabla^2 u^n + f^n, \tag{5.7}$$

which rearranged gives us

$$u^{n+1} = u^n + \Delta t \left(-u^n \cdot \nabla u^n - \frac{\nabla p^n}{\rho} + \nu \nabla^2 u^n + f^n \right). \tag{5.8}$$

This is relatively simple, however it immediately raises a few issues. For one, we get no expression for how we should update the pressure, i.e. what should p^{n+1} be? Secondly, we have no reason to assume that $\nabla \cdot u^{n+1} = 0$.

One way to overcome this is to introduce a tentative solution

$$u^* = u^n + \Delta t \left(-u^n \cdot \nabla u^n - \frac{\nabla p^n}{\rho} + \nu \nabla^2 u^n + f^n \right), \tag{5.9}$$

then saying that the real solution u^{n+1} should satisfy

$$u^{n+1} = u^n + \Delta t \left(-u^n \cdot \nabla u^n - \frac{\nabla p^{n+1}}{\rho} + \nu \nabla^2 u^n + f^n \right). \tag{5.10}$$

Subtracting Eq. (5.9) from Eq. (5.10), we get

$$u^{n+1} - u^* = -\frac{\Delta t}{\rho} \nabla \left(p^{n+1} - p^n \right). \tag{5.11}$$

As we should have $\nabla \cdot u^{n+1} = 0$, taking the divergence of Eq. (5.11) gives us

$$\nabla \cdot u^* = \frac{\Delta t}{\rho} \nabla^2 \left(p^{n+1} - p^n \right). \tag{5.12}$$

Letting now $\phi = p^{n+1} - p^n$, we can write this as

$$\nabla^2 \phi = \frac{\rho}{\Delta t} \nabla \cdot u^*, \tag{5.13}$$

which we recognize as a Poisson problem, as u^* is known.

This basically gives us all the ingredients we need for a scheme, which we summarize as

1. Compute the tentative velocity

$$u^* = u^n + \Delta t \left(-u^n \cdot \nabla u^n - \frac{\nabla p^n}{\rho} + \nu \nabla^2 u^n + f^n \right). \tag{5.14}$$

2. Solve the Poisson problem

$$-\nabla^2 \phi = -\frac{\rho}{\Delta t} \nabla \cdot u^* \tag{5.15}$$

   in order to find the update for the pressure.

3. Update the velocity by

$$u^{n+1} = u^* - \frac{\Delta t}{\rho} \nabla \phi. \tag{5.16}$$

4. Update the pressure by

$$p^{n+1} = p^n + \phi. \tag{5.17}$$

This scheme is rather simple when stated like this, however there are some unadressed issues. Firstly is the issue of boundary trouble. As we need to solve a Poisson problem in the pressure update, we require boundary conditions along the entire boundary, even though this isn't strictly necessary in the original problem.

Another approach is to use an implicit scheme, given where we get the tentative velocity by

$$\frac{u^* - u^n}{\Delta t} + u^* \cdot \nabla u^* = -\frac{\nabla p^n}{\rho} + \nu \nabla^2 u^* + f^{n+1}. \tag{5.18}$$

Rearranging this again leads to

$$u^* - \Delta t \left((-u^* \cdot \nabla u^*) - \frac{1}{\rho} \nabla p^n + \nu \nabla^2 u^* \right) = u^n + \Delta t f^{n+1}. \tag{5.19}$$

The term $u^* \cdot \nabla u^*$ is problematic as it is non-linear, however we can apply a typical linearization technique by replacing it with $u^n \cdot \nabla u^*$ to get

$$u^* - \Delta t \left((-u^n \cdot \nabla u^*) - \frac{1}{\rho} \nabla p^n + \nu \nabla^2 u^* \right) = u^n + \Delta t f^{n+1}. \tag{5.20}$$

What we really want is however

$$u^{n+1} - \Delta t \left((-u^n \cdot \nabla u^{n+1}) - \frac{1}{\rho} \nabla p^n + \nu \nabla^2 u^{n+1} \right) = u^n + \Delta t f^{n+1}. \tag{5.21}$$

Subtracting the first equation from the second again now gives us a more complicated expression.

For simplicity, we introduce the convection-diffusion operator

$$s(u^c) = \Delta t \left(-u^n \cdot \nabla u^c + \nu \nabla^2 u^c\right). \tag{5.22}$$

With

$$u^{n+1} = u^* + u^c, \tag{5.23}$$

we can now write

$$\begin{aligned} u^c - s(u^c) + \frac{\Delta t}{\rho}\nabla\phi &= 0 \\ \nabla \cdot u^c &= -\nabla \cdot u^*. \end{aligned} \tag{5.24}$$

So far, we haven't gotten any closer to our deisred goal, as it it just as hard as the original equations. However, note that we are using a first order approximation to the time derivative, and as the leading term in $s(u^c)$ is first order, we can drop it while still having a first order approximation. We then get

$$\begin{aligned} u^c + \frac{\Delta t}{\rho}\nabla\phi &= 0 \\ \nabla \cdot u^c &= -\nabla \cdot u^*, \end{aligned} \tag{5.25}$$

where we can again rewrite the first equation as

$$-\nabla^2\phi = -\frac{\rho}{\Delta t}\nabla \cdot u^*. \tag{5.26}$$

This second approach amounts to the Incremental Pressure Correction Scheme (IPCS), which consists of the four steps

1. Compute the tentative velocity

$$u^* - s(u^*) + \frac{\Delta}{\rho}\nabla p^n = f^{n+1}. \tag{5.27}$$

2. Solve the Poisson problem for the pressure

$$-\nabla^2\phi = -\frac{\rho}{\Delta t}\nabla \cdot u^* \tag{5.28}$$

3. Update the velocity

$$u^{n+1} = u^* - \frac{\Delta t}{\rho}\nabla\phi \tag{5.29}$$

4. Update the pressure

$$p^{n+1} = p^n + \phi \tag{5.30}$$

5.2.2 Issues with operator splitting

Operator splitting introduces trouble along the boundary. NS in itself, in 3D, requires 3 conditions in every point at the boundary. IPCS however requries 4 conditions along the boundary, 3 for the tentative velocity, and 1 for the Poisson equation.

We can derive boundary conditions for ϕ in two ways:

1. From the scheme

   (a)
   $$u^{n+1} = u^* - \frac{\Delta t}{\rho}\nabla\phi$$

   (b) As u^{n+1} and u^* have the same BCs, we obtain homogenous Neumann conditions for ϕ, as
   $$\nabla\phi \cdot n = \frac{\rho}{\Delta t}(u^{n+1} - u^*) \cdot n = 0 \tag{5.31}$$

2. From the Navier–Stokes equations we have that

   (a)
   $$\nabla p^n = -\rho\left(\frac{\partial u^n}{\partial t} + u^n \cdot \nabla u^n\right) + \mu\nabla^2 u^n + f$$

   (b) As $\phi = p^{n+1} - p^n$ and $p^{n+1} \neq p^n$ we obtain a non-homogeneous condition.

In conclusion, we arrive at two different conditions which both seem reasonable. The difference between the two is first order.

5.2.3 Algebraic splitting

Here, we firstly seek the weak form of the NS equations. We use $\langle\cdot,\cdot\rangle$ to denote the L^2-inner product. Multiplying the momemtum equation by a test and integrating by parts then yields

$$\langle\rho\frac{\partial u}{\partial t}, v\rangle + \langle\rho u \cdot \nabla u, v\rangle - \langle p, \nabla \cdot v\rangle + \langle\nabla u, \nabla v\rangle = \langle f, v\rangle + \langle t_N, v\rangle_{\Gamma_N}, \tag{5.32}$$

doing the same with the continuity equation yields

$$\langle\nabla \cdot u, q\rangle = 0. \tag{5.33}$$

6 Other formulations

Consider the various formulations of the Poisson problem. Show that minimization of energy

$$\int_\Omega \frac{1}{2}(\nabla u)^2 - f\,u\,\mathrm{d}x \tag{6.1}$$

corresponds to a weak formulation that through Green's lemma gives a strong formulation. Do the same thing for the least square formulation and arrive at a bi-harmonic equation with some additional boundary conditions. Consider both the mixed formulation and least square formulation of the mixed formulation, and note that only the first requires the Brezzi conditions.

6.1 Short-form answer

The strong formulation of the Poisson problem is

$$-\Delta u_S = f \quad \text{in } \Omega, \quad u_S = g \quad \text{on } \partial\Omega. \tag{6.2}$$

This formulation requires that $u_S \in C^2(\Omega)$, given an $f \in C^0(\Omega)$. The weak formulation on the other hand is to find $u_W \in H^1_g(\Omega)$ given an $f \in L^2(\Omega)$ such that

$$\int_\Omega \nabla u_W \cdot \nabla v = f\,v\,\mathrm{d}x. \tag{6.3}$$

We have now eased the continuity requirement on u_W.

We can also derive the weak solution from the energy minimization problem

$$u_W = \underset{u\in H^1_g(\Omega)}{\arg\min} \int_\Omega \frac{1}{2}(\nabla u)^2 - f\,u\,\mathrm{d}x = \underset{u\in H^1_g(\Omega)}{\arg\min} E(u,f). \tag{6.4}$$

In order to see this, note that $H^1_g(\Omega)$ is known to be separable, which means that it is spanned by a countable set of basis functions $\{\psi_i\}_{i=1}^\infty$. This means that we can write any $v \in H^1_g(\Omega)$ as

$$v = \sum_i v_i \psi_i, \tag{6.5}$$

and we then have

$$\frac{\partial v}{\partial v_i} = \psi_i, \tag{6.6}$$

or specifically in our case $u_W = \sum_i u_i^W \psi_i$. We then seek to find u_W such that

$$\frac{\partial E}{\partial u_W} = 0. \tag{6.7}$$

This gives us

$$\frac{\partial E}{\partial u_i^W} = \int_\Omega \nabla u_W \cdot \nabla \psi_i - f\,\psi_i \,\mathrm{d}x = 0, \tag{6.8}$$

which is exactly the weak formulation we had before.

As is a general strategy when approximating a function, we can also formulate a least squares problem based on the PDE. Here we seek to minimize the functional

$$E_{LS}(v, f) = \int_\Omega (-\Delta v - f)^2 \,\mathrm{d}x. \tag{6.9}$$

If u_S is the strong solution, then we clearly have that $E_{LS}(u_S, f) = 0$. If however we have a weak solution u_W, then we may have that $E_{LS}(u_W, f) = \infty$, as $u_W \in H^1(\Omega)$, meaning that we only have control over the first derivative. We must instead have that $u_{LS} \in H^2(\Omega)$, in order the get the proper regularity. This means that we can write the least squares problem as

$$u_{LS} = \operatorname*{arg\,min}_{v \in H^2(\Omega)} E_{LS}(v, f). \tag{6.10}$$

Solving this in the same manner as before, we find

$$\begin{aligned}
\frac{\partial E_{LS}}{\partial u_i^{LS}} &= \frac{\partial}{\partial u_i^{LS}} \int_\Omega (-\Delta u_{LS} - f)^2 \,\mathrm{d}x \\
&= \int_\Omega \frac{\partial}{\partial u_i^{LS}} (-\Delta u_{LS} - f)^2 \,\mathrm{d}x \\
&= 2 \int_\Omega (-\Delta u_{LS} - f)(-\Delta \psi_i) \,\mathrm{d}x
\end{aligned}$$