Topics in Computational Mechanics
MEK4250

Ivar Stangeby

June 12, 2017

Abstract

In this document we consider the different topics in computational mechanics covered in the course MEK4250. A lot of emphasis is put on showing the well-posedness of the various variational and finite element formulations of the problems. Error estimation methods are given for the various problems. This document is hopelessly void of any mathematical rigour.

Contents

Weak Formulation And Finite Element Error Estimation

1

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

1.1 Finite Element Formulation

The Poisson problem is formulated as:

$$-\nabla\cdot(\kappa\,\nabla u) = f \text{ in } \Omega, \tag{1.1}$$

$$u = u_0 \text{ on } \Gamma_D, \tag{1.2}$$

$$-\kappa\,\nabla u\cdot n = g \text{ on } \Gamma_N. \tag{1.3}$$

Here u denotes the unknown field. We associate to this strong formulation of the problem, the bilinear operator $a\colon \hat{V}\times V\to\mathbb{R}$, and the linear operator $L\colon \hat{V}\to\mathbb{R}$ as follows:

$$a(u,v) \coloneqq \langle -\nabla\cdot(\kappa\,\nabla u), v\rangle, \tag{1.4}$$

$$L(v) \coloneqq \langle f, v\rangle. \tag{1.5}$$

Weak Formulation

We can therefore consider instead the *weak formulation* of the Poisson problem, which is: Find $u \in V$ such that

$$a(u, v) = L(v) \text{ for all } v \in \hat{V}. \tag{1.6}$$

We need to place some requirements to the spaces V and $\hat{V}$ in order to satisfy the boundary conditions. However, one requirement we impose immediately is that $v = 0$ on Γ_D for all $v \in \hat{V}$. What these spaces should be is not immediate from the current formulation. Throughout the following, $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product on V. Expanding a using among others the Gauss–Green lemma yields:

$$a(u, v) = \langle -\nabla \cdot (\kappa \nabla u), v \rangle = -\langle \kappa \Delta u, v \rangle \tag{1.7}$$

$$= \langle \kappa \nabla u, \nabla v \rangle - \int_{\partial\Omega} -v\kappa \frac{\partial u}{\partial x} \cdot n \, dS. \tag{1.8}$$

The boundary integral above can be rewritten using the partitioning of the boundary:

$$\int_{\partial\Omega} -v\kappa \frac{\partial u}{\partial x} \cdot n \, dS = -\int_{\Gamma_D} v\kappa \frac{\partial u}{\partial x} \cdot n \, dS + \int_{\Gamma_N} -v\kappa \frac{\partial u}{\partial x} \cdot n \, dS. \tag{1.9}$$

Applying the boundary conditions over respective boundaries and condition on $\hat{V}$ we get in total that:

$$a(u, v) = \langle \kappa \nabla u, \nabla v \rangle - \int_{\Gamma_N} gv \, dS. \tag{1.10}$$

Since the boundary term in a is independent of u we decide to transfer this term to L:

$$a(u, v) = \langle \kappa \nabla u, \nabla v \rangle, \quad L(v) = \langle f, v \rangle + \int_{\Gamma_N} gv \, dS. \tag{1.11}$$

We now have what we need to determine the spaces $\hat{V}$ and V. Note that we in the weak form only require u and v to be once differentiable. Furthermore, we require u to reduce to u_0 on Γ_D by the boundary conditions, while we need v to vanish identically at Γ_D. We therefore decide on the test and trial spaces

$$V = H_g^1(\Omega) \coloneqq \left\{ u \in H^1(\Omega) : u = g \text{ on } \Gamma_D \right\} \subseteq H^1(\Omega), \tag{1.12}$$

$$\hat{V} = H_0^1(\Omega) \coloneqq \left\{ u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_D \right\} \subseteq H^1(\Omega). \tag{1.13}$$

Finite Element Formulation

In order to compute with these spaces, we need to introduce a basis. However, the spaces might be infinite dimensional, so we approximate by finite subspaces $\hat{V}_h$ and V_h respectively[1]. Since we have $\hat{V} \subseteq V$ we have $\hat{V}_h \subseteq V_h$, and we therefore use the same basis vectors for both test and trial functions. We seek a solution $u = \sum_i c_i \varphi_i$ such that

$$\sum_{i=1}^{N} c_i a(\varphi_i, \varphi_j) = L(\varphi_j) \text{ for } j = 1, \dots, M \tag{1.14}$$

where N, M denotes the dimensions of $V_h, \hat{V}_h$ respectively. This determines a linear system

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b} \tag{1.15}$$

where the matrix entries are determined as follows:

$$A_{i,j} = a(\varphi_i, \varphi_j), \quad b_j = L(\varphi_j). \tag{1.16}$$

1.2 Well Posedness of Weak Formulation

In this section we discuss whether the problem in its weak formulation is in fact well posed. Does there exist a solution, and if it does, is this solution unique? The Lax–Milgram theorem provides sufficient conditions for this problem to be well posed, and hence the solution to exist and be unique. We verify the three properties in turn:

Boundedness of a: Let $u, v \in H^1(\Omega)$. Then we have

$$a(u,v) = \langle \kappa \nabla u, \nabla v \rangle \underbrace{\leq}_{\text{Using the Cauchy–Schwartz inequality}} \overbrace{\|\kappa\|}^{\text{subordinate matrix norm}} \| \nabla u\|_0 \| \nabla v\|_0 = \|\kappa\| \|u\|_1 \|v\|_1 \tag{1.17}$$

This is close to proving boundedness of a, however, I do not know how to deal with the $\|\kappa\|$. However, for $\kappa = 1$ this proves the claim.

[1] Much of the finite element theory amounts to determining what the error in this specific approximation is.

Coercivity of a: Let $u \in H^1(\Omega)$ and consider the following:

$$\|u\|_1^2 \overset{\text{def}}{=} \|u\|_0^2 + \|\nabla u\|_0^2 \underbrace{\leq}_{\text{Using the Poincaré inequality on } H_0^1(\Omega)} \geq (C^2+1)\|\nabla u\|_0^2 \tag{1.18}$$

This shows that $a(u,u)$ $(C^2+1)^{-1}\|u\|_1^2$ in the case where $\kappa = 1$ identically.

Boundedness of L:

$$L(v) = \langle f, v\rangle + \int_{\Gamma_N} gv\, dS \tag{1.19}$$

Using Cauchy–Schwartz on both terms yields

$$\leq \|f\|_0\|v\|_0 + \|g\|_{L^2(\Gamma_N)}\|v\|_{L^2(\Gamma_N)}; \tag{1.20}$$

Using the fact that v is defined on the entirety of Ω, the L^2 norm is certainly larger over the whole domain

$$\leq \|f\|_0\|v\|_0 + \|g\|_{L^2(\Gamma_N)}\|v\|_0; \tag{1.21}$$

Using that the H^1 norm greater than or equal to the L^2 norm

$$\leq (\|f\|_0 + \|g\|_{L^2(\Gamma_N)})\|v\|_1, \tag{1.22}$$

$$= D\|v\|_1. \tag{1.23}$$

This proves the boundedness of L.

This means that the weak formulation of the Poisson problem satisfies the Lax–Milgram theorem, hence is well posed. In addition, the solution u satisfies

$$\|u\|_1 \leq \frac{(C^2+1)^{-1}}{\|\kappa\|}\|f\|_{-1}. \tag{1.24}$$

In the abstract framework, this corresponds to the operator $A : \hat{V} \to \hat{V}^\star$ given by

$$\langle Au, v\rangle = L(v) \tag{1.25}$$

is an isomorphism.

1.3 Extensions to Other Problems

Recall that the equation of linear elasticity is given as

$$-2\mu(\nabla\cdot\varepsilon(u)) - \lambda\,\nabla(\nabla\cdot u) = f \tag{1.26}$$

where $\varepsilon(u) \coloneqq (\nabla\, u + (\nabla\, u)^T)/2$ is the strain tensor. Ignoring the second term, first term is supposedly a Poisson-equation[2]. However, in order to check whether we can apply the results found for the abstract Poisson-problem we need to determine what the $\nabla(\nabla\cdot u)$ in the second term is. The Helmholtz decomposition theorem states that any u in $L^2(\Omega)$ can be decomposed into a curl-free part ψ and a divergence-free part φ:

$$u = \psi + \varphi \tag{1.27}$$

where $\nabla\cdot\varphi = 0$ and $\nabla\times\psi = 0$. In the light of this, we can consider the two special cases where either φ or ψ are zero.

(i) Assume that $\varphi = 0$. This means that $u = \psi$ consists only of the divergence free part. We then have that $\nabla(\nabla\cdot u) = 0$. Consequently, the second term vanishes in equation (1.26).

(ii) Assume that $\psi = 0$, i.e., u consists only of the curl free part. Then using the identity

$$\Delta\, u = \nabla(\nabla\cdot u) - \nabla\times(\nabla\times u) \tag{1.28}$$

we see that $u = \nabla(\nabla\cdot(u))$. Hence, equation (1.26) reduces to a Poisson problem.

In these two special cases, we have that the linear elasticity problem reduces to a Poisson problem of the form

$$-(\mu+\lambda)\,\Delta\, w = f. \tag{1.29}$$

for some source term f.

[2]This needs to be verified.

1.4 A Priori Error Estimates

In the following, we work over arbitrary test and trial spaces V, as these estimates are general.

Energy Norm

We now discuss some a priori estimates. Recall that any arbitrary inner product induces a norm by

$$\|x\| \coloneqq \sqrt{\langle x, x\rangle}. \tag{1.30}$$

It turns out that the bilinear form a may, or may not, constitute an inner product. If it does, then we may talk about the norm induced by this bilinear form. We call this the *energy norm*. We first need to verify that a does indeed constitute an inner product. The only condition that is not trivial is the coercivity, however we already proved this for the Lax–Milgram theorem. We therefore define the energy norm

$$\|u\|_E \coloneqq \sqrt{a(u,u)}. \tag{1.31}$$

Consider now the error $e \coloneqq u - u_h$. Let $v \in \hat{V}$ be arbitrary. In the energy norm, we have

$$\|e\|_E^2 = a(e,e) = a(e, u - u_h) = a(e, u - v + v - u_h) \tag{1.32}$$

$$= a(e, u - v) + a(e, \underbrace{u_h - v}_{\in \hat{V}}) = a(e, u - v). \tag{1.33}$$

Using the Cauchy–Schwartz inequality, we have that

$$\|e\|_E^2 \leq \|e\|_E \|u - v\|_E \implies \|e\|_E \leq \|e - v\|_E \tag{1.34}$$

for all $v \in \hat{V}$. This does however not give any sharp bounds, so it is hard to quantify exactly what magnitude the error has. We can however combine this result with an interpolation error estimate:

$$\|e\|_E \leq \|u - \pi_{q,h} u\|_E \leq C(q) \|h^{q+1} D^{q+1} u\|. \tag{1.35}$$

Here, $\pi_{q,h} u$ denotes the q-th order interpolant to u, and h is the maximum mesh size.

Error estimate without coercivity of a

The symmetry of a is quite a strong requirement, and this is not always satisfied. In this section, we do *not* assume that a is symmetric. In the cases where a is bounded however we can consider the error in the vector space norm as follows. Let $v \in \hat{V}$ be arbitrary. Then using the coercivity of a we have:

$$\|e\|_V^2 \leq \frac{1}{\alpha} a(e,e) = \frac{1}{\alpha} a(e, u - v + v - u_h) \tag{1.36}$$

$$= \frac{1}{\alpha} a(e, u - v) \underbrace{\leq}_{\text{Using boundedness of } a} \frac{D}{\alpha} \|e\|_V \|u - v\|_V. \tag{1.37}$$

Dividing by $\|e\|_V$ and combining with an interpolation error estimate, as we did above,

$$\|e\|_V \leq \frac{DC(q)}{\alpha} \|h^{q+1} D^{q+1} u\|. \tag{1.38}$$

1.5 Error Approximation

In the following we assume we solve a constructed problem with known solution. Let N denote the number of basis elements over the domain. Then we can consider the error as a function of N. Using degree p elements over a mesh with maximal mesh size of h. Let u denote the analytic solution, and u_N the computed solution with N basis elements. Denote by e_N the error in the corresponding approximation. We wish to approximate the convergence rate β in the following:

$$\|e\|_V \leq \|h^\beta D^\beta u\|. \tag{1.39}$$

We can rewrite this as a linear equation in h with slope β and constant term $\log(D^\beta u)$. Computing e_N for various N we can find a linear regression line through the data and approximate β.

Discretization Of Convection-Diffusion

2

Problem. 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 Streamline diffusion / Petrov–Galerkin methods resolve these problems. Discuss also approximation properties in light of Ceas lemma.

2.1 Finite Element Formulation

The convection diffusion problem is given as:

$$-\mu \Delta u + \omega \cdot \nabla u = f \text{ in } \Omega, \tag{2.1}$$

$$u = g \text{ on } \Gamma_D. \tag{2.2}$$

Here u is the unknown, μ is the diffusivity, and ω is a velocity. We associate to this problem the bilinear operator $a\colon V \times \hat{V} \to \mathbb{R}$ and the linear operator $L\colon \hat{V} \to \mathbb{R}$ given by

$$a(u, v) \coloneqq \langle -\mu \Delta u, v \rangle + \langle \omega \cdot \nabla u, v \rangle; \tag{2.3}$$

$$L(v) \coloneqq \langle f, v \rangle. \tag{2.4}$$

Weak Formulation

We can now consider the *weak formulation* of the Convection-Diffusion problem. That is: Find u in V such that:

$$a(u, v) = L(v) \text{ for all } v \in \hat{V}. \tag{2.5}$$

Again, we need to later properly determine the spaces V and $\hat{V}$, however, as always, we require $v = 0$ on Γ_D for all $v \in \hat{V}$. In order to get rid of the Laplacian term, we employ the Gauss–Green lemma, yielding:

$$a(u, v) = \langle \mu \nabla u, \nabla v \rangle - \underbrace{\int_{\Gamma_D} \mu \frac{\partial u}{\partial x} v \cdot n \, dS}_{\text{This is zero due to } v \in \hat{V}} + \langle \omega \cdot \nabla u, v \rangle \tag{2.6}$$

$$= \langle \mu \nabla u, \nabla v \rangle + \langle \omega \cdot \nabla u, v \rangle. \tag{2.7}$$

We again only require one derivative for u and v so we let

$$V = H^1_g(\Omega) \subseteq H^1(\Omega), \tag{2.8}$$

$$\hat{V} = H^1_0(\Omega) \subseteq H^1(\Omega), \tag{2.9}$$

as in the Poisson problem. We delay the finite element formulation until after the Streamline diffusion / Petrov–Galerkin method has been discussed.

We now need to verify that this abstract problem is well posed.

2.2 Well Posedness of Weak Formulation

While our problem is not *homogeneous* in the Dirichlet conditions, we can reduce it to a homogeneous problem. The reason for doing this is to employ the Poincaré inequality, which only holds for $H^1_0(\Omega)$. Again, the Lax–Milgram theorem gives sufficient conditions for the problem being well posed. We have two cases for this specific problem: (i) Incompressible flow, $\nabla \cdot \omega = 0$; or (ii) compressible flow, $\nabla \cdot \omega \neq 0$. We deal with these two cases separately. Furthermore, for simplicity, we define

$$b(u, v) \coloneqq \langle \mu \nabla u, \nabla v \rangle, \tag{2.10}$$

$$c_\omega(u, v) \coloneqq \langle \omega \cdot \nabla u, v \rangle, \tag{2.11}$$

and note that $a(u, v) = b(u, v) + c_\omega(u, v)$.

Incompressible Flow

For the incompressible case, we have $\nabla \cdot \omega = 0$. In addition to this assumption, we also assume that the flow velocities are bounded, i.e., $D_\omega \coloneqq \|\omega\|_\infty < \infty$. It can then be shown that the bilinear form $c_\omega(u, v)$ is *skew-symmetric*, that is

$$c_\omega(u, v) = -c_\omega(v, u). \tag{2.12}$$

We now show that the conditions in the Lax–Milgram theorem is satisfied.

Coercivity of a: Using the skew-symmetric property of $c_\omega(u, v)$, we get that $c_\omega(u, u) = -c_\omega(u, u)$ which implies that $c_\omega(u, u) = 0$. Therefore, we have

$$a(u, u) = b(u, u) + c_\omega(u, u) = b(u, u). \tag{2.13}$$

So, a is coercive, as

$$b(u, u) = \mu \int_\Omega (\nabla u)^2 \, d\Omega \geq \mu \left(\int_\Omega \nabla u \, d\Omega \right)^2 = \mu |u|_1^2. \tag{2.14}$$

Boundedness of a: Applying the Cauchy–Schwartz inequality we have

$$a(u, v) = \langle \mu \nabla u, \nabla v \rangle + \langle \omega \cdot \nabla u, v \rangle \tag{2.15}$$

$$\leq |\mu| \| \nabla u\|_0 \| \nabla v\|_0 + \|\omega \cdot \nabla u\|_0 \|v\|_0 \tag{2.16}$$

Using the assumption of bounded flow velocities; and that the problem has been reduced to homogeneous Dirichlet conditions — so we can apply the Poincaré inequality with domain dependent factor C_Ω — we get:

$$\leq |\mu| |u|_1 |v|_1 + D_\omega |u|_1 \|v\|_1 \tag{2.17}$$

$$\leq (\mu + D_\omega C_\Omega) |u|_1 |v|_1. \tag{2.18}$$

Consequently, a is bounded.

Boundedness of L: Applying the Cauchy–Schwartz inequality we get

$$L(u,v) = \langle f, v\rangle \leq \|f\|_0\|v\|_0 \leq \|f\|_1\|v\|_1. \tag{2.19}$$

The Lax–Milgram conditions are satisfied, hence the weak formulation of the convection-diffusion problem is well posed. In addition, the solution u satisfies

$$\|u\|_1 \leq \frac{\mu + D_\omega C_\Omega}{\mu}\|f\|_{-1}. \tag{2.20}$$

Compressible Flow

In the case where the flow is compressible, i.e., $\nabla\cdot\omega \neq 0$, we need to put some extra restrictions on the flow velocities ω in order to ensure well posedness. This is because in the general case, we have $c_\omega(u,u) \neq 0$. The coercivity of a was the only property where we assumed incompressibility, hence the two other properties remain the same.

Coercivity of a with compressible fluids: If $D_\omega C_\Omega \leq B\mu$ where $B < 1$ we obtain

$$a(u,u) = \langle \mu\,\nabla u, \nabla u\rangle + \langle \omega\cdot\nabla u, u\rangle \tag{2.21}$$

$$\geq \mu(1 - D_\omega)\|u\|_1^2, \tag{2.22}$$

however, it is not clear to me exactly how this result is obtained.

2.3 Oscillations in the Solution

Assume for now we are working with the one dimensional convection diffusion problem on a mesh with h denoting the largest mesh element. Solving this with first order Lagrangian elements corresponds to the central finite difference scheme

$$-\frac{\mu}{h^2}[u_{i+1} - 2u_i + u_{i-1}] - \frac{\omega}{2h}[u_{i+1} - u_{i-1}] = 0 \tag{2.23}$$

for $i = 1,\ldots,N-1$, where N denotes the number of elements. Assume the boundary conditions are $u_0 = 0$ and $u_N = 1$. Examining the above expression we see that in

the limit $\mu \to 0$ that the scheme reduces to

$$\frac{\omega}{2h}[u_{i+1} - u_{i-1}] = 0 \tag{2.24}$$

for $i = 1, \ldots, N$ with $u_0 = 0$ and $u_N = 1$. Here we see that u_{i+1} is coupled to u_{i-1} but not u_i. This means that we may get a numerical solution consisting of two sequences $(u_{2i})_i$ and $(u_{2i+1})_i$ that have no relation to each other. This may very well cause oscillations in the solution.

Finite Difference Upwinding

One remedy is to introduce the concept of *upwinding*. Instead of using a central finite difference scheme as above, one employs either a forward or a backward first order scheme, based on the velocity ω. That is:

$$u'(x_i) \approx \frac{1}{h}[u_{i+1} - u_i] \text{ if } \omega < 0, \tag{2.25}$$

$$u'(x_i) \approx \frac{1}{h}[u_i - u_{i-1}] \text{ if } \omega > 0. \tag{2.26}$$

This upwinding scheme can be seen as a special case of *artificial diffusion*, where one solves the "artificial" problem

$$-(\mu + \varepsilon)\,\Delta u + \omega \cdot \nabla u = f, \tag{2.27}$$

with $\varepsilon > 0$ some arbitrary real number. In particular, choosing $\varepsilon = h/2$ one regains the upwinding scheme mentioned above.

The fact that the finite element method coincides with the finite difference method in the case of one dimensional convection diffusion, and first order Lagrangian elements is not something that holds in general. Our question in the following is then: How do we implement artificial diffusion in our finite element method? This leads us to the Streamline diffusion / Petrov–Galerkin methods.

2.4 Streamline Diffusion / Petrov–Galerkin

Our goal here is to add artificial in a consistent way that does not changes the solution as h tends to zero. It turns out that naively adding artificial diffusion to our current finite element formulation does not give us what we want. We first examine why.

Naive Artificial Diffusion

Recall that our problem reads: Find $u \in V$ such that

$$\langle \mu \nabla u, \nabla v \rangle + \langle \omega \cdot \nabla u, v \rangle = \langle f, v \rangle \text{ for all } v \in \hat{V}. \tag{2.28}$$

Replacing μ by $\mu + \varepsilon$ yields a new bilinear operator $\tilde{a}$:

$$\tilde{a}(u, v) \coloneqq \langle \mu \nabla u, \nabla v \rangle + \langle \varepsilon \nabla u, \nabla v \rangle + \langle \omega \cdot \nabla u, v \rangle. \tag{2.29}$$

This can be written succinctly as

$$\tilde{a}(u, v) = a(u, v) + \varepsilon \langle \nabla u, \nabla v \rangle. \tag{2.30}$$

If we let $\varepsilon = h/2$ we see that in the limit as $h \to 0$, we have $\tilde{a}(u, v) \to a(u, v)$, and the scheme is consistent in this sense. However, it is not *strongly consistent* as it does not satisfy the Galerkin-orthogonality as a does:

$$a(u - u_h, v) = 0 \text{ for all } v \in \hat{V}_h, \tag{2.31}$$

namely that this equation is zero for *all* discretization, and not just in the limit.

We can however make the scheme strongly consistent by employing different spaces for the test functions and the trial functions.

Petrov–Galerkin method

The only difference between the Petrov–Galerkin and the standard Galerkin formulation is that the trial and test functions differ. In the standard Galerkin method, the same basis is used for both test and trial functions. In the Petrov-Galerkin method the test functions are tailored to ensure a strongly consistent scheme.

Finite Element Formulation

Letting $\hat{V}_h$ and V_h denote the finite dimensional subspaces of $\hat{V}$ and V respectively. Assume these are given by bases $(\psi_j)_{i=1}^M$ and $(\varphi_i)_{i=1}^N$. The finite element formulation then gives rise to a linear system

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b}, \tag{2.32}$$

where the matrix elements are given as

$$A_{i,j} = \langle \mu \nabla \varphi_i, \nabla \psi_j \rangle + \langle \omega \cdot \nabla \varphi_i, \psi_j \rangle, \tag{2.33}$$

$$b_j = \langle f, \psi_j \rangle. \tag{2.34}$$

How do we choose the basis $(\psi_j)_{j=1}^M$ such that we can add diffusion consistently? It turns out that setting

$$\psi_j \coloneqq \varphi_j + \varepsilon \omega \cdot \nabla \varphi_j \tag{2.35}$$

does the trick. Expanding the matrix elements with this new basis yields

$$A_{i,j} = \langle \mu \nabla \varphi_i, \nabla \varphi_j \rangle + \varepsilon \langle \mu \nabla \varphi_i, \nabla(\omega \cdot \nabla \varphi_j) \rangle + \langle \omega \cdot \nabla \varphi_i, \varphi_j \rangle + \varepsilon \langle \omega \cdot \nabla \varphi_i, \omega \cdot \nabla \varphi_j \rangle \tag{2.36}$$

$$b_j = \langle f, \varphi_j \rangle + \varepsilon \langle f, \omega \cdot \nabla \varphi_j \rangle. \tag{2.37}$$

The terms not containing ε correspond to the standard Galerkin method. In order keep A a square matrix, we make sure $N = M$.

2.5 Error estimates

We consider two different error estimates. One being an estimate for the error in the standard Galerkin approximation. The other one is tailored for the Streamline diffusion / Petrov–Galerkin method.

Standard Galerkin Method

We employ the same trick as in section 1.4 on page 9 where we assume boundedness of a. Let u_h be the computed solution. Using the coercivity of the bilinear form a

and the Galerkin orthogonality property we get

$$\|e\|_1^2 \leq \frac{1}{\alpha} a(e,e) = \frac{1}{\alpha} a(e, u - v + v - u_h) \tag{2.38}$$

$$= \frac{1}{\alpha} a(e, u - v) \leq \frac{C}{\alpha} \|e\|_1 \|u - v\|_1 \tag{2.39}$$

for all $v \in \hat{V}$. Dividing both sides yield

$$\|e\|_1 \leq \frac{C}{\alpha} \|u - v\|_1. \tag{2.40}$$

The Bramble–Hilbert lemma yields a bound on the interpolation error of a certain type of interpolation operator, denoted $\pi_{p,h}u$ of order p. This can be combined with the above error estimate to yield

$$\|e\|_1 \leq \frac{C}{\alpha} \|u - \pi_{p,h}u\|_1 \leq \frac{CB}{\alpha} \|h^p u\|_{p+1}. \tag{2.41}$$

Recall that the constant α, coming from the coercivity of a, is given as $\alpha = \mu(1 - D_\omega)$. If μ is very small, i.e., in convection dominated problems, then our error bound becomes very bad. This can be fixed by looking at a more specifically tailored error estimate.

Petrov–Galerkin method

Introduce the *SUPG-norm* defined as follows:

$$\|u\|_{\text{SUPG}} \coloneqq \left(h \|\omega \cdot \nabla u\|^2 + \mu |\nabla u|^2 \right)^{1/2} \tag{2.42}$$

It turns out that solving the Petrov–Galerkin problem on a finite element space of order 1 with the same assumptions as above, then

$$\|u - u_h\|_{\text{SUPG}} \leq C h^{3/2} \|u\|_2. \tag{2.43}$$

This is stated without proof.

Discretization of Stokes Problem

3

Problem. 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 circumvent the inf-sup condition by stabilization.

3.1 Finite Element Formulation

The Stokes problem deals with the flow of incompressible Newtonian fluids slowly moving in a domain $\Omega \subseteq \mathbb{R}^n$. The strong formulation of the problem is given as:

$$-\Delta u + \nabla p = f \text{ in } \Omega, \tag{3.1}$$

$$\nabla \cdot u = 0 \text{ in } \Omega, \tag{3.2}$$

$$u = g \text{ on } \Gamma_D, \tag{3.3}$$

$$\frac{\partial u}{\partial n} - pn = h \text{ on } \Gamma_N. \tag{3.4}$$

Here $p : \Omega \to \mathbb{R}$ denotes the fluid pressure, $u : \Omega \to \mathbb{R}^n$ denotes the fluid velocity. The body force is denoted f. Note that this problem has two unknowns, namely u

and p. To this strong formulation we associate the two bilinear forms $a\colon V \times V \to \mathbb{R}$ and $b\colon W \times V \to \mathbb{R}$ given by

$$a(u, v) \coloneqq \langle \nabla u, \nabla v \rangle, \tag{3.5}$$

$$b(p, v) \coloneqq \langle p, \nabla \cdot v \rangle. \tag{3.6}$$

In addition, we also introduce the linear form $L\colon V \to \mathbb{R}$ given by

$$L(v) \coloneqq \langle f, v \rangle + \int_{\Gamma_N} hv \, dS. \tag{3.7}$$

Weak Formulation

We can therefore instead consider the weak form of the Stokes problem, namely: Find $u \in V$ and $p \in W$ such that

$$a(u, v) + b(p, v) = f(v) \text{ for all } v \in \hat{V}, \tag{3.8}$$

$$b(q, u) = 0 \text{ for all } q \in \hat{W}. \tag{3.9}$$

Again we need to precicely state what the trial spaces V, W and respective test spaces are. We require u to equal g on the Dirichlet boundary, while v should vanish on the Dirichlet boundary. Furthermore, we require only one derivative in u and v and no derivatives in p. We therefore decide on the spaces

$$V \coloneqq H^1_g(\Omega) \subseteq H^1(\Omega), \qquad \hat{V} \coloneqq H^1_0(\Omega) \subseteq H^1(\Omega), \tag{3.10}$$

$$W \coloneqq L^2(\Omega), \qquad \hat{W} \coloneqq L^2_0(\Omega) \subseteq L^2(\Omega). \tag{3.11}$$

Finite Element Formulation

In order to compute with these spaces we need to introduce a basis. Assume that the finite dimensional velocity space V_h is spanned by basis elements $(\varphi_i)_{i=1}^N$ and that the finite dimensional pressure space W_h is spanned by basis elements $(\psi_j)_{j=1}^M$. Making the ansatz

$$u = \sum_{i=1}^{N} c_i \varphi_i, \qquad p = \sum_{j=1}^{M} d_i \psi_j \tag{3.12}$$

we end up with the linear system

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{B} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ 0 \end{bmatrix}, \tag{3.13}$$

where the matrix elements are given as

$$A_{i,j} = \langle \nabla \varphi_i, \nabla \varphi_j \rangle, \tag{3.14}$$

$$B_{i,j} = \langle \psi_i, \nabla \varphi_j \rangle, \tag{3.15}$$

$$b_j = \langle f, \varphi_j \rangle + \int_{\Gamma_N} h\varphi_j \, dS. \tag{3.16}$$

3.2 Well Posedness of Weak Formulation

In the abstract setting, the Stokes problem is a mixed saddle point problem. There is a set of four conditions — analogous to the Lax–Milgram theorem used in the Poisson and Convection-Diffusion problems — for the Stokes problem that ensure the existence and uniqueness of the continuous solution. In the following, we assume that the Dirichlet boundary condition can be reduced to a homogenous one, in order to be able to apply the Poincaré inequality.

Boundedness of a: Applying the Cauchy–Schwartz inequality on a yields:

$$a(u, v) = \langle \nabla u, \nabla v \rangle \leq \| \nabla u \|_0 \| \nabla v \|_0. \tag{3.17}$$

Noting that the L^2-norm of the gradient is always smaller than the H^1 norm, the condition holds.

Boundedness of b: Applying the Cauchy–Schwartz inequality on b yields:

$$b(p, v) = \langle p, \nabla \cdot v \rangle \leq \|p\|_0 \| \nabla \cdot v \|_0. \tag{3.18}$$

To show boundedness, it suffices to show that $\| \nabla \cdot v \|_0 \leq \|v\|_1$. To this end, note that

$$\| \nabla \cdot v \|_0 = \left(\int_\Omega \sum_{i=1}^{n} \left(\frac{\partial v_i}{\partial x_i} \right)^2 d\Omega \right)^{1/2}. \tag{3.19}$$

This is merely a subset of the terms occuring in the expression for $(\nabla u)^2$, hence we can conclude outright that $\| \nabla v\|_0 \leq \|v\|_1$, and consequently b is bounded.

Coercivity of a: Start by noting that $\|u\|_1^2 = \|u\|_0^2 + \| \nabla u\|_0^1$. By the Poincaré inequality, we have that this satisfies

$$\|u\|_1^2 \leq (C^2 + 1)|u|_1^2. \tag{3.20}$$

Furthermore, we have that $|u|_1^2 = a(u, u)$, and hence

$$a(u, u) \geq \frac{1}{C^2 + 1}\|u\|_1^2 \tag{3.21}$$

which shows that a is indeed coercive.

The inf-sup **condition:** In order for the discretized Stokes problem to be well posed, we also need to satisfy the inf-sup condition, namely that

$$\sup_{v \in \hat{V}} \frac{b(q, v)}{\|v\|_1} > K\|q\|_0 \text{ for all } q \in \hat{W}. \tag{3.22}$$

This can be thought of as the "coercivity" of b. This condition ensures that B is surjective, or that B^T is injective. This in turn ensures that the solution exists and is unique. We will not show that this holds for the Stokes problem, as it is tricky.

3.3 Oscillations in the pressure

Consider the matrix equations in equation (3.13) on page 20. Writing these out yields the system of two equations

$$\boldsymbol{A}\boldsymbol{c} + \boldsymbol{B}^T\boldsymbol{d} = \boldsymbol{b}, \tag{3.23}$$

$$\boldsymbol{B}\boldsymbol{c} = 0. \tag{3.24}$$

Recall that $\boldsymbol{c}$ are the degrees of freedom for the velocity, and that $\boldsymbol{d}$ are the degrees of freedom for the pressure. Since $a \rightsquigarrow \boldsymbol{A}$ we have that $\boldsymbol{A}$ is invertible. Under the assumption that $\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T$ is invertible we can solve for the pressure $\boldsymbol{d}$, yielding

$$\boldsymbol{d} = (\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{b}. \tag{3.25}$$

Under what cirumstances is $\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T$ invertible? Since $\boldsymbol{A}$ is invertible, we need only consider whether $\boldsymbol{B}\boldsymbol{B}^T$ is invertible. This is equivalent to verifying that $\mathrm{Ker}(\boldsymbol{B}) = 0$, and it is exactly this the inf-sup condition in equation (3.22) on page 21 gives sufficient conditions for. Hence if this is not satisfied, we may be solving a non-invertible system.

If n and m denotes the number of degrees of freedom for the velocity and the pressure respectively, the block matrix equation is non-singular if n is sufficiently large compared to m. This is because the zero-matrix is of dimension $m \times m$. Having $n \gg m$ ensures that this is comparatively small.

3.4 Error Estimates

In the following, set $e_u \coloneqq u - u_h$ and $e_p \coloneqq p - p_h$. For finite element pairs that satisfy the inf-sup condition equation (3.22) we have an error estimate that reads

$$\|e_u\|_1 + \|e_p\|_1 \leq Ch^k \|u\|_{k+1} + Dh^{\ell+1}\|p\|_{\ell+1} \tag{3.26}$$

where k and ℓ denotes the polynomial degree of the velocity and pressure. Finding finite element pairs that satisfies the conditions for this error estimate is a difficult task. Below are a few examples of such finite element pairs:

The Taylor–Hood element: This element pair consists of a quadratic element for the velocity, and a linear element for the pressure. This yields the error estimate

$$\|e_u\|_1 + \|e_p\|_1 \leq h^2 \left(C\|u\|_3 + D\|p\|_2\right). \tag{3.27}$$

The Crouzeix–Raviart element: This element consists of a linear element in velocity, and a constant element in pressure, yielding the error estimate:

$$\|e_u\|_1 + \|e_p\|_1 \leq h^1 \left(C\|u\|_2 + D\|p\|_1\right) \tag{3.28}$$

The Mini element: This element employs linear elements in both velocity and pressure, however the velocity element also contains a cubic bubble in order to yield enough degrees of freedom to satisfy the inf-sup condition. We get the error estimate

$$\|e_u\|_1 + \|e_p\|_1 \leq Ch^1 \|u\|_2 + Dh^2 \|p\|_2 \tag{3.29}$$

3.5 Stabilization Techniques

Instead of solving the system given in equation (3.13), we may solve an alternative system given as

$$\boldsymbol{A}\boldsymbol{c} + \boldsymbol{B}^T\boldsymbol{d} = \boldsymbol{b}, \tag{3.30}$$

$$\boldsymbol{B}\boldsymbol{c} - \varepsilon\boldsymbol{D}\boldsymbol{d} = \varepsilon d. \tag{3.31}$$

This alternative, perturbed, system has coupled the solution of $\boldsymbol{c}$ to the solution of $\boldsymbol{d}$ in a way that lets us control the coupling, through both the matrix $\boldsymbol{D}$ which we have yet to define, and the parameter ε. Solving this for the pressure, we get:

$$\boldsymbol{d} = (-\varepsilon D)^{-1}(\varepsilon d - \boldsymbol{B}\boldsymbol{c}) \tag{3.32}$$

Using this to solve for the velocity $\boldsymbol{c}$:

$$\boldsymbol{c} = (\boldsymbol{d} + \boldsymbol{D}^{-1}d)\left(\boldsymbol{A} + \frac{1}{\varepsilon}\boldsymbol{B}^T\boldsymbol{D}^{-1}\boldsymbol{B}\right)^{-1} \tag{3.33}$$

provided that the matrix on the right is infact invertible. It can be verified that is is by noting that $\boldsymbol{A}$ is positive by construction, and $\boldsymbol{B}^T\boldsymbol{D}^{-1}B$ is positive if $\boldsymbol{D}$ is positive. Hence, in chosing $\boldsymbol{D}$ we need to make sure it is infact positive. Three choices for $\boldsymbol{D}$ are proposed, all based on perturbed versions of the equation of continuity $\nabla\cdot u = 0$:

(i) Setting $\boldsymbol{D} \coloneqq \boldsymbol{A}$, this corresponds to pressure stabilization where $\nabla\cdot v = \varepsilon\,\Delta\, p$;

(ii) setting $\boldsymbol{D} \coloneqq \boldsymbol{M}$, corresponding to the penality method where $\nabla\cdot v = \varepsilon p$; or

(iii) setting $\boldsymbol{D} \coloneqq (1/\Delta t)M$ corresponding to artificial compressibility, $\nabla\cdot v = -\varepsilon(\partial p/\partial t)$.

A problem with such techniques lies in the choice of the parameter ε. Choosing ε too small, pressure oscillations occur in the solution. Choosing ε too large the accuracy of the solution deteriorates.

Discretization Of Navier–Stokes

4

Problem. Explain the difference between operator splitting and algebraic splitting in the context of incompressible Navier–Stokes equations. Show disadvantages of operator splitting schemes associated with boundary conditions. Explain the advantage with operator splitting schemes as compared to algebraic splitting in the context of Richardson iteration and spectral equivalence.

Recall that the incompressible Navier–Stokes equations are given by

$$\frac{\partial v}{\partial t} + v \cdot \nabla v = -\frac{1}{\varrho} \nabla p + \nu \nabla^2 v + g, \tag{4.1}$$

with the equation of continuity, or *incompressibility constraint*,

$$\nabla \cdot v = 0. \tag{4.2}$$

Here v is the velocity field, p is the pressure, ϱ is the fluid density, and g is an umbrella term describing body forces.

4.1 Operator Splitting

The basic operator splitting algorithm can be summarized in three steps. These steps will be motivated in the following.

(i) Compute the velocity prediction $v^\star$ from an explicit equation involving the previous velocity v^n.

(ii) Solve the obtained Poisson equation for the pressure difference in time.

(iii) Compute the new velocity v^{n+1} and the new pressure p^{n+1} from explicit equations.

Explicit Scheme

Starting with a forward step in time in equation (4.1):

$$v^{n+1} = v^n - \mathrm{d}t \left(v^n + \frac{1}{\varrho} \nabla p^n - \nu \nabla^2 v^n - g^n \right). \tag{4.3}$$

This cannot be the velocity at time level $n+1$ because it does not satisfy the equation of continuity equation (4.2), i.e., $\nabla \cdot v^{n+1} \neq 0$. We instead use this as an intermediate step, denoted $v^\star$, in computing the new velocity. We may attempt to use the incompressibility constraint to compute a correction term v^{c} such that $v^{n+1} = v^\star + v^{\mathrm{c}}$. In order to gain some control over the pressure data in the equation, we also introduce the factor β to the pressure term:

$$v^\star = v^n - \mathrm{d}t \left(v^n + \frac{\beta}{\varrho} \nabla p^n - \nu \nabla^2 v^n - g^n \right). \tag{4.4}$$

The computed velocity v^{n+1} should also solve equation (4.3) where the pressure is evaluated at time level $n+1$. That is

$$v^{n+1} = v^n - \mathrm{d}t \left(v^n + \frac{1}{\varrho} \nabla p^{n+1} - \nu \nabla^2 v^n - g^n \right). \tag{4.5}$$

Subtracting these two equations yield an expression for v^{c}:

$$v^{\mathrm{c}} \coloneqq v^{n+1} - v^\star = -\frac{\mathrm{d}t}{\varrho} \nabla \left(p^{n+1} - \beta p^n \right), \tag{4.6}$$

or equivalently:

$$v^{n+1} = v^\star - v^{\mathrm{c}} = v^\star - \frac{\mathrm{d}t}{\varrho} \nabla \left(p^{n+1} - \beta p^n \right). \tag{4.7}$$

Setting $\varphi \coloneqq p^{n+1} - \beta p^n$ and requiring that $\nabla \cdot v^{n+1} = 0$ yields a Poisson equation in φ:

$$\nabla^2 \varphi = \frac{\varrho}{\mathrm{d}t} \nabla \cdot v^\star. \tag{4.8}$$

After solving this for φ we may update our solution in both velocity and pressure as follows:

$$p^{n+1} = \beta p + \varphi, \tag{4.9}$$

$$v^{n+1} = v^\star - \frac{\mathrm{d}t}{\varrho} \nabla \varphi. \tag{4.10}$$

Boundary Conditions

A question arises, namely: How do we solve the Poisson equation for the pressure difference φ? We are short on boundary conditions. Two remedies are proposed:

(i) Computing $\partial p/\partial n$ from equation (4.1) by multiplying the equation by the unit normal vector. This gives us Neumann boundary conditions for the pressure difference, $\partial\varphi/\partial n$.

(ii) Since v^{n+1} is supposed to satisfy the Dirichlet boundary conditions, then from equation (4.9) we must have

$$\nabla \varphi|_{\partial\Omega} = \frac{\mathrm{d}t}{\varrho}(v^{n+1} - v^\star)|_{\partial\Omega} = 0. \tag{4.11}$$

Hence φ must be constant on the boundary.

Implicit Scheme

In the above derivations an explicit scheme was used for the stepping. It is perfectly reasonable to instead use implicit schemes, like a θ-scheme. This will generally lead to a solution that involves an advection-diffusion equation, a Poisson equation, and then performing two implicit updates of the velocity and pressure. We will not go into detail here.

4.2 Algebraic Splitting

Note that in the operator splitting scheme, we discretize in time before discretizing in space. This leads to the need for more boundary conditions in order to solve

the Poisson-problem for the pressure difference. An alternative approach discretizes in space before we discretize in time. This removes the need for construction of additional boundary conditions as these are baked into the algebraic constraints. Discretizing the spatial operators in equations (4.1) and (4.2) using for instance a finite element method yields a set of linear systems:

$$\boldsymbol{M}\dot{\boldsymbol{u}} + \boldsymbol{K}(\boldsymbol{u})\boldsymbol{u} = -\boldsymbol{Q}\boldsymbol{p} + \boldsymbol{A}\boldsymbol{u} + \boldsymbol{f}, \tag{4.12}$$

$$\boldsymbol{Q}^T\boldsymbol{u} = 0. \tag{4.13}$$

Explicit Schemes

We can employ the same methodology as in section 4.1. Compute a tentative velocity $\boldsymbol{u}^\star$ from

$$\boldsymbol{M}\boldsymbol{u}^\star = \boldsymbol{M}\boldsymbol{u}^n + \mathrm{d}t\left(-\boldsymbol{K}(\boldsymbol{u}^n)\boldsymbol{u}^n - \beta\boldsymbol{Q}\boldsymbol{p}^n + \boldsymbol{A}\boldsymbol{u}^n + \boldsymbol{f}^n\right). \tag{4.14}$$

Again, this tentative velocity does not necessarily satisfy the equation of continuity, $\boldsymbol{Q}^T\boldsymbol{u} = 0$. We use $\boldsymbol{u}^\star$ to compute a correction term $\boldsymbol{u}^{\mathrm{c}}$ such that $\boldsymbol{u}^{n+1} \coloneqq \boldsymbol{u}^\star + \boldsymbol{u}^{\mathrm{c}}$ satisfies $\boldsymbol{Q}^T\boldsymbol{u}^{n+1} = 0$. In analogous fashion to the operator splitting method we can formulate a discrete Poisson equation for the pressure difference φ:

$$\boldsymbol{Q}^T\boldsymbol{M}^{-1}\boldsymbol{Q}\varphi = \frac{1}{\mathrm{d}t}\boldsymbol{Q}^T\boldsymbol{u}^\star. \tag{4.15}$$

Solving this for φ we update the pressure and velocity for the next time level:

$$\boldsymbol{p}^{n+1} = \beta\boldsymbol{p}^n + \varphi, \tag{4.16}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^\star - \mathrm{d}t\boldsymbol{M}^{-1}\boldsymbol{Q}\varphi. \tag{4.17}$$

Classical Schemes

We may employ a more general implicit scheme, known as the θ-scheme. Here θ is a *blending parameter*, where $\theta = 0$ corresponds to the Backward Euler scheme, $\theta = 1/2$ corresponds to the Crank–Nicolson scheme, and $\theta = 0$ yields the explicit Forward Euler Scheme as above. It turns out that these classical schemes can be viewed in the light of iterative methods.

Discretizing in time yields what we will refer to as the *fully implicit system*:

$$\boldsymbol{N}\boldsymbol{u}^{n+1} + \mathrm{d}t\boldsymbol{Q}\boldsymbol{p}^{n+1} = \boldsymbol{q}, \tag{4.18}$$

$$\boldsymbol{Q}^T\boldsymbol{u}^{n+1} = 0. \tag{4.19}$$

Here we have abbreviated as follows:

$$\boldsymbol{N} \coloneqq \boldsymbol{M} + \theta\mathrm{d}t\boldsymbol{R}(\boldsymbol{u})^n, \tag{4.20}$$

$$\boldsymbol{R}(\boldsymbol{u}^n) \coloneqq \boldsymbol{K}(\boldsymbol{u}^n) - \boldsymbol{A}, \tag{4.21}$$

$$\boldsymbol{q} \coloneqq (\boldsymbol{M} - (1-\theta)\mathrm{d}t\boldsymbol{R}(\boldsymbol{u}^n))\boldsymbol{u}^n + \mathrm{d}t\boldsymbol{f}^n. \tag{4.22}$$

Assuming invertibility of $\boldsymbol{N}$ we can solve this for $\boldsymbol{u}^{n+1}$ — and inserting this into the second equation yields what is known as the *Schur complement pressure equation*:

$$\boldsymbol{Q}^T\boldsymbol{N}^{-1}\boldsymbol{Q}\boldsymbol{p}^{n+1} = \frac{1}{\mathrm{d}t}\boldsymbol{Q}^T\boldsymbol{N}^{-1}\boldsymbol{q}. \tag{4.23}$$

The system $\boldsymbol{N}^{-1}$ can often be computationally costly to solve. Since $\boldsymbol{N}$ is a sparse matrix, $\boldsymbol{N}^{-1}$ is dense, so we we would like some other way of dealing with this equation. Luckily, we can note that equation (4.23) is a linear system of the form

$$\boldsymbol{B}\boldsymbol{p}^{n+1} = \boldsymbol{b}. \tag{4.24}$$

A preconditioned Richardson iteration can be formulated for $\boldsymbol{p}^{n+1}$

$$\boldsymbol{p}^{n+1,k+1} = \boldsymbol{p}^{n+1,k} - \boldsymbol{C}_1^{-1}(\boldsymbol{B}\boldsymbol{p}^{n+1,k} - \boldsymbol{b}), \tag{4.25}$$

where $\boldsymbol{C}_1^{-1}$ is a preconditioner similar to $\boldsymbol{N}^{-1}$ however in some sense, simpler to solve, and k is an iteration counter. For each time level we start the iteration with $\boldsymbol{p}^{n+1,0} = \boldsymbol{p}^n$. For the Schur complement pressure equation this yields an iteration of the form

$$\boldsymbol{p}^{n+1,k+1} = \boldsymbol{p}^{n+1,k} - \boldsymbol{C}_1^{-1}(\boldsymbol{Q}^T\boldsymbol{N}^{-1}\boldsymbol{Q}\boldsymbol{p}^{n+1,k} - \frac{1}{\mathrm{d}t}\boldsymbol{Q}^T\boldsymbol{N}^{-1}\boldsymbol{q}). \tag{4.26}$$

Formulating a similar procedure for the velocity, with preconditioner $\boldsymbol{C}_2^{-1}$ we end up with a preconditioned system similar to the fully implicit system:

$$\begin{bmatrix} \boldsymbol{C}_1 & 0 \\ 0 & \boldsymbol{C}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{N} & \boldsymbol{Q} \\ \boldsymbol{Q}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{0} \end{bmatrix}. \tag{4.27}$$

Iterative Methods

5

Problem. Describe the Richardson iteration. Explain spectral equivalence and show how spectral equivalence may lead to convergence in a constant number of iterations. Explain what to expect for the Poisson problem when using Conjugate Gradient methods combined with Jacobi, ILU, and AMG based on experiments with FEniCS. How does it compare with direct methods?

In this section we will be dealing with linear systems of the form

$$Au = b, \tag{5.1}$$

where A is a $N \times N$ square matrix. In applications N might be very large, typically between 10^6 and 10^9, and is often sparse. For instance collocation matrices with basis functions that exhibit local support properties will typically only have $\mathcal{O}(N)$ nonzero elements. This means, that in general, solving systems of the form as in equation (5.1) by computing A^{-1} is computationally very expensive. In light of this we will examine alternative methods, namely *iteration methods*, and *preconditioning*.

5.1 The Richardson Iteration

The simplest iteration we consider is the *Richardson iteration*. This is given as

$$u^n = u^{n-1} - \tau \left(Au^{n-1} - b \right), \tag{5.2}$$

where τ is a relaxation parameter that gives us some control over the iteration. This value must be determined. The Richardson iteration is consistent in the sense that if

we happen to have converged to the correct solution, i.e., that $u^{n-1} = \boldsymbol{u}$, then $u^n = \boldsymbol{u}$ as well. Every iteration consists of a matrix-vector product, hence the procedure has a time complexity of $\mathcal{O}(n)$ FLOPS.

Error Analysis

Denote by $e^n \coloneqq u^n - u$ the error in the n-th iteration. Subtracting u from both sides of equation (5.2) yields

$$e^n = e^{n-1} - \tau A e^{n-1}. \tag{5.3}$$

We can analyze this iterative error in terms of the L^2-norm, which we in the following denote by $\|\cdot\|$:

$$\|e^n\| = \|e^{n-1} - \tau A e^{n-1}\| \leq \|I - \tau A\|_M \|e^{n-1}\|. \tag{5.4}$$

Here $\|\cdot\|_M$ denotes the induced matrix norm. If $\|I - \tau A\|_M < 1$, then the iteration converges to the exact solution. The question now is, how do we ensure this. And in the case of convergence, what is the convergence rate? For a symmetric and positive definite matrix A, the matrix norm of A, defined as

$$\|A\|_M \coloneqq \max_x \frac{\|Ax\|}{\|x\|}, \tag{5.5}$$

is equal to the largest eigenvalue of A, denoted $\lambda_{\max}$. We can use this fact to discuss the norm of $I - \tau A$. We have that

$$\|I - \tau A\|_M = \max_x \frac{\|(I - \tau A)x\|}{\|x\|}. \tag{5.6}$$

We can find the optimal relaxation parameter τ by noting that the minimum value for $\|I - \tau A\|_M$ is attained when $(1 - \tau\lambda_{\min}) = -(1 - \tau\lambda_{\max})$. Solving for τ yields

$$\tau = \frac{2}{\lambda_{\max} + \lambda_{\min}}. \tag{5.7}$$

We denote this by τ_{optimal}. The matrix norm of $I - \tau_{\text{optimal}} A$ is equal to its largest eigenvalue, and by choice of the relaxation parameter, we have that the largest

eigenvalue and the smallest eigenvalue is equal, hence

$$\|I - \tau_{\text{optimal}} A\|_M = 1 - \tau_{\text{optimal}} \lambda_{\min} = 1 - \frac{2\lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}. \tag{5.8}$$

Recalling that the *condition number* κ of a matrix A is the largest eigenvalue divided by the smallest, we have from the above that

$$\|I - \tau_{\text{optimal}} A\|_M = \frac{\kappa - 1}{\kappa + 1}, \tag{5.9}$$

and note that this number is always strictly smaller than one. We have therefore shown that the Richardson iteration converges, and that the rate of convergence *depends* on the eigenvalues of the matrix A.

Iteration Stopping Criteria

In order for this iteration method to be useful, we need to somehow know how many iterations to perform until a certain error tolerance has been met. Assuming we want to reduce the error by a factor ε, i.e., we require $\|e^n\|/\|e^0\| \leq \varepsilon$. To this end, recall that

$$\|e^n\| \leq \frac{\kappa - 1}{\kappa + 1} \|e^{n-1}\|. \tag{5.10}$$

Repeatedly applying one iteration yields

$$\|e^n\| \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^n \|e^0\|. \tag{5.11}$$

Dividing by $\|e^0\|$ and requiring this to be smaller than ε yields

$$\frac{\|e^n\|}{\|e^0\|} \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^n \leq \varepsilon. \tag{5.12}$$

We can solve for n which yields

$$n \geq \frac{\log(\varepsilon)}{\log(\kappa - 1) - \log(\kappa + 1)}, \tag{5.13}$$

where the inequality sign is flipped due to dividing by something negative. So, we could for instance set n to be the integer ceiling of this value to ensure that our error ratio is as small as we wanted:

$$n \coloneqq \left\lceil \frac{\log(\varepsilon)}{\log(\kappa - 1) - \log(\kappa + 1)} \right\rceil. \tag{5.14}$$

Note now that our required number of iterations n is a function of the condition number κ, and the condition number is entirely dependent on the eigenvalues of the matrix A. In applications we often have mesh dependent eigenvalues which means that our error estimates change depending on what type of mesh we choose. One example is particularity enlightening, namely that of the Poisson equation on the open domain $(0,1) \subset \mathbb{R}$ where the corresponding eigenvalues of A are

$$\lambda_i = \frac{4}{h^2} \sin^2\left(\frac{\pi i h}{2}\right)$$

yielding $\lambda_{\min} = \pi^2$ and $\lambda_{\max} = 4/h^2$. The corresponding condition number is then $\kappa = 4/(\pi h)^2$. This means that by refining the mesh leads to the need for *more* iterations until convergence, which is *not* a good trait.

In order to remedy this we turn to the notion of *spectral equivalence* and the method of *preconditioning*.

5.2 Spectral Equivalence and Preconditioning

The idea of preconditioning is to instead of solving the system $Au = b$ by inverting A, solve the system $BAu = Bb$, where B is some suitable matrix, called the *preconditioner*, that is both easy to store and easy to compute. The defining criteria is that the matrix BA should have a smaller condition number than the matrix A. Performing the same analysis as in section 5.1 on page 30, we see that the error in the n-th Richardson iteration of this new system is

$$e^n = e^{n-1} - \tau BAe^{n-1}. \tag{5.15}$$

Consequently, the iteration converges if $\|I - \tau BA\| < 1$. We list some criteria for choosing a preconditioner:

1. The evaluation of B on a vector should be $\mathcal{O}(N)$,
2. the storage of B should be $\mathcal{O}(N)$, and
3. the matrix B should be spectrally equivalent with A^{-1}.

The notion of *spectral equivalence* is defined as follows:

Definition 1 (Spectral Equivalence)**.** Two symmetric and positive definite linear operators A^{-1} and B are called *spectrally equivalent* if there exists constants c_1 and c_2 such that

$$c_1\langle A^{-1}v, v\rangle \leq \langle Bv, v\rangle \leq c_2\langle A^{-1}v, v\rangle \tag{5.16}$$

for all v. If A^{-1} and B are spectrally equivalent, then the condition number κ of the matrix BA is bounded as $\kappa \leq c_2/c_1$.

If we choose a matrix B spectrally equivalent matrix to A^{-1}, we know that the Richardson iteration is order optimal, as the condition number is bounded independently of the discretization.

5.3 Krylov Methods

The Richardson iteration discussed above is a linear iteration. It turns out that any alternative linear iteration method can be written as a Richardson iteration with a preconditioner. There are however non-linear iteration methods, where, for instance, the need to determine the relaxation parameter τ beforehand is removed. Some honorable mentions:

1. The Conjugate Gradient method — Used when the matrix is symmetric and positive definite. Requires a symmetric and positive definite preconditiner;
2. the Minimal Residual method — Used when the matrix is symmetric but indefinite. Also requires a symmetric and positive definite preconditioner;
3. GMRES with either ILU or AMG — Used for positive matrices, i.e., in convection-diffusion problems.
4. BiCGStab / GMRES — Used for nonsymmetric and indefinite matrices.