Lecture Note #11 **MTH652: Advanced Numerical Analysis**
Winter 2025 Instructor: Will Pazner

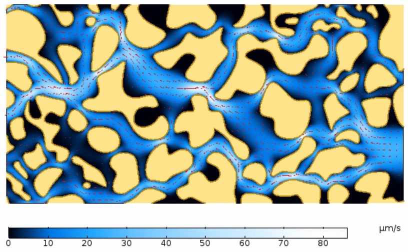
Saddle-Point Systems and Iterative Methods

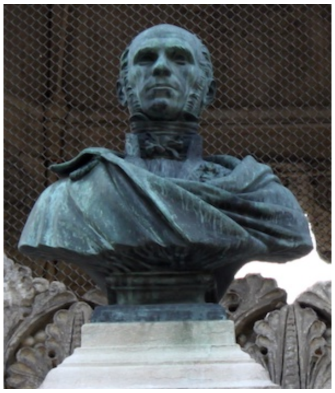
1 Darcy's Law

Last time, we considered finite element discretizations for the Stokes equations

$$\begin{aligned} -\Delta \boldsymbol{u} + \nabla p &= \boldsymbol{f}, \\ \nabla \cdot \boldsymbol{u} &= 0, \\ \boldsymbol{u} &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

Using the inf–sup condition, we showed that this problem is well-posed. In order to obtain convergent finite element methods, we require the inf–sup condition to hold at the discrete level. We gave an example ($\mathcal{P}^1$–$\mathcal{P}^0$) where this doesn't hold (the matrix will not be invertible) We also gave an example ($\mathcal{P}^2$–$\mathcal{P}^0$) that does work; the inf–sup condition can be verified by constructing a Fortin operator. There are many other stable choices of spaces for Stokes. One of the most common is the "Taylor–Hood" element, which is simply $\mathcal{P}^k$–$\mathcal{P}^{k-1}$ for $k \geq 2$.





The theory is of course not limited just to Stokes. We give another example of a set of equations that give rise to mixed methods that necessitate inf–sup stability. Consider fluid flow through a porous medium. For example, consider the flow of water through a bed of sand.

Henry Darcy established (experimentally) the relation between the *flux* (i.e. *fluid velocity*) and the pressure. This is known as Darcy's law.

This relationship is given by

$$\boldsymbol{u} = -\frac{k}{\mu}\nabla p,$$

where $\boldsymbol{u}$ is the velocity, k is the permeability tensor, and μ is the viscosity. (It is of course possible to add forcing terms to the right-hand side).

Many fluids are *incompressible* (in reality, very well-approximated as incompressible). This means that the density of a given "parcel" of fluid remains constant throughout the flow. This can be written as

$$\nabla \cdot \boldsymbol{u} = 0.$$

If we consider a "control volume" V, divergence theorem implies that

$$\int_V \nabla \cdot \boldsymbol{u} \, dx = \int_{\partial V} \boldsymbol{u} \cdot \boldsymbol{n} \, ds.$$

This relationship tells us that the fluid flowing **into** the control volume has to be exactly balanced by the fluid flowing **out** of the control volume (and hence the density will remain constant).

Combining Darcy's law with incompressibility, we obtain the following equation for the pressure:

$$-\nabla \cdot \left(\frac{k}{\mu} \nabla p \right) = 0.$$

Note that this is essentially a Poisson problem for the pressure. For such problems, we have the concept of a **conservative flux approximation**. Note that testing the equation

$$-\Delta p = f$$

with a constant test function $v \equiv 1$ and integrating over a control volume gives that

$$-\int_{\partial V} \nabla p \cdot \boldsymbol{n} \, ds = \int_V f \, dx.$$

(In the context of e.g. Darcy flow, this means that the velocity is *locally incompressible*). However, our standard finite element approximation **does not** satisfy this conservation property. One way to recover finite elements method that are conservative is to use **mixed methods**.

Consider the first order system

$$\begin{aligned} \boldsymbol{u} + \nabla p &= 0 \\ \nabla \cdot \boldsymbol{u} &= g \end{aligned}$$

with boundary conditions $p = 0$ on $\partial\Omega$. This is equivalent to the Poisson problem $-\Delta p = g$. Without being precise about the test and trial spaces, we could try to formulate the variational formulation for this problem by multiplying the first and second equations by test functions, and integrating one of the terms by parts.

Let $\boldsymbol{v}$ be a vector-valued test function, and let q be a scalar test function. Then, we obtain

$$\begin{aligned} \int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} \, dx + \int_\Omega \nabla p \cdot \boldsymbol{v} \, dx &= 0 \\ \int_\Omega (\nabla \cdot \boldsymbol{u}) q \, dx &= \int_\Omega g q \, dx \end{aligned}$$

Integrating the second term in the first equation by parts and multiply the second equation by -1, we have

$$\begin{aligned} \int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} \, dx - \int_\Omega (\nabla \cdot \boldsymbol{v}) p \, dx &= 0 \\ -\int_\Omega (\nabla \cdot \boldsymbol{u}) q \, dx &= -\int_\Omega g q \, dx \end{aligned}$$

(we use a homogeneous Dirichlet condition of p to omit the boundary term). Once we specify the spaces $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}$ and $p, q \in P$, we have a mixed variational problem. The previously developed theory of inf–sup stability will tell us about well-posedness of this problem. In what follows, we will consider a slightly generalized version of this problem by introducing a forcing term $\boldsymbol{f}$ on the first equation,

$$\begin{aligned} \boldsymbol{u} + \nabla p &= \boldsymbol{f} \\ \nabla \cdot \boldsymbol{u} &= g. \end{aligned}$$

Note that in the case of the standard Galerkin formulation for the Poisson problem, we chose the space $H^1(\Omega)$, which is the minimal regularity space to make sense of the bilinear forms. *Let's do the same thing here:*

- $\boldsymbol{u}$ and $\boldsymbol{v}$ need to be square-integrable and to have well-defined square-integrable divergence
- p and q need to be square-integrable

This means we choose $p, q \in L^2(\Omega)$ and

$$\boldsymbol{u}, \boldsymbol{v} \in \left\{ \boldsymbol{w} : \boldsymbol{w} \in \boldsymbol{L}^2(\Omega) \text{ and } \nabla \cdot \boldsymbol{w} \in L^2(\Omega) \right\}$$

This is a Sobolev-type space called $\boldsymbol{H}(\text{div}, \Omega)$. The natural norm for this space is

$$\|\boldsymbol{v}\|^2_{H(\text{div},\Omega)} := \|\boldsymbol{v}\|^2_{L^2(\Omega)} + \|\nabla \cdot \boldsymbol{v}\|^2_{L^2(\Omega)}$$

Note that the **normal trace** of functions in $H(\text{div}, \Omega)$ can be defined (using integration by parts), and such functions possess **normal continuity** across subdomains.

Let's define some bilinear forms

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) &= \int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} \, dx \\ b(\boldsymbol{u}, q) &= -\int_\Omega (\nabla \cdot \boldsymbol{u}) q \, dx \end{aligned}$$

Let $\boldsymbol{V} = \boldsymbol{H}(\text{div}, \Omega)$ and $P = L^2(\Omega)$. Then, the variational problem can be written as: find $(\boldsymbol{u}, p) \in \boldsymbol{V} \times P$ such that

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) &= F(\boldsymbol{v}) \\ b(\boldsymbol{u}, q) &= G(q) \end{aligned}$$

for all test functions $(\boldsymbol{v}, q) \in \boldsymbol{V} \times P$. Here, the linear forms are defined by

$$\begin{aligned} F(\boldsymbol{v}) &= \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, dx \\ G(q) &= \int_\Omega g q \, dx. \end{aligned}$$

To ensure that the mixed problem is well-posed, we can check the following conditions:

(i) $a(\cdot\,,\cdot)$ and $b(\cdot\,,\cdot)$ are continuous

(ii) $F(\cdot)$ and $G(\cdot)$ are bounded

(iii) $a(\cdot\,,\cdot)$ is coercive on the nullspace of $b(\cdot\,,\cdot)$

(iv) $b(\cdot\,,\cdot)$ satisfies the inf–sup condition

First, we note that the bilinear forms are continuous (why?):

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{v}) &\lesssim \|\boldsymbol{u}\|_{\boldsymbol{V}}\|\boldsymbol{v}\|_{\boldsymbol{V}} \\ b(\boldsymbol{u},q) &\lesssim \|\boldsymbol{u}\|_{\boldsymbol{V}}\|q\|_{P} \end{aligned}$$

Similarly, the linear forms F and G are bounded.

The nullspace Z of $b(\cdot\,,\cdot)$ consists of those function $\boldsymbol{v} \in \boldsymbol{V}$ such that $b(\boldsymbol{v},q) = -(\nabla\cdot\boldsymbol{v},q) = 0$ for all $q \in P$. Setting $q = \nabla\cdot\boldsymbol{v}$ shows that

$$Z = \{\boldsymbol{v} \in \boldsymbol{V} : \nabla\cdot\boldsymbol{v} = 0\}.$$

For any $\boldsymbol{v} \in Z$,

$$\|\boldsymbol{v}\|^2_{H(\mathrm{div},\Omega)} = \|\boldsymbol{v}\|^2_{L^2(\Omega)} + \|\nabla\cdot\boldsymbol{v}\|^2_{L^2(\Omega)} = \|\boldsymbol{v}\|^2_{L^2(\Omega)} = a(\boldsymbol{v},\boldsymbol{v}),$$

showing coercivity of $a(\cdot\,,\cdot)$ on Z (with constant 1).

It remains to show that the inf–sup condition

$$\inf_{q\in P}\sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\boldsymbol{V}}\|q\|_P} \geq \beta > 0$$

is satisfied. The argument is the same as for Stokes. Let $q \in P$ be arbitrary. We have showed that there exists $\boldsymbol{v} \in \boldsymbol{V}$ such that $\nabla\cdot\boldsymbol{v} = q$, and $\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)} \lesssim \|q\|_{L^2(\Omega)}$. We conclude using that $\|\boldsymbol{v}\|^2_{H(\mathrm{div},\Omega)} \lesssim \|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)}$.

At the discrete level, we need to choose inf–sup-stable pairs of spaces $\boldsymbol{V}_h \subseteq \boldsymbol{V}$ and $P_h \subseteq P$. It is possible to construct some different $\boldsymbol{H}(\mathrm{div},\Omega)$-conforming finite element spaces ("Raviart–Thomas" spaces are one of the standard ones); these constructions will be discussed later in this course.

2 Linear Algebra

After discretization of a variational problem

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},p) &= F(\boldsymbol{v}) \\ b(\boldsymbol{u},q) &= G(q) \end{aligned}$$

(whether it is Stokes or Darcy), we obtain a linear system of equations

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

The solution to this system is the **saddle point** of the functional

$$\mathcal{L}(u,p) = \frac{1}{2}u^TAu - f^Tu + (Bu-g)^Tp,$$

which is the Lagrangian associated with the minimization problem

$$\begin{aligned} \frac{1}{2}u^T Au - f^T u &\to \min \\ \text{subject to } Bu &= g \end{aligned}$$

Here, the variable p plays the role of the Lagrange multiplier.

In our case, we have that A is SPD and B is full row rank, which imply invertibility of $\mathcal{A}$. Clearly, the matrix

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$$

is symmetric. What can we say about its spectrum? Note that we can factorize $\mathcal{A}$ as follows

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & A^{-1}B^T \\ 0 & I \end{pmatrix}$$

where the **Schur complement** S is defined by

$$S = -BA^{-1}B^T.$$

Since A is symmetric, we have

$$\mathcal{A} = \mathcal{B}\mathcal{S}\mathcal{B}^T,$$

where $\mathcal{S} = \operatorname{blockdiag}(A, S)$. This means that $\mathcal{A}$ and $\mathcal{S}$ are **congruent matrices** (not the same thing as similar matrices). We reference (without proof) a result from linear algebra

Theorem 1 (Sylvester's Law of Intertia)**.** *Two symmetric matrices are congruent if and only if they have the same number of positive, zero, and negative eigenvalues.*

Since $\mathcal{A}$ and $\mathcal{S}$ are invertible, they both have no zero eigenvalues. We can easily count the number of positive and negative eigenvalues of $\mathcal{S}$, since it is block diagonal. The matrix A is SPD, and has only positive eigenvalues. It is clear from the form of the Schur complement S that it is **negative definite**, and hence has only negative eigenvalues. Therefore $\mathcal{S}$, and hence also $\mathcal{A}$, has n positive eigenvalues and m negative eigenvalues. In particular, $\mathcal{A}$ must necessarily be an **indefinite** matrix. Although it is symmetric, we cannot apply methods like conjugate gradient.

In order to iteratively solve $\mathcal{A}\boldsymbol{x} = \boldsymbol{b}$, we can try to use the same idea as GMRES: minimize the residual $\|\boldsymbol{b} - \mathcal{A}\boldsymbol{x}\|$ where x is taken from the mth Krylov subspace $\mathcal{K}_m(A, \boldsymbol{b})$. However, we can try to use symmetric of $\mathcal{A}$ to our advantage.

Recall from GMRES, the important **Arnoldi identity**

$$\mathcal{A}Q_m = Q_{m+1}H_m.$$

and multiply on the left by Q_m^T to obtain

$$Q_m^T \mathcal{A} Q_m = Q_m^T Q_{m+1} H_m =: H_m'.$$

Since $\mathcal{A}$ is symmetric, the left-hand side is symmetric, and so H'_m is also symmetric. By orthogonality of Q_m, we see that H'_m is upper Hessenberg (it is the first m rows of H_m). Since it is also symmetric, it must be tridiagonal, leading to the simplified relationship

$$\mathcal{A}\boldsymbol{q}_m = H_{m-1,m}\boldsymbol{q}_{m-1} + H_{mm}\boldsymbol{q}_m + H_{m+1,m}\boldsymbol{q}_{m+1}.$$

This is called the **Lanczos iteration**. In this situation, orthogonalizing $\mathcal{A}\boldsymbol{q}_m$ to find the new vector $\boldsymbol{q}_{m+1}$ requires only orthogonalizing against the **previous two** basis vectors; this gives rise to a **short-term recurrence**. The resulting method is called **MINRES**. No restarting is required, since the storage and per-iteration cost do not increase with further iterations.

The convergence of MINRES can be characterized in terms of the ratio of maximum and minimum (absolute) eigenvalues,

$$\kappa(\mathcal{A}) = \frac{|\lambda|_{\max}(\mathcal{A})}{|\lambda|_{\min}(\mathcal{A})}.$$

In the special case when the matrix is SPD (i.e. **not** applicable to the discussion of saddle-point systems), then the residuals satisfy the following estimate, which can be compared to the estimate from conjugate gradient

$$\|\boldsymbol{r}_m\| \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \|\boldsymbol{r}_0\|.$$

As in the case of CG, we can **precondition** MINRES; we cannot simply apply the same algorithm to $\mathcal{M}\mathcal{A}$ for a preconditioner $\mathcal{M}$, since this will destroy the symmetry of the problem. However, if $\mathcal{M}$ is SPD, we can consider the algorithm in the inner product induced by $\mathcal{M}$. A particularly good preconditioner for saddle-point problems is

$$\mathcal{M} = \begin{pmatrix} A^{-1} & 0 \\ 0 & -S^{-1} \end{pmatrix}.$$

Each of A^{-1} and $-S^{-1}$ may be replaced themselves by a spectrally equivalent preconditioner to A and the Schur complement, respectively.