Lecture Note #6 **MTH652: Advanced Numerical Analysis**
Winter 2025 Instructor: Will Pazner

Non-Symmetric Problems

1 Well-Posedness for Symmetric Problems

So far, in this course and in the previous term, we have been dealing with PDEs that give rise to symmetric variational formulations. The two main examples we have seen are the Poisson problem and the equations of linear elasticity.

The **Poisson problem**

$$-\Delta u = f$$

gives rise to the variational problem: find $u \in V$ such that

$$(\nabla u, \nabla v) = (f, v)$$

for all $v \in V$. The equations of **linear elasticity** (in displacement form) are:

$$-\nabla \cdot (2\mu \nabla^s \boldsymbol{u} + \lambda(\nabla \cdot \boldsymbol{u}) I) = \boldsymbol{f},$$

which gives rise to the variational problem: find $\boldsymbol{u} \in \boldsymbol{V}$ such that

$$(2\mu \nabla^s \boldsymbol{u}, \nabla^s \boldsymbol{v}) + (\lambda \nabla \cdot \boldsymbol{u}, \nabla \cdot \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v})$$

for all $\boldsymbol{v} \in \boldsymbol{V}$. In both cases above, we assume homogeneous Dirichlet boundary conditions for simplicity.

In each case, the variational problem can be written as: find $u \in V$ such that

$$a(u, v) = F(v) \tag{1}$$

for all $v \in V$, where $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is a bilinear form, and $F(\cdot) : V \to \mathbb{R}$ is a linear form. Written this way, we have

$$\begin{aligned} \text{Poisson:} \quad & a(u, v) := (\nabla u, \nabla v) \\ \text{Elasticity:} \quad & a(\boldsymbol{u}, \boldsymbol{v}) := (2\mu \nabla^s \boldsymbol{u}, \nabla^s \boldsymbol{u}) + (\lambda \nabla \cdot \boldsymbol{u}, \nabla \cdot \boldsymbol{v}) \end{aligned}$$

Checking well-posedness of the variational formulation and applicability of the Galerkin/ finite element method is a matter of verifying conditions (i), (ii), and (iii) from Section 2.1 of the textbook:

(i) $a(\cdot, \cdot)$ is symmetric, $a(u, v) = a(v, u)$

(ii) $a(\cdot, \cdot)$ is continuous, $a(u, v) \leq \gamma \|u\|_V \|v\|_V$

(iii) $a(\cdot, \cdot)$ is coercive (or V-elliptic), $a(u, u) \geq \alpha \|u\|_V^2$

Under these conditions, $a(\cdot, \cdot)$ defines an inner product that is equivalent (up to the constants γ and α) to the standard inner product on the Hilbert space V, and so the Riesz Representation Theorem ensures a unique solution to the variational problem (1) (assuming that the linear form F is also continuous).

2 Convection–Diffusion

We would like to relax condition (i) (symmetry), and consider some relevant examples of non-symmetric problems. The **convection–diffusion**(–reaction) equation (also known as advection–diffusion) is:

$$-\mu\Delta u + \boldsymbol{\beta}\cdot\nabla u + \varepsilon u = f,$$

for coefficients $\mu, \varepsilon > 0$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. The coefficient μ is called the *diffusion coefficient*, and represents the strength of diffusion relative to advection. The vector coefficient $\boldsymbol{\beta}$ is the *advection velocity*. The term εu is usually called the reaction term. This equation models the movement of a quantity (for example, concentration) that is advecting (i.e. moving in the direction of $\boldsymbol{\beta}$) and simultaneously diffusing (spreading out). Suppose that homogeneous Dirichlet boundary conditions are applied, i.e.

$$u = 0 \text{ on } \partial\Omega.$$

Then, the variational formulation can be obtained through the usual procedure by multiplying by a test function $v \in H^1_0(\Omega)$ and integrating the Laplace term by parts, obtaining: find $u \in H^1_0(\Omega)$ such that

$$\mu(\nabla u, \nabla v) + (\boldsymbol{\beta}\cdot\nabla u, v) + \varepsilon(u, v) = (f, v) \tag{2}$$

for all $v \in H^1_0(\Omega)$. While the diffusion and reaction terms are symmetric, the advection term is not symmetric. However, the bilinear form

$$a(u, v) = \mu(\nabla u, \nabla v) + (\boldsymbol{\beta}\cdot\nabla u, v) + \varepsilon(u, v)$$

is still continuous and coercive. Continuity can be shown straightforwardly using the standard arguments, and coercivity can be seen as follows (making some assumptions on the velocity $\boldsymbol{\beta}$). Suppose $\boldsymbol{\beta}$ is spatially constant. Then,

$$\begin{aligned} a(u, u) &= \mu(\nabla u, \nabla u) + (\boldsymbol{\beta}\cdot\nabla u, u) + \varepsilon(u, u) \\ &\gtrsim \|u\|^2_{H^1(\Omega)} + (\boldsymbol{\beta}\cdot\nabla u, u) \end{aligned}$$

Integrating the second term on the right-hand side by parts, we obtain

$$\begin{aligned} \int_\Omega (\boldsymbol{\beta}\cdot\nabla u) u \, dx &= -\int_\Omega u \nabla\cdot(\boldsymbol{\beta} u)\, dx + \int_{\partial\Omega} u^2 \boldsymbol{\beta}\cdot\boldsymbol{n}\, ds \\ &= -\int_\Omega u(\boldsymbol{\beta}\cdot\nabla u)\, dx, \end{aligned}$$

using that $u \in H^1_0(\Omega)$ vanishes on $\partial\Omega$ and $\nabla\cdot(\boldsymbol{\beta} u) = \boldsymbol{\beta}\cdot\nabla u$. Therefore,

$$(\boldsymbol{\beta}\cdot\nabla u, u) = -(\boldsymbol{\beta}\cdot\nabla u, u) = 0,$$

and coercivity of $a(\cdot,\cdot)$ follows. The assumption that $\boldsymbol{\beta}$ is spatially constant can be relaxed. Suppose instead that $\boldsymbol{\beta}$ is *divergence-free*, that is, $\nabla\cdot\boldsymbol{\beta} = 0$. In this case, we use the vector calculus identity (analogous to the product rule)

$$\nabla\cdot(\boldsymbol{\beta} u) = (\nabla\cdot\beta)u + \boldsymbol{\beta}\cdot\nabla u,$$

and the same conclusion follows.

Can we infer well-posedness of the variational problem given continuity and coercivity of $a(\cdot,\cdot)$, even without symmetry? The positive answer to this question is given by the **Lax–Milgram** theorem, which is a generalization of the Riesz Representation Theorem.

3 Lax–Milgram and Céa

Theorem 1 (Lax–Milgram)**.** *Let V be a Hilbert space, let $a : V \times V \to \mathbb{R}$ be a continuous and coercive bilinear form (with constants γ and α as above), and let $F : V \to \mathbb{R}$ be a continuous linear functional. Then, there exists a unique $u \in V$ such that*

$$a(u, v) = F(v)$$

for all $v \in V$.

Proof. Let V' denote the dual space of V (i.e. the space of continuous linear functionals defined on V, so that $F \in V'$). For any $u \in V$, define $A_u : V \to \mathbb{R}$ by

$$A_u : v \mapsto a(u, v).$$

Bilinearity of $a(\cdot,\cdot)$ implies that A_u is linear. Furthermore,

$$A_u(v) = a(u, v) \leq \gamma \|u\|_V \|v\|_V,$$

and so A_u is a *continuous* linear functional (with constant of continuity bounded by $\gamma\|u\|_V$), and hence $A_u \in V'$. The mapping $A : u \mapsto A_u$ is also a linear map from $V \to V'$.

Our goal is to find a (unique) $u \in V$ such that $a(u, v) = F(v)$ for all $v \in V$. Equivalently, $A_u(v) = F(v)$ for all v, which is the same as $A_u = F$ in the dual space V'. By the Riesz Representation Theorem, for every linear functional $\phi \in V'$ there is a unique element $\tau(\phi) \in V$ (its *Riesz representative*) such that $(\tau(\phi), v) = \phi(v)$ for all $v \in V$. Since this representation is unique, $A_u = F$ in V' is equivalent to $\tau(A_u) = \tau(F)$ in V. Note that the mapping $\tau : V' \to V$ is itself linear, and $\|\tau(\phi)\|_V = \|\phi\|_{V'}$ for all $\phi \in V'$.

Define the mapping $T : V \to V$ by

$$T : v \mapsto v - \rho(\tau(A_v) - \tau(F))$$

for some nonzero $\rho \in \mathbb{R}$. If we could show that T is a contraction mapping, then T would have a unique fixed point u, which satisfies $T(u) = u$, and hence $\rho(\tau(A_u) - \tau(f)) = 0$, implying $\tau(A_u) = \tau(F)$, or equivalently $A_u = F$, which is the conclusion we are trying to prove. Therefore, it suffices to prove that there exists $\rho \neq 0$ such that T is a contraction.

Take arbitrary $v_1, v_2 \in V$, and let $v = v_1 - v_2$. Then,

$$\begin{aligned}
\|T(v_1) - T(v_2)\|^2 &= \|v_1 - v_2 - \rho(\tau(A_{v_1}) - \tau(A_{v_2}))\|^2 \\
&= \|v - \rho\tau(A_v)\|^2 && \text{(linearity of } \tau \text{ and } A) \\
&= \|v\|^2 - 2\rho(\tau(A_v), v) + \rho^2\|\tau(A_v)\|^2 \\
&= \|v\|^2 - 2\rho A_v(v) + \rho^2 A_v(\tau(A_v)) && \text{(definition of } \tau) \\
&= \|v\|^2 - 2\rho a(v, v) + \rho^2 a(v, \tau(A_v)) && \text{(definition of } A_u) \\
&\leq \|v\|^2 - 2\rho\alpha\|v\|^2 + \rho^2\gamma\|v\|\|\tau(A_v)\| && \text{(continuity and coercivity)} \\
&= \|v\|^2 - 2\rho\alpha\|v\|^2 + \rho^2\gamma\|v\|\|A_v\| && (\tau \text{ is an isometry}) \\
&\leq \|v\|^2 - 2\rho\alpha\|v\|^2 + \rho^2\gamma^2\|v\|^2 && \text{(continuity of } A_u) \\
&= (1 - 2\rho\alpha + \rho^2\gamma^2)\|v\|^2 \\
&= (1 - 2\rho\alpha + \rho^2\gamma^2)\|v_1 - v_2\|^2
\end{aligned}$$

If there exists ρ such that $1 - 2\rho\alpha + \rho^2\gamma^2 < 1$, then the corresponding T is a contraction. This occurs when $\rho(\rho\gamma^2 - 2\alpha) < 0$. Choosing ρ such that

$$0 < \rho < \frac{2\alpha}{\gamma^2},$$

then $1 - 2\rho\alpha + \rho^2\gamma^2 < 1$, and T is contraction, from which the result follows. □

From the Lax–Milgram theorem, we see that conditions (ii) and (iii) are sufficient for well-posedness. Consequently, the convection–diffusion problem (2) has a unique solution.

In the (Galerkin) finite element method, the variational formulation (1) is solved on a (finite-dimensional) subspace $V_h \subseteq V$. In the case that the bilinear form $a(\cdot, \cdot)$ is symmetric, then there is an induced energy norm $\|\cdot\|_a$, and the solution $u_h \in V_h$ minimizes the error in the energy norm,

$$\|u - u_h\|_a \to \min.$$

Continuity and coercivity give that

$$\begin{aligned}
\|u - u_h\|_V &\leq \frac{1}{\sqrt{\alpha}}\|u - u_h\|_a \\
&= \frac{1}{\sqrt{\alpha}} \min_{v_h \in V_h} \|u - v_h\|_a \\
&= \sqrt{\frac{\gamma}{\alpha}} \min_{v_h \in V_h} \|u - v_h\|_V.
\end{aligned}$$

In the case that $a(\cdot, \cdot)$ is not symmetric, the form does not define an inner product, and so it does not induce an energy norm. However, we still can obtain quasi-optimal error estimates.

Theorem 2 (Céa's Lemma)**.** *Let conditions (ii) and (iii) hold, and suppose u solves* (1) *on V, and u_h solves* (1) *on V_h. Then,*

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V. \tag{3}$$

Proof. For any $v_h \in V_h$, it holds that

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = F(v_h) - F(v_h) = 0,$$

so Galerkin orthogonality holds even in the non-symmetric case. Let $v_h \in V_h$ be arbitrary. We calculate,

$$\begin{aligned} \alpha\|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) && \text{(coercivity)} \\ &= a(u - u_h, u - v_h) && \text{(Galerkin orthogonality)} \\ &\leq \gamma\|u - u_h\|_V\|u - v_h\|_V. && \text{(continuity)} \end{aligned}$$

Dividing both sides by $\alpha\|u - u_h\|_V$ gives

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha}\|u - v_h\|_V.$$

Since $v_h \in V_h$ was arbitrary, the result holds. □

In both the symmetric and non-symmetric cases, the error is quasi-optimal (that is, no more than a constant factor larger than the optimal error), but in the symmetric case the constant is improved by taking square root.

4 Convection-Dominated Problems

Convection-dominated problems: the convection–diffusion problem above models the situation where the dynamics of the system are governed by two processes: convection and diffusion. In some cases, one of the processes may be dominant. For example, the advection velocity may be very small, and the problem is called diffusion-dominant. When the diffusion coefficient is very small (and the advection velocity is moderately sized), the problem is called *convection-dominated*. Note that when $\mu \ll 1$, then we may have $\alpha \ll 1$, where α is the coercivity constant. In this case, the error bound (3) in Céa's lemma will blow up, and we are not guaranteed that the standard finite element method will deliver accurate results. Handling convection-dominated problems well is one of the motivations for the *discontinuous Galerkin method*, which we have seen in the context of time integration methods, but is also (mainly) used for spatial discretizations.

5 Linear Solvers

Given a basis $\{\phi_i\}$ for the finite element space V_h, we can form the matrix A that corresponds to the bilinear form $a(\cdot, \cdot)$, defined by

$$A_{ij} = a(\phi_j, \phi_i).$$

If $u_h = \sum_i u_i\phi_i \in V_h$ and $v_h = \sum_i v_i\phi_i \in V_h$, then we consider the associated vectors of coefficients $\boldsymbol{u} = (u_1, \ldots, u_N), \boldsymbol{v} = (v_1, \ldots, v_N)$, and obtain the relationship

$$\boldsymbol{v}^T A\boldsymbol{u} = a(u_h, v_h).$$

Since in general $a(u_h, v_h) \neq a(v_h, u_h)$, we see that the matrix A will not be symmetric, and so A is not SPD. This means that solvers such as the conjugate gradient method cannot be used, and LU factorization instead of Cholesky needs to be used as a direct method. The classical iterative methods (Jacobi, Gauss–Seidel, etc.) still may be applied, but the convergence theory is not the same as in the SPD case. One of the most common iterative methods for non-symmetric problems is the "Generalized Minimum Residual Method" (GMRES), which is a *Krylov subspace method*, like the conjugate gradient method. This method converges for any invertible matrix, but its computational cost is generally higher than for CG.