MATH 272A: Numerical PDE
Take Home Final

University of California, San Diego

Zihan Shao

1. Use a finite difference method to discretize the problem

$$-a(x)u''(x) + b(x)u'(x) = f(x), \quad x \in (0,1),$$

$$u(0) = u(1) = 0,$$

where $a(x)$ and $b(x)$ are continuous functions and $a(x) \geq a_0 > 0$. Discuss the truncation error and numerical stability of your method (you are allowed to make assumptions on $b(x)$).

Solution.

Finite difference discretization: We choose a uniform gird on $[0,1]$. $x_i = ih$, $i = 0, 1, \ldots, N$ with $Nh = 1$. We use the following finite difference approximation of derivatives

$$u'(x_i) \approx \frac{u(x_{i+1}) - u(x_i)}{2h} \tag{1}$$

$$u''(x_i) \approx \frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} \tag{2}$$

We are to find $u_h : \{x_i\}_{i=0}^N \to \mathbb{R}$ such that

$$-a(x_i)\frac{u_h(x_{i+1}) + u_h(x_{i-1}) - 2u(x_i)}{h^2} + b(x_i)\frac{u_h(x_{i+1}) - u_h(x_{i-1})}{2h} = f(x_i), \qquad i = 1, \ldots, N \tag{3}$$

$$u_h(0) = u_h(1) = 0 \tag{4}$$

This is equivalent to solving the following linear system

$$AU = F \tag{5}$$

where

$$A_h = \frac{1}{h^2}\begin{bmatrix} 2a(x_1) & -a(x_1) + \frac{b(x_1)h}{2} & 0 & \cdots & \cdots & 0 \\ -a(x_2) - \frac{b(x_2)h}{2} & 2a(x_2) & -a(x_2) + \frac{b(x_2)h}{2} & \cdots & \cdots & 0 \\ 0 & -a(x_3) - \frac{b(x_3)h}{2} & 2a(x_3) & -a(x_3) + \frac{b(x_3)h}{2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -a(x_{N-1}) - \frac{b(x_{N-1})h}{2} & 2a(x_{N-1}) \end{bmatrix}$$

and $U = [u_h(x_1), \ldots, u_h(x_N)]^T$ and $F = [f(x_1), \ldots, f(x_N)]^T$. Note that here we don't have correction to F since the boundary is zero.

Trucation error: We state the following theorem for truncation error

Theorem (Truncation error)**.** *Let differential operator L be defined by $Lu := -a(x)u''(x) + b(x)u'(x)$ and*

L_h be its finite difference discretization defined in (3) i.e.

$$(L_h u)(x) := -a(x)\frac{u(x+h)+u(x-h)-2u(x)}{h^2} + b(x)\frac{u(x+h)-u(x-h)}{2h}$$

For $u \in C^4$ we have

$$\|Lu - L_h u\|_{L^\infty(0,1)} \leq Ch \tag{6}$$

for some finite $C \in \mathbb{R}$.

Proof. The proof is quite routine (by Taylor's theorem) so that we just provide a sketch here. Expanding $u'(x \pm h)$ to the 4th order

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\xi_1), \tag{7}$$

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\xi_2), \tag{8}$$

We have

$$|\frac{u(x+h)+u(x-h)-2u(x)}{h^2} - u''(x)| \leq \frac{\max_{y\in[0,1]}|u^{(4)}(y)|}{12}h^2.$$

Now only use Taylor's theorem to the 2nd order, we have

$$|\frac{u(x+h)-u(x-h)}{2h} - u'(x)| = \frac{\max_{y\in[0,1]}|u^{(2)}(y)|}{2}h$$

Since a and b are both continous, thus being bounded on $[0,1]$. Substituting the truncation errors for $u''(x)$ and $u'(x)$, the total error becomes:

$$|Lu - L_h u| \leq \max_{x\in[0,1]}(|a(x)| + |b(x)|) \cdot (\frac{\max_{y\in[0,1]}|u^{(4)}(y)|}{12} + \cdot\frac{\max_{y\in[0,1]}|u^{(3)}(y)|}{6})h.$$

□

Numerical stability: If $b = 0$ and a is a positive constant, then this system is identical to Poisson's equation. The intuition is if bh is small, then we have the numerical stability in the same way as Poisson's equation. In other words, the first order term must not dominate the diffusion term.

Since it is unrealistic to deduce stability through a Maximum principle-style argument (Is it possible?), we try to analyze the resulting matirx A_h.

If we have $|b| < \frac{2a_0}{h}$ for all x, A_h satisfies (P1) and (P2) (defined in lecture notes)

(P1) $a_{ii} > 0$, $a_{ij} < 0$, $i \neq j$.

(P2) A_h is irreduciably diagonally dominant

With these 2 properties, we are able to have stability. (I failed to figure out the details).

2. We consider the Dirichlet problem:

$$\begin{aligned} -\sum_{j,k=1}^{d} \frac{\partial}{\partial x_k}\left(a_{jk}(x)\frac{\partial u}{\partial x_j}(x)\right) + c(x)u(x) = f(x), \quad x \in \Omega, \\ u(x) = g(x), \quad x \in \partial\Omega, \end{aligned} \tag{9}$$

where $c(x) \in L^\infty(\Omega)$, $c(x) \geq 0$, and

$$A(x) := (a_{jk}(x))_{j,k=1}^{d} \in L^\infty(\Omega; \mathbb{R}^{d\times d})$$

is symmetric and uniformly positive definite, satisfying

$$\xi^T A(x)\xi \geq \lambda|\xi|^2, \quad \text{for all } \xi \in \mathbb{R}^d, \text{ and some } \lambda > 0.$$

Solution.

Weak formulation: The weak formulation of this problem is to find $u \in H^1(\Omega)$ such that $u = g \in H^{1/2}(\Omega)$ on $\partial\Omega$ (in the sense of traces) and

$$\int_\Omega (-\sum_{j,k=1}^{d} \frac{\partial}{\partial x_k}\left(a_{jk}(x)\frac{\partial u}{\partial x_j}(x)\right) + c(x)u(x))v(x)dx = \int_\Omega f(x)v(x)dx, \quad \forall v \in H_0^1(\Omega) \tag{10}$$

With integration by parts, we have for $v \in H_0^1(\Omega)$

$$\int_\Omega (-\sum_{j,k=1}^{d} \frac{\partial}{\partial x_k}\left(a_{jk}(x)\frac{\partial u}{\partial x_j}(x)\right) dx = \int_\Omega \left(\sum_{j,k=1}^{d} a_{jk}(x)\frac{\partial u}{\partial x_j}(x)\frac{\partial v}{\partial x_k}(x)\right) dx \tag{11}$$

Note that the boudary term vanishes since $v \in H_0^1(\Omega)$. We then have the weark formulation for the Dirichlet problem

$$a(u,v) = \int_\Omega fv, \quad f \in H^{-1}(\Omega), \forall v \in H_0^1(\Omega), \tag{12}$$

where

$$a(u,v) := \int_\Omega \left(\sum_{j,k=1}^{d} a_{jk}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_k} + c(x)uv\right) dx$$

Existence and Uniqueness via Lax-Milgram Theorem The bilinear form $a(\cdot,\cdot)$ satisfy the conditions of the Lax-Milgram theorem:

- Boundedness: $a(u,v)$ is bounded because $a_{jk}(x)$, $c(x) \in L^\infty(\Omega)$:
  We may first divide $a(u,v)$ into 2 parts.
  Boundedness of the first term $a_1(u,v) = \int_\Omega \left(\sum_{j,k=1}^{d} a_{jk}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_k} + c(x)uv\right) dx$ can be proved by applying Cauchy-Schwarz twice. Let $M = \max_{j,k} \|a_{jk}\|_{L^\infty}$, we have

$$|a_1(u,v)| \leq M \int_\Omega \sum_{j,k} |\frac{\partial u}{\partial x_j}||\frac{\partial v}{\partial x_k}|dx \tag{13}$$

$$= M \int_\Omega (\sum_{j}^{d} |\frac{\partial u}{\partial x_j}|)(\sum_{k}^{d} |\frac{\partial v}{\partial x_k}|)dx \tag{14}$$

Using Cauchy-Schwarz on the summation, we have $\sum_j^d |\frac{\partial u}{\partial x_j}| \leq \sqrt{d}|\nabla u|$ and $\sum_k^d |\frac{\partial v}{\partial x_k}| \leq \sqrt{d}|\nabla v|$. Note that here we use $|\cdot|$ to denote the Euclidean norm in $\mathbb{R}^d$ to avoid confusion. Substituting back and apply Cauchy-Schwarz again, we have

$$|a_1(u,v)| \leq Md \int_\Omega |\nabla u||\nabla v| dx \tag{15}$$

$$\leq Md(\int_\Omega |\nabla u|^2 dx)^{1/2}(\int_\Omega |\nabla v|^2 dx)^{1/2} = Md\|\nabla u\|_2 \|\nabla v\|_2 \tag{16}$$

Boundedness of the second term $a_2(u,v) = \int_\Omega c(x)uvdx$ is given by Cauchy-Schwarz.

$$a_2(u,v) = \int_\Omega c(x)uv\, dx \tag{17}$$

$$\leq \|c\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \tag{18}$$

Overall, we have the boundedness,

$$|a(u,v)| \leq |a_1(u,v)| + |a_2(u,v)| \tag{19}$$

$$\leq \max(Md, \|c\|_{L^\infty(\Omega)}) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \tag{20}$$

- Coercivity: If we further assume that $|c(x)| > c_0 > 0$ Given uniform positive definiteness of A

$$a(u,u) = \int_\Omega \nabla u^T A(x) \nabla u + c(x)u^2 dx \tag{21}$$

$$\geq \lambda \|\nabla u\|^2_{L^2(\Omega)} + c_0 \|u\|^2_{L^2(\Omega)} \tag{22}$$

$$\geq \min(\lambda, c_0) \|u\|^2_{H^1(\Omega)} \tag{23}$$

Remark: It seems that we should have the assumption that $|c(x)| > c_0 > 0$ otherwise the coercivity does not hold. Otherwise we will need u to be zero on the boundary (in terms of trace), which is not the case here, and we may use Poincare's inequality again.

With Lax-Milgram theorem, we have existence and uniqueness of the weak solution.

3. State the Galerkin approximation to the weak formulation of equation (9). Do we have the quasi-optimal approximation property of the Galerkin approximation? Prove your claim. Suppose we use a p-th order finite element space (i.e., the space consists of continuous piecewise p-th order polynomials). What is the order of convergence of the finite element method (you may quote any result from any FEM book (e.g., Brenner-Scott))?

Solution.

Galerkin Approximation: Choose a finite dimensional space $V_h \subseteq V = H_0^1(\Omega)$, then the Galerkin approximation is: Given $f \in (H^1)^*$, find $u_h \in V_h$ such that $u_h = g \in H^{1/2}(\Omega)$ on $\partial\Omega$, such that $a(u_h, v) = \langle f, v\rangle$ for all v.

Quasi-Optimal property: We still have quasi-optimal property since we still have Galerkin orthogonality. We provide the proof here:

Proof. We have Galerkin orthogonality here:

$$a(u, v) = \langle f, v\rangle \quad \forall v \in V \tag{24}$$
$$a(u_h, v) = \langle f, v\rangle \quad \forall v \in V_h \tag{25}$$

which implies

$$a(u - u_h, v) = 0 \quad \forall v \in V_h$$

Then we have, for arbitrary $v \in V_h$

$$\begin{aligned} r\|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \quad \text{(coercivity)} && (26)\\ &= a(u - u_h, u - v) - a(u - u_h, v - u_h) && (27)\\ &= a(u - u_h, u - v) \quad \text{(Galerkin orthogonality)} && (28)\\ &\leq C\|u - u_h\|_V\|u - v\|_V \quad \text{(Boundedness)} && (29) \end{aligned}$$

Rearrange and take infimum over v, we obtain the quasi-optimal approximation error

$$\|u - u_h\|_V \leq \frac{C}{r} \inf_{v \in V_h} \|u - v\|_V \tag{30}$$

□

Order of Convergence: As is argued in *Brenner-Scott* Chapter 5.4, we have

$$\|u - u_h\|_{H^1(\Omega)} = O(h^p)|u|_{H^{p+1}(\Omega)} \tag{31}$$
$$\|u - u_h\|_{L^2(\Omega)} = O(h^{p+1})|u|_{H^{p+1}(\Omega)} \tag{32}$$

4. Consider the saddle point problem: find $(u, p) \in V \times Q$ such that

$$\begin{cases} a(u, v) + b(v, p) = \langle f, v \rangle, \\ b(u, q) = \langle g, q \rangle \end{cases} \tag{33}$$

for all $(v, q) \in V \times Q$.

(a) Suppose we take finite-dimensional spaces $V_h := \text{span}\{\phi_j\}_{j=1}^n \subset V$ and $Q_h := \text{span}\{\varphi_j\}_{j=1}^m \subset Q$ and define the Galerkin approximation with respect to $V_h \times Q_h$. Write down the resulting linear system

$$\begin{pmatrix} A_h & B_h^T \\ B_h & 0 \end{pmatrix} U = F. \tag{34}$$

Solution. Here $A_h \in \mathbb{R}^{n \times n}$ where $A_{ij} = a(\phi_i, \phi_j)$; $B_h \in \mathbb{R}^{m \times m}$ where $B_{ij} = b(\varphi_i, \varphi_j)$. $F \in \mathbb{R}^{n+m}$. The first n entries is given by $F_i = \langle f, \phi_i \rangle$, the next m entries is given by $F_{n+j} = \langle g, \varphi_j \rangle$. By solving the linear system, we obtain $u_h = \sum_{i=1}^n U_i \phi_i$ and $v_h = \sum_{j=1}^m U_{n+j} \varphi_j$.

(b) For the Stokes system, verify that we have

$$a(v_h, v_h) \geq \gamma \|v_h\|_V^2 \quad \forall v_h \in V_h,$$

for some $\gamma > 0$. Show that for the Stokes system, the above inequality is equivalent to the positive definiteness of A_h (i.e., $x^T A_h x \geq c\|x\|^2$ for some $c > 0$).

Proof. We first show a is coercive for Stokes equation. For Stokes system, we have

$$a(u, v) = \int_\Omega \nabla u : \nabla v \, dx \tag{35}$$

Here V_h is a subspace of $H_0^1(\Omega)$. With Poincare's inequality, we have $\int_\Omega \|u\|^2 \leq C_p \int_\Omega \|\nabla u\|^2$ ($\|\cdot\|$ denote Euclidean norm). Therefore,

$$a(u, v) = \int_\Omega \nabla u : \nabla u \, dx \geq 1/C_p \int_\Omega \|u\|^2 + \int_\Omega \|\nabla u\|^2 dx \geq (1 + 1/C_p)\|u\|_{H^1(\Omega)}^2 \tag{36}$$

Without loss of generality, we assume ϕ_i are orthonormal. Let $v_h = \sum_{i=1}^n x_i \phi_i$. We have

$$a(v_h, v_h) = \sum_i \sum_j x_i x_j a(\phi_i, \phi_j) \tag{37}$$

$$= x^T A_h x \tag{38}$$

Then we see $a(v_h, v_h) \geq \gamma \|v_h\|_V^2$ is equivalent to $x^T A_h x \geq \gamma \|v_h\|_V^2 = \gamma x^T x$ and our conclusion follows. If the basis is not orthonormal, then $\|v_h\|_V^2 = x^T G x$ where $G_{ij} = (\phi_i, \phi_j)$ where $(\cdot, \cdot)$ is the associated inner product of V. Clearly, the gram matrix G is positive definite by construction [1] hence $\|x\|_G = x^T G x$ defines a new norm. The conclusion then follows since all norms are equivalent on finite dimensional spaces.

□

[1] $x^T G x = \|v_h\|_V^2 \geq 0$

(c) Under the condition that A_h is positive definite, what condition do we need on B_h for equation (34) to be solvable?

Solution. According to the Brezzi's theorem, we additionally need B_h to be **surjective**, which is equivalent to the following inf-sup condition

$$\inf_{q \in Q_h} \sup_{v \in V_h} \frac{b(v,q)}{\|v\|_V \|q\|_Q} = \beta > 0 \tag{39}$$

5. Prove the Lax-Milgram theorem using the Riesz representation theorem and the Banach fixed point theorem (for a bounded and coercive bilinear form $a : U \times U \to \mathbb{R}$.)

Theorem (Lax-Milgram)**.** *Given a Hilbert space $V, (\cdot,\cdot)$, a continuous, coercive bilinear form $a(\cdot,\cdot)$ and a continuous linear functional $F \in V'$, there exists a unique $u \in V$ such that*

$$a(u, v) = F(v) \quad \forall v \in V. \tag{40}$$

To prove this theorem, we need the following lemma

Lemma (Banach fixed point)**.** *Let V be a Banach space and $T : V \to V$ a mapping satisfying*

$$\|Tv_1 - Tv_2\| \leq M\|v_1 - v_2\|, \tag{41}$$

for all $v_1, v_2 \in V$ and a fixed M with $0 \leq M < 1$. Then, there exists a unique $u \in V$ such that

$$u = Tu, \tag{42}$$

i.e., the contraction mapping T has a unique fixed point u.

Proof. We prove the theorem with the following steps

Step 1: $A : V \to V'$, $(Au)(v) = a(u, v)$ is a well-defined bounded linear map

- $Au \in V'$. Clearly, Au is linear. Continuity is guaranteed by bonudedness of binlinear form a:

$$\|Au\|_{V'} = \sup_{v\in V} \frac{|Au(v)|}{\|v\|_V} = \sup_{v\in V} \frac{|a(u, v)|}{\|v\|_V} \leq C\|u\|_V < \infty \tag{43}$$

- $A \in L(V, V')$. Clearly, A is linear. For continuity, we have

$$\|A\| = \sup_{u\in V} \frac{\|Au\|_{V'}}{\|u\|_V} \leq \sup_{u\in V} \frac{C\|u\|_V}{\|u\|_V} = C \tag{44}$$

Step 2: We want to show that there exists a unique u such that $a(u, v) = F(v)$ for all v, which is equivalent to the existence and uniqueness of u such that $Au = F$ in V' where A is defined in Step 1. With Riesz representation theory, it is also the same as $\tau Au = \tau F$ where τ is the bijection from V' to V. We define the following map between V and itself

$$Tv = v - \rho(\tau Av - \tau F) \tag{45}$$

where ρ is a nonzero constant. Its fixed point, if exists, is the anticipated u. Given the lemma, it remains to show that there exists $\rho \neq 0$ such that T is a contraction.

Step 3: We show such ρ exists.

$$\begin{aligned}
\|Tv_1 - Tv_2\|^2 &= \|v_1 - v_2 - \rho(\tau A v_1 - \tau A v_2)\|^2 \\
&= \|v - \rho(\tau A v)\|^2 \\
&= \|v\|^2 - 2\rho(\tau A v, v) + \rho^2 \|\tau A v\|^2 \\
&= \|v\|^2 - 2\rho A v(v) + \rho^2 A v(\tau A v) \\
&= \|v\|^2 - 2\rho a(v, v) + \rho^2 a(v, \tau A v) \\
&\leq \|v\|^2 - 2\rho r \|v\|^2 + \rho^2 C \|v\| \|\tau A v\| \\
&\leq (1 - 2\rho r + \rho^2 C^2) \|v\|^2 \\
&= (1 - 2\rho r + \rho^2 C^2) \|v_1 - v_2\|^2 \\
&= M^2 \|v_1 - v_2\|^2.
\end{aligned} \tag{46}$$

Here r is the coercivity constant. It remains to find ρ that makes $M < 1$ i.e.

$$0 < 1 - 2\rho r + \rho^2 C^2 < 1$$

We can choose $\rho \in (0, \frac{C^2}{2r})$ to complete the proof.

□