CSC2310 Computational Methods for Partial Differential Equations

1 Classification of PDEs

There are three kinds of PDEs of interest:

1. Elliptic: $\Delta u = 0$ or $\nabla^2 u = 0$
2. Parabolic: $u_t = \nabla^2 u$
3. Hyperbolic: $u_{tt} = \nabla^2 u$

1.1 Elliptic PDEs

Laplace: $\nabla^2 u = 0$
Poisson: $\nabla^2 u = f$

e.g. Distribution of heat. Let Ω be the domain. The boundary conditions on $\partial\Omega$ could be

1. Dirichlet: $u = f$
2. Neumann: $\frac{\partial u}{\partial x} = g(x)$
3. Robin: $\alpha u + \beta \frac{\partial u}{\partial x} = h(x)$
4. Or mixed based on domain.

Consider Laplace's equation on a rectangular region:

$$\begin{cases} \nabla u = 0 \\ u(0, y) = f(y) \\ u(L, y) = g(y) \\ u(x, 0) = u(x, H) = 0 \end{cases}$$

The solution is $u(x, y) = h(x)\phi(y)$ with basis $\sinh\left(\frac{n\pi y}{H}\right)$ $\begin{bmatrix} \sinh\left(\frac{n\pi x}{H}\right) / \sinh\left(\frac{n\pi L}{H}\right) \\ \sinh\left(\frac{n\pi (L-x)}{H}\right) / \sinh\left(\frac{n\pi L}{H}\right) \end{bmatrix}$. Then the solution is

$$u(x, y) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi y}{H}\right) \frac{\sinh\left(\frac{n\pi (L-x)}{H}\right)}{\sinh\left(\frac{n\pi x}{H}\right)} + \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi y}{H}\right) \frac{\sinh\left(\frac{n\pi x}{H}\right)}{\sinh\left(\frac{n\pi x}{H}\right)}$$

$$A_n = \frac{2}{H} \int_0^H f(y) \sin\left(\frac{n\pi y}{H}\right) dy$$

$$B_n = \frac{2}{H} \int_0^H g(y) \sin\left(\frac{n\pi y}{H}\right) dy$$

The solution is perfectly stable. *i.e.* if f, g are perturbed slighly, the solution is also perturbed slightly; the solution is smoothed out inside the domain.

Suppose we solve:

$$\begin{cases} \nabla u = 0 \\ u(0,y) = f(y) \\ u_x(0,y) = g(y) \end{cases}$$

The solution is

$$u(x,y) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi y}{H}\right) \cosh\left(\frac{n\pi L}{H}\right) + \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi y}{H}\right) \sinh\left(\frac{n\pi x}{H}\right) \frac{H}{n\pi}$$

1.2 Parabolic PDEs (Heat Equation)

$$\begin{cases} u_t = k u_{xx} \\ u(0,t) = u(L,t) = 0 \\ u(x,0) = f(x) \end{cases}$$

Separation of variable gives:

$$u(x,t) = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right) e^{-k\left(\frac{n\pi}{L}\right)^2 t}$$
$$B_n = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx$$

Also, notice if $B_n \neq 0$ for some large n, then u_t is huge at $t = 0$, since $\frac{d}{dt} e^{-k\left(\frac{n\pi}{L}\right)^2 t}|_{t=0} = -k\left(\frac{n\pi}{L}\right)^2$

1.3 Hyperbolic PDEs

$$\begin{cases} \partial_t^2 u = c^2 u \partial_x^2 u \\ u(x,0) = f(x) \\ u_t(x,0) = g(x) \end{cases}$$

D'Alembert's solution:

$$u(x,t) = \frac{1}{2c} \int_{x-ct}^{x+ct} g(x)dx + \frac{f(x+ct) + f(x-ct)}{2}$$

If $g = 0$, then $f(x)$ splits into 2 waves with velocity c. The wave equation does not smooth out initial data. It is also not stiff. It's stable but derivatives blow up when perturbing ICs.

1.4 General PDEs

How to relate general PDE to one of the three types?

$$au_{xx} + bu_{yy} + cu_{xy} + du_x + eu_y + fu = g$$

Compare ∂_x^2, ∂_x applied to $f(x)$ in $[-0.5, 0.5]$. The Fourier expansion of $f(x)$ is

$$f(x) = \sum_{-\infty}^{\infty} A_n e^{2\pi i n x}$$

Applying ∂_x^2 multiplies A_n by $-4\pi^2 n^2$. Applying ∂_x multiplies A_n by $2\pi i n$.

$$\partial_x^2 = \begin{bmatrix} -4\pi n^2 & \cdots & 0 \\ 0 & -4\pi n^2 & 0 \\ 0 & \cdots & -4\pi n^2 \end{bmatrix} A_n$$

$$\partial_x = \begin{bmatrix} 2\pi i n & \cdots & 0 \\ 0 & 2\pi i n & 0 \\ 0 & \cdots & 2\pi i n \end{bmatrix} A_n$$

$$\partial_x^2 + \partial_x = \begin{bmatrix} -4\pi n^2 & \cdots & 0 \\ 0 & -4\pi n^2 & 0 \\ 0 & \cdots & -4\pi n^2 \end{bmatrix} \left(I + \begin{bmatrix} 2\pi i n & \cdots & 0 \\ 0 & 2\pi i n & 0 \\ 0 & \cdots & 2\pi i n \end{bmatrix} \right) A_n$$

If higher order terms are all that matter, how to classify?

2nd order linear PDEs can always be written as

$$\begin{bmatrix} \partial x_1 \\ \vdots \\ \partial x_n \end{bmatrix}^T A \begin{bmatrix} \partial x_1 \\ \vdots \\ \partial x_n \end{bmatrix}$$

with A symmetric. For any A symmetric, there is always an invertible P s.t. $P^T A P = \text{diag}(1, -1, 0)$. Let the number of 1s be n_+, number of -1s be n_-, number of 0s be n_0.

Theorem: 1.1: Sylvester's Law of Inertia

The set (n_+, n_-, n_0) is the signature of A. It is invariant over all transformation P. If we pick P s.t. $P^T A P$ has the form $P^T A P = \text{diag}(1, -1, 0)$. Let $v = Px$, then the PDE is

$$\left(\frac{\partial^2}{\partial v_1^2} + \cdots + \frac{\partial^2}{\partial v_{n_+}^2} - \frac{\partial}{\partial v_{n_+ + 1}} - \cdots - \frac{\partial}{\partial v_{n_+ + n}} \right) u + \text{lower order term} = 0$$

2 Finite Difference Method

If we want to solve PDE numerically, we need to somehow represent derivatives. One such way is finite difference:

$$\frac{d}{dx}f(x) \approx \frac{f(x+h)-f(x-h)}{2h}$$

Consider the Taylor expansion:

$$f(x+h) = f(x) + hf'(x) + \frac{f''(x)}{2}h^2 + O(h^3)$$
$$f(x-h) = f(x) - hf'(x) + \frac{f''(x)}{2}h^2 + O(h^3)$$

Then $\frac{d}{dx}f(x) \approx f'(x) + O(h^2)$

There are two sources of errors

1. Truncation error (mathematical error): $O(h^2)$ remainder.
2. Cancellation error (floating point approximation error): $\frac{\epsilon|f|}{h}$, where ϵ is the machine precision. $o(x) = x(1+\epsilon)$.

If $\left|\frac{f^{(m)}(x)}{m!}\right| \approx 1$, total error is approximately $E(h) = \frac{\epsilon}{h} + h^2$. The minimum is achieved at $-\frac{\epsilon}{h^2} + 2h = 0$ or $h = \epsilon^{1/3}$. $E(\epsilon^{1/3}) = 2\epsilon^{2/3} \approx 10^{-10}$ for $\epsilon \approx 10^{-16}$ (typical machine epsilon).

Suppose the truncation error is $O(h^m)$, we get $E(h) = \frac{\epsilon}{h} + h^m$, the minimum is achieved at $h \approx \epsilon^{\frac{1}{m}+1}$, so $E(h) \approx \epsilon^{\frac{m}{m+1}}$. A small perturbation $R(x) + \delta(x)$ ($\delta(x) = \epsilon \sin \omega x$ for example) will significantly change $\frac{d}{dx}(R(x) + \delta(x))$.

What kinds of FDs can we take? Consider a 3×3 grid, $\frac{\partial f_{0,0}}{\partial x} = \frac{1}{2h}(f_{1,0} - f_{-1,0}) + O(h^2)$. $\frac{\partial^2 f_{0,0}}{\partial x \partial y} = \frac{1}{4h^2}(f_{1,1} - f_{1,-1} - f_{-1,1} + f_{-1,-1}) + O(h^2)$.

2.1 Heat Equation

Suppose we want to solve $\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}$, where k is diffusivity with $u(0,t) = u(L,t) = 0$, $u(x,0) = f(x)$. For simplicity, $L = \pi$, $f(x) = \begin{cases} \frac{2}{\pi}x, x \le \frac{\pi}{2} \\ 2 - \frac{2}{\pi}x, x > \frac{\pi}{2} \end{cases}$. Then the Fourier series is

$$u(x,t) = \sum_{n=1}^{\infty} A_m \sin(mx) e^{-km^2 t}$$
$$A_m = \frac{1}{\pi}\int_0^{\pi} f(x)\sin(mx)dx$$

For $m > 1$, $A_m = \frac{4}{\pi(2n+1)^2}(-1)^n$. Easy to see that $\sum_{m=1}^{\infty} |A_m| < \infty$.

Place a grid on the domain with increments Δx and Δt, with $x = j\Delta x$ and $t = n\Delta t$, $j = 0, 1, ..., J$. Consider the grid/mesh functions: let $u_j^n = u(j\Delta x, n\Delta t)$. Write

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = k\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}$$

Apply BC by setting $u_0^n = 0, u_J^n = 0$ and IC $u_j^0 = f(j\Delta x)$.

Consider $\frac{k\Delta t}{(\Delta x)^2}$, the larger the value, the worse the approximation. Reducing Δx and Δt may help.

1. Stability: what happens to $\left|u_j^n - u(j\Delta x, n\Delta t)\right|$ as $n \to \infty$? Whether it stays bounded.
2. Consistency: what happens to $\left|u_j^n - u(j\Delta x, n\Delta t)\right|$ as $\Delta x, \Delta t \to 0$?

Stability:
Suppose we substitute $u(x,t) = \sin(mx)\xi(m)^n$ into the finite difference scheme.

$$\frac{\sin(mj\Delta x)(\xi(m)^{n+1} - \xi(m)^n)}{\Delta t} = k\frac{\sin(m(j+1)\Delta x) - 2\sin(mj\Delta x) + \sin(m(j-1)\Delta x)}{(\Delta x)^2}\xi(m)^n$$

(Split $\sin(m(j+1)\Delta x)$ and $\sin(m(j-1)\Delta x)$)

$$\frac{\sin(mj\Delta x)(\xi(m)-1)\xi(m)^n}{\Delta t} = \frac{\sin(mj\Delta x)(\cos(m\Delta x)-1)}{(\Delta x)^2}\xi(m)^n$$

$$\xi(m) = \frac{2k\Delta t}{(\Delta x)^2}(\cos(m\Delta x) - 1) + 1$$

$$u_j^n = \sum_{m=1}^{\infty} A_m \sin(mj\Delta x)\xi(m)^n$$

For actual solution to the heat equation, $\xi(m) = e^{-km^2\Delta}$. Compare the Taylor expansions:

$$\xi(m) = 1 - \frac{2k\Delta t}{(\Delta x)^2}(1 - \cos(m\Delta x)) = 1 - m^2 k\Delta t + \frac{1}{12}m^4 k\Delta t(\Delta x)^2 + \cdots$$

$$e^{-km^2\Delta t} = 1 - m^2 k\Delta t + \frac{1}{2}m^4 k^2(\Delta t)^2 + \cdots$$

For solution u_j^n to stay bounded, need $\max_m |\xi(m)| \le 1$, equivalently, $\left|1 - \frac{2k\Delta t}{(\Delta x)^2}(1 - \cos(m\Delta x))\right| \le 1$. Since $1 - \cos(m\Delta x) \in (-1, 1)$, we need $\frac{2k\Delta t}{(\Delta x)^2} > 0$ to be bounded. $\Rightarrow 1 - \frac{4k\Delta t}{(\Delta x)^2} \ge -1$, so $\frac{2k\Delta t}{(\Delta x)^2} \le 1$.
$u(x,t)$ is always bounded, so error is bounded if and only if u_j^n is bounded.

Consistency:
Suppose we have $\frac{2k\tilde{\Delta}t}{(\tilde{\Delta}x)^2} = L$ for $\Delta x = \frac{\tilde{\Delta}x}{k}$, $\Delta t = \frac{\tilde{\Delta}t}{k^2}$, then $\frac{2k\Delta t}{(\Delta x)^2} = L$. Now we show that if $L \le 1$, then $|u_j^n - u(j\Delta x, n\Delta t)| \to 0$ as $k \to \infty$.

Proof. Recall that

$$u(x,t) = \sum_{m=1}^{\infty} A_m \sin(mx)e^{-km^2 t}$$

$$u_j^n = \sum_{m=1}^{\infty} A_m \sin(mx)\xi(m)^n$$

Then

$$\begin{aligned} u_j^n - u(j\Delta x, n\Delta t) &= \sum_{m=1}^{\infty} A_m \sin(mx)(\xi(m)^n - e^{-km^2 n\Delta t}) \\ &= \sum_{m=1}^{m_0} A_m \sin(mx)(\xi(m)^n - e^{-km^2 n\Delta t}) + \sum_{m=m_0+1}^{\infty} A_m \sin(mx)(\xi(m)^n - e^{-km^2 n\Delta t}) \\ &:= \Sigma_1 + \Sigma_2 \end{aligned}$$

Choose m_0 big enough s.t.

$$\Sigma_1 \leq 2 \sum_{m=m_0+1}^{\infty} |A_m| < \frac{\epsilon}{2}$$

Now we bound $\Sigma_1 = \sum_{m=1}^{m_0} A_m \sin(mx)(\xi(m)^n - e^{-km^2 n\Delta t})$: Since $a^n + b^n = (a-b)(a^{n-1} + a^{n-2}b + \cdots + b^{n-1})$ and $|\xi(m)| \leq 1$, $|e^{-km^2 n\Delta t}| \leq 1$,

$$\left|\xi(m)^n - e^{-km^2 n\Delta t}\right| \leq \left|\xi(m) - e^{-km^2\Delta t}\right| n$$

$$\xi(m) = 1 - km^2\Delta t + \frac{1}{12}m^4 k\Delta t(\Delta x)^2$$

$$e^{-km^2\Delta t} = 1 - km^2\Delta t + \frac{1}{2}m^4 k^2(\Delta t)^2$$

Since $(\Delta x)^2 = \frac{2k\Delta t}{L}$, we get $|\xi(m) - e^{-km^2\Delta t}| \leq m^4(\Delta t)^2 B$ for some constant B.

$$\Sigma_1 \leq \sum_{m=1}^{m_0} m^4(\Delta t)^2 Bn|A_m| \leq m_0^4 t\Delta t B \sum_{m=1}^{m_0} |A_m|$$

Take Δt small enough s.t. $\Sigma_1 < \frac{\epsilon}{2}$.
Therefore $\Sigma_1 + \Sigma_2 < \epsilon$. □

Implicit Equations:
Let $\delta f_j = f\left(\left(j + \frac{1}{2}\right)\Delta x\right) - f\left(\left(j - \frac{1}{2}\right)\Delta x\right)$ so $\delta^2 f_j = f((j+1)\Delta x) - 2f(j\Delta x) + f((j-1)\Delta x)$. Take a new finite difference scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{k\theta(\delta^2 u)_j^{n+1} + (1-\theta)(\delta^2 u)_j^n}{(\Delta x)^2}$$

with $\theta \in [0,1]$. If $\theta = 0$, we have **explicit** scheme. If $\theta = 1$, we have **implicit** scheme.

Substitute $u(x, n\Delta t) = \sin(mx)\xi(m)^n$ into the new FD scheme, get $\xi(m) = \frac{1-(1-\theta)L(1-\cos(m\Delta x))}{1+\theta L(1-\cos(m\Delta x))}$.
For $\theta \geq \frac{1}{2}$, $|\xi(m)| \leq 1$ for any Δx and Δt. The method if unconditionally stable.

2.1.1 Error Analysis

Local truncation error
Suppose $\tilde{u}(x,t)$ is the exact solution to the PDE. Substitute into $\frac{u_j^{n+1} - u_j^n}{\Delta t} = k\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}$.

$$\text{LHS} = \frac{\tilde{u}_j^n + \left(\frac{\partial \tilde{u}}{\partial t}\right)_j^n \Delta t + \frac{1}{2}\left(\frac{\partial^2 \tilde{u}}{\partial t^2}\right)_j^n (\Delta t)^2 + \mathcal{O}(\Delta t^3) - \tilde{u}_j^n}{\Delta t} = \left(\frac{\partial \tilde{u}}{\partial t}\right)_j^n + \frac{1}{2}\left(\frac{\partial^2 \tilde{u}}{\partial t^2}\right)_j^n \Delta t + \mathcal{O}(\Delta t^2)$$

$$\text{RHS} = k\left(\frac{\partial^2 \tilde{u}}{\partial x^2}\right)_j^n - \frac{k}{12}(\Delta x)^2\left(\frac{\partial^4 \tilde{u}}{\partial t^4}\right)_j^n + \mathcal{O}(\Delta x^4)$$

Note $\left(\frac{\partial \tilde{u}}{\partial t}\right)_j^n = k\left(\frac{\partial^2 \tilde{u}}{\partial x^2}\right)_j^n$, since $\tilde{u}(x,t)$ is the exact solution. The remainders are the local truncation error

$$\text{LTE} = \frac{1}{2}\Delta t\left(\frac{\partial^2 \tilde{u}}{\partial t^2}\right)_j^n + \frac{k}{12}(\Delta x)^2\left(\frac{\partial^4 \tilde{u}}{\partial t^4}\right)_j^n + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4)$$

Global Errors: Let $\epsilon_j^n = u_j^n - \tilde{u}_j^n$, then

$$\frac{\epsilon_j^{n+1} - \epsilon_j^n}{\Delta t} = k\frac{\epsilon_{j+1}^n - 2\epsilon_j^n + \epsilon_{j-1}^n}{(\Delta x)^2} + \text{LTE}$$

$$\epsilon_j^{n+1} - \epsilon_j^n = \frac{L}{2}(\epsilon_{j+1}^n - 2\epsilon_j^n + \epsilon_{j-1}^n) + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2 \Delta t), \text{ where} L = \frac{2k\Delta t}{(\Delta x)^2}$$

$$\epsilon_j^{n+1} = \frac{L}{2}\epsilon_{j+1}^n + (1-L)\epsilon_j^n + \frac{L}{2}\epsilon_{j-1}^n + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2 \Delta t)$$

If $L \leq 1$, then $\max_j \left|\epsilon_j^{n+1}\right| \leq \max_j \left|\epsilon_j^n\right| + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2 \Delta t)$.
Since $\epsilon_j^0 = 0$ and $n = \frac{t}{\Delta t}$,

$$\max_j \left|\epsilon_j^n\right| \leq n(\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2 \Delta t)) \leq t(\mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2))$$

***Definition:* 2.1: Stability**

Contributions of local errors stay bounded as grid size $\to 0$.

***Definition:* 2.2: Consistency**

LTE$\to 0$ as $\Delta x, \Delta t \to 0$.

***Theorem:* 2.1: Lax Equivalence Theorem**

Stability + Consistency $\Leftrightarrow$ Convergence

Improving LTE:
Notice in the formula for LTE

$$\text{LTE} = \frac{1}{2}\Delta t \left(\frac{\partial^2 \tilde{u}}{\partial t^2}\right)_j^n + \frac{k}{12}(\Delta x)^2 \left(\frac{\partial^4 \tilde{u}}{\partial t^4}\right)_j^n + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4)$$

Choose $\frac{k\Delta t}{(\Delta x)^2} = \frac{1}{6}$, then the first two terms cancel. This is called **compact finite difference scheme**. However, this is typically not used in practice, due to lack of robustness.

2.1.2 Rate of Convergence

***Theorem:* 2.2: Convergence of Fourier Coefficients**

If $f \in C^p([0,\pi])$, the Fourier coefficients $A_m = \frac{1}{\pi}\int_0^\pi f(x)\sin(mx)dx$ decay to zero like $A_m = \mathcal{O}\left(\frac{1}{m^{p+1}}\right)$ as $m \to \infty$.

Earlier, we have the bounds:

$$\Sigma_2 \leq 2\sum_{m=m_0+1}^{\infty} |A_m| = \mathcal{O}\left(\frac{1}{m_0^p}\right)$$

$$\Sigma_1 \leq \sum_{m=1}^{m_0} m^4(\Delta t)^2 B\frac{t}{\Delta t}|A_m| \leq C_0 Bt\Delta t \sum_{m=1}^{m_0} m^{4-p-1} \leq \begin{cases} C_1 m_0^{4-p}\Delta t, p \leq 3 \\ C_2 \log(m_0)\Delta t, p = 4 \\ C_3 \Delta t, p \geq 5 \end{cases}$$

For $p \geq 5$, the sum $\Sigma_1 + \Sigma_2$ decays following Σ_1.

To minimize $\Sigma_1 + \Sigma_2$ when $p \leq 4$, need to choose optimal m_0, which turns out to be $\mathcal{O}\left(\Delta t^{-\frac{1}{4}}\right)$.

Assuming $\Delta t \propto (\Delta x)^2$,

$$\Sigma_1 + \Sigma_2 \leq \begin{cases} \mathcal{O}\left(\Delta t^{p/4}\right) = \mathcal{O}\left(\Delta x^{p/2}\right), p \leq 3, \\ \mathcal{O}\left(\Delta t |\log \Delta t|\right) = \mathcal{O}(\Delta x^2 |\log \Delta x|), p = 4 \\ \mathcal{O}(\Delta t) = \mathcal{O}(\Delta x^2), p \geq 5 \end{cases}$$

Proof. For $p \leq 3$. The error term is $E(m_0) = m_0^{4-p}\Delta t + m_0^{-p}$.
To find the minimum, $E'(m_0) = 0$.

$$\begin{aligned} m_0^{4-p-1}\Delta t - m_0^{-p-1} &= 0 \\ \Rightarrow m_0 &= (\Delta t)^{-\frac{1}{4}} \end{aligned}$$

□

2.2 Equilibrium Heat Equation

Consider the heat equation:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} \\ u(0,t) = \alpha, u(1,t) = \beta, u(x,0) &= g(x) \end{aligned}$$

Let $Q(x,t) = -f(x)$. At equilibrium $\frac{\partial u}{\partial t} = 0$. This gives a Poisson equation:

$$\frac{\partial^2 u}{\partial x^2} = f(x), u(0) = \alpha, u(1) = \beta$$

The general form in high dimension is $\frac{\partial u}{\partial t} = \nabla^2 u + Q$ or $\nabla^2 u = f(x)$ for equilibrium. The numerical form is:

$$\frac{u_{j-1} - 2u_j + u_{j+1}}{(\Delta x)^2} f_j \text{ for } j = 1, 2, ..., m$$

where $u_j \approx u(j\Delta x), f_j \approx f(j\Delta x), u_0 = \alpha, u_{m+1} = \beta$

Let $h = \Delta x$. We get $AU = F$, where

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & 0 \\ 0 & 1 & -2 & \ddots & \vdots \\ \ddots & \ddots & \ddots & -2 & 1 \\ 0 & \cdots & \cdots & 1 & -2 \end{bmatrix}, U = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, F = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

To deal with BC, $F = \begin{bmatrix} f_1 - \frac{\alpha}{h^2} \\ \vdots \\ f_m - \frac{\beta}{h^2} \end{bmatrix}$.
This comes from $\frac{\alpha - 2u_1 + u_2}{h^2} = f_1$.

To compute the LTE (τ_j), substitute $\tilde{u}(j\Delta x)$ into the finite difference formula, and apply Taylor expansion.

$$\tau_j = \frac{\tilde{u}_{j-1} - 2\tilde{u}_j + \tilde{u}_{j+1}}{h^2} - f_j = \partial^2 \tilde{u}_j + \frac{h^2}{12}\partial^4 \tilde{u}_j + \mathcal{O}(h^4) - f_j = \mathcal{O}(h^2)$$

Consistency: $\tau_j \to 0$ as $h \to 0$
Stability: Since $AU = F$, we have $U = A^{-1}F$. Thus $\|U\| \leq \left\|A^{-1}\right\| \|F\|$
Let $E = u - \tilde{u}$, $AE = A(u - \tilde{u}) = -\tau$, so $E = -A^{-1}\tau$, $\|E\| \leq \left\|A^{-1}\right\| \|\tau\|$.
Since $\|\tau\|_\infty = \mathcal{O}(h^2)$, $\|\tau\|_2 = \mathcal{O}\left(\frac{h^2}{\sqrt{h}}\right) = \mathcal{O}\left(h^{\frac{3}{2}}\right)$.
If $\left\|A^{-1}\right\| \leq c$ as $h \to 0$, then $\|E\| \leq C\mathcal{O}\left(h^{\frac{3}{2}}\right)$, so consistency and stability $\Rightarrow$ convergence.

2.2.1 Closer Look at Stability

Since A is symmetric, $\|A\|_2 = \max_p |\lambda_p|$, where λ_p is the p-th eigenvalue.
Suppose that $u_j^p = \sin(p\pi jh)$, $u_j \approx u(j\Delta x)$.

$$(Au^p)_k = \frac{\sin(p\pi(j-1)h) - 2\sin(p\pi jh) + \sin(p\pi(j+1)h)}{h^2} = \frac{2}{h^2}(\cos(p\pi h) - 1)\sin(p\pi jh)$$

so $\lambda_p = \frac{2}{h^2}(\cos(p\pi h) - 1)$, where $p \leq m$, $h = \mathcal{O}\left(\frac{1}{m}\right)$. When $ph \approx 1$, we get minimum.

$$\cos(p\pi h) - 1 = -\frac{p^2\pi^2 h^2}{2} + \mathcal{O}(h^4 p^4), \lambda_p = \frac{2}{h^2}(\cos(p\pi h) - 1) = -\pi^2 + \mathcal{O}(h^2)$$

$|\lambda_p|$ increases w.r.t. p. Take $p = 1$ to get $\lambda_1 \approx -\pi^2 + \mathcal{O}(h^2)$.
Eigenvalues of A^{-1} are λ_p^{-1}, so the order will reverse, $|\lambda_1^{-1}|$ will be the largest.
True eigenvalues and eigenfunctions are $\tilde{u}^p(x) = \sin(p\pi x)$ and $\tilde{\lambda_p} = -p^2\pi^2$, $\left\|A^{-1}\right\|_2 \leq \left|\frac{1}{\lambda_1}\right| = \frac{1}{\pi^2}$.
$\|E\|_\infty \leq \sqrt{m}\|E\|_2 = \mathcal{O}(h)$. To get a bound on $\|E\|_\infty$, we need to construct or approximate A^{-1}.

For $\frac{d^2u}{dx^2} = f(x)$ with $u(0) = \alpha, u(L) = \beta$, the Green's function $G(x, x_0)$ satisfies $\frac{d^2}{dx^2}G(x, x_0) = \delta(x - x_0)$ and $G(0, x_0) = 0, G(L, x_0) = 0$.

Notice that $\frac{d^2}{dx^2}G(x, x_0) = 0$ at all $x \neq x_0$, $G(0, x_0) = G(L, x_0) = 0$. Thus, $G(x, x_0) = \begin{cases} bx, x < x_0 \\ d(x - L), x > x_0 \end{cases}$.

Since we require $\frac{d^2}{dx^2}G(x, x_0) = \delta(x - x_0)$, integrating both sides, we get $\frac{d}{dx}G = H(x - x_0)$ and G must be continuous.

Solve for b and d, $\begin{cases} d - b = 1 \\ bx_0 = d(x_0 - L) \end{cases}$. This gives:

$$G(x, x_0) = G(x_0, x) = \begin{cases} -\frac{x}{L}(L - x_0), x < x_0 \\ -\frac{x_0}{L}(L - x), x > x_0 \end{cases}$$

$f_j(x) = \begin{cases} \frac{1}{h}, j = 1 \\ 0, \text{ otherwise} \end{cases}$. Write $F = (0, ..., \frac{1}{h}, ..., 0)$, where ith row is $\frac{1}{h}$.
$U = A^{-1}F$ is the ith column of A^{-1}, multiplied by $\frac{1}{h}$.

Recall that $\lim_{h\to 0} f_h(x) = \delta(x - x_0)$. Let $\tilde{u}$ be the true solution. As $h \to 0$, $f_h(x) \to G(x, x_0)$.

We might think that $\frac{1}{h}U \to G(jh, x_0)$, but f_h is changing as $h \to 0$.

Recall $\frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = 0$ for $i \neq j$, $\frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f_h(x) = \frac{1}{h}$.
For $j \neq i$, the differential relation has two solutions $u_j = 1$ and $u_j = j$.

At $j = i$, $\frac{1}{h}\left(\frac{u_{i-1}-u_j}{h} - \frac{u_i - u_{i+1}}{h}\right) < \frac{1}{h}$.

By inspection, $U_j = G_{i,j} = \begin{cases} -\frac{x_j}{L}(L - x_i), j \le i \\ -\frac{x_i}{L}(L - x_j), j > i \end{cases}$

$$AE = -\tau = (0, ..., \epsilon, ..., 0)^T,\ E = \epsilon h G_{i,j} = \sum_{i=1}^{m} h G_{i,j}\tau_i = \mathcal{O}(mh\left\|\tau\right\|_\infty) = \mathcal{O}(\left\|\tau\right\|_\infty)$$

2.3 Elliptic Equations in 2 or More Dimensions

Consider $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f$ on Ω with $u = 0$ on $\partial\Omega$. Let $x_i = i\Delta x, y_j = j\Delta y$:

$$\nabla^2 u = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{\Delta x^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{\Delta y^2} = f_{i,j}$$

Typically, we have $\Delta x = \Delta y$, so we have a 5-point stencil.
Let $u^{(j)} = (u_{1,j}, ..., u_{m,j})^T$, $u = (u^{(1)}, ..., u^{(m)})^T$.

$$A = \begin{bmatrix} T & I & 0 & 0 & 0 \\ I & T & I & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T & I \\ 0 & 0 & 0 & I & T \end{bmatrix} \text{ with } T = \begin{bmatrix} -4 & 1 & 0 & 0 & 0 \\ 1 & -4 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -4 & 1 \\ 0 & 0 & 0 & 1 & -4 \end{bmatrix}$$

Let $N = m^2$, $A \in \mathbb{R}^{N\times N}$, the band size (# of foward and backward pass in Gaussian elimination) is $d = \sqrt{N} = m$. The cost of Gaussian elimination is $\mathcal{O}(Nd^2) = \mathcal{O}(N^2) = \mathcal{O}(m^4)$.

It has been proven that the best possible cost of a direct method for Poisson in 2D is $\mathcal{O}(m^3) = \mathcal{O}(N^{3/2})$.

Actually, best possible cost for d-dimensional problem is $\mathcal{O}(N^{1+\frac{1}{d}})$, where $N = m^d$.

The 9-point Laplacian Stencil:

$$\nabla_9^2 u = \nabla^2 u + \frac{h^2}{12}(u_{xxxx} + 2u_{xxyy} + u_{yyyy}) + \mathcal{O}(h^4)$$

In the bracket, it is equivalent to $\nabla^2(\nabla^2 u) = \nabla^2 f$. Since we have f, we can just add a correction to RHS to cancel 2nd order term and increase to fourth order.

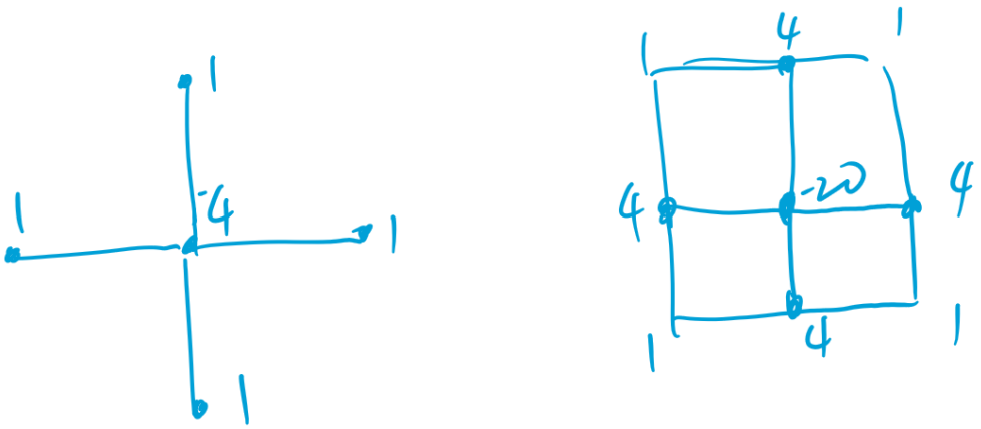
2.4 Iterative Methods

Suppose we have a system $Ax = y$. Define $F(x) = x^T y - x^T A x$, $\nabla F(x) = 0$ solves $Ax = y$.

Gradient descent is $x_{n+1} = x_n - \alpha\nabla F(x_n)$.

Conjugate gradient: gradient descent + orthogonalization w.r.t. $\langle x, y\rangle_A = x^T A y$ where A is p.d.
Convergence depends on the condition number $\kappa(A)$.
The error is $\|e_n\| \le 2\left(1 - \frac{2}{\sqrt{\kappa(A)}}\right)^n \|e_0\|$. After $\mathcal{O}(\sqrt{\kappa(A)}) \approx \mathcal{O}(m)$ steps, e_n will be small.
The cost of multiplication is $\mathcal{O}(m^2)$. Total cost is $\mathcal{O}(m^3) = \mathcal{O}(N^{3/2})$. In d-dim, it is $\mathcal{O}(m^{d+1}) = \mathcal{O}(N^{1+\frac{1}{d}})$.

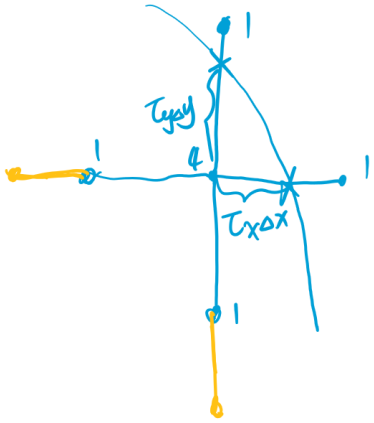
Figure 1: 5-point and 9-point Laplacian stencils



2.5 Non-rectangular Domains

Suppose we solve $\nabla^2 u = f$ on Ω with $u = 0$ on $\partial\Omega$, and Ω is non-rectangular. In 5-point stencil ∇_5, some points (at most 2) may be out of the region. If more than 2 are out of region, then finite difference will fail.

Figure 2: Non-rectangular Domain



$0 < \tau_y \leq 1, 0 < \tau_x \leq 1$. Apply Taylor expansion:

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= \frac{1}{(\Delta x)^2}\left(\frac{2}{\tau+1}u(x_0-\Delta x) - \frac{2}{\tau}u(x_0) + \frac{2}{\tau(\tau+1)}u(x_0+\tau\Delta x)\right) \\ &= \frac{\partial^2 u}{\partial x^2}(x_0) + \frac{1}{3}(\tau-1)u_{xxx}(x_0)\Delta x + \mathcal{O}((\Delta x)^2\end{aligned}$$

As soon as the symmetry disappears, the first order term appears, and we lose second order convergence. It is possible to show that after we extend as in Figure 2, we get back second order convergence:

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= \frac{1}{(\Delta x)^2}\left(\frac{\tau-1}{\tau+2}u(x_0-2\Delta x) + \frac{2(2-\tau)}{\tau+1}u(x-x_0) + \frac{3-\tau}{\tau}u(x_0) + \frac{6}{\tau(\tau+1)(\tau+2)}u(x_0+\tau\Delta x)\right) \\ &= \frac{\partial^2 u}{\partial x^2}(x_0) + \mathcal{O}((\Delta x)^2)\end{aligned}$$

Finite difference requires regularity of the domain. Some smooth transformations may be required for finite difference to work nicely.

3 Direct and Iterative Methods

Definition: 3.1: Fill-in

Fill-ins are entries of L, where $LL^T = A$, not appearing in A. Some orderings are better than others. Finding the optimal permutation P s.t. $LL^T = PAP^T$ is NP-hard.

Optimal cost of a direct solve for Poisson's equation in 2D is $\mathcal{O}(N^{\frac{3}{2}})$ where $N = m^2$. Optimal fill-in for Cholesky factorization of a tri-diagonal matrix is $\mathcal{O}(N \log N)$, $A = L^T L$. In general, for a m^d grid problem in d-dimensions, optimal cost of direct method is $\mathcal{O}\left(N^{\frac{3(d-1)}{d}}\right)$ and optimal fill-in is $\mathcal{O}\left\{N^{\frac{2(d-1)}{d}}\right\}$. Nested dissection provides optimal ordering that achieves the above bounds.
Iterative methods typically have complexity $\mathcal{O}\left(N^{1+\frac{1}{d}}\right)$ when $d = 2$ or 3, they are about the same as directly methods, but typically have larger constants.

3.1 Sparse Direct Solvers

Sparse matrix data structure (**triplet form**)

1. $i[\,] = \{i_1, i_2, ...\}$, row index
2. $j[\,] = \{j_1, j_2, ...\}$, column index
3. $X[\,] = \{x_1, x_2, ...\}$, entries

They represent a matrix with (i, j)-position of value X, we only store non-zero entries.

Compressed-Column form:

1. $p[\,] = \{p_1, p_2, ...\}$, stores the number of non-empty cells in each column (or prefix sum of it), length= #columns
2. $i[\,] = \{i_{11}, ..., i_{1p_1}, i_{21}, ...\}$, stores the indices of the non-zero entries in each columns, length= #non-zero entries
3. $X[\,]$, entries

The cost of accessing a column is $\mathcal{O}(1)$.

Sparse triangular solvers:
Let A be a large triangular sparse matrix with nonzero diagonal. We want to solve $Ax = b$ for x.Back substitution:

$$A_{11}x_1 = b_1 \Rightarrow x_1 = \frac{b_1}{A_{11}}$$
$$A_{21}x_1 + A_{22}x_2 = b2 \Rightarrow x_2 = \frac{b_2 - A_{21}x_1}{A_{22}}$$

Consider $Lx = b$, where $L = \begin{bmatrix} l_{11} & 0 \\ l_{21} & L_{22} \end{bmatrix}$, $l_{11} \in \mathbb{R}$, $l_{21} \in \mathbb{R}^{(N-1)\times 1}$, $L_{22} \in \mathbb{R}^{(N-1)\times(N-1)}$, $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, $x_1, b_1 \in \mathbb{R}$, $x_2, b_2 \in \mathbb{R}^{N-1}$.
We can solve recursively $x_1 = \frac{b_1}{l_{11}}$, $L_{22}x_2 = b_2 - l_{21}x_1$, where L_{22} is the same structure as L.

The cost of Algorithm 1 is $\mathcal{O}(n + f)$, where n is dimension, f is the number of floating point calculations. For dense b, $f \approx |L|$.

Algorithm 1 Sparse Triangular Solve Dense x

```
x = b
for j = 0 : n − 1 do
    x_j = x_j / l_jj
    for each i > j where l_ij ≠ 0 do
        x_i = x_i − l_ij x_j
    end for
end for
```

Suppose x is sparse, we modify the algorithm as in Algorithm 2. The cost is still $\mathcal{O}(n + |b| + f)$, where $|b|$ is the sparsity of b.

Algorithm 2 Sparse Triangular Solve Sparse x

```
x = b
for j = 0 : n − 1 do
    if x_j ≠ 0 then
        x_j = x_j / l_jj
        for each i > j where l_ij ≠ 0 do
            x_i = x_i − l_ij x_j
        end for
    end if
end for
```

We use n sparse triangular solves to factor $A = LL^T$. A generic algorithm will require $\mathcal{O}(n^2)$. To improve from $\mathcal{O}(n + |b| + f)$, suppose we know in advance the set $X = \{j : x_j \neq 0\}$. Then we modify to Algorithm 3. The cost is now $\mathcal{O}(|b| + f)$.

Algorithm 3 Sparse Triangular Solve Sparse x Improved

```
x = b
for j ∈ X do
    x_j = x_j / l_jj
    for each i > j where l_ij ≠ 0 do
        x_i = x_i − l_ij x_j
    end for
end for
```

Reachability: How do we determine X?
Non-zero entries of x follows two rules (without potential cancellation effects):

1. $b_i \neq 0 \Rightarrow x_i \neq 0$
2. $x_j \neq 0 \wedge \exists i(l_{ij} \neq 0) \Rightarrow x_i \neq 0$

This can be described as Graph traversal problems:
Let $G_L = (V, E)$ be a directed graph with $V = \{1, 2, ..., n\}$, $E = \{(j, i) : l_{ij} \neq 0\}$ (non-zero entries in j propagates to $i > j$) G_L is acyclic by construction. We want to mark all nodes in X. The following rules apply:

1. Mark $i \in B = \{i : b_i \neq 0\}$
2. If j is marked, and $(j, i) \in E$, then i is marked.

Let `Reach`(B) denote set of all nodes reachable from $i \in B$ by paths in G_L. Then $X =$`Reach`(B). This can be done by DFS. Cost is $\mathcal{O}(|\tilde{V}| + |\tilde{E}|)$ where $\tilde{V}$ and $\tilde{E}$ are the vertices and edges the algorithm visits. Each edge corresponds to a required floating point operation, so the cost is $\mathcal{O}(|x| + f) = \mathcal{O}(|b| + f)$.

DFS does not return a sorted array X. We don't want to sort it because we don't want $|x| \log |x|$ cost. However, DFS returns X in topological order. *i.e.* if $(j, i) \in E$, then i must appear after j in X. Since we only update X_i when we have a j s.t. $(j, i) \in E$ and $x_j \neq 0$, we will always have applied all updates by the time we get to x_i.

Note: DFS does not give the exact set X, it gives $\tilde{X} \supset X$. Also, some topological orders may not be valid. A simple modification is by prepending node n to list only after considering all other nodes that depend on n.

3.1.1 Cholesky Factorization

Suppose A is symmetric and positive definite, we want to find L s.t. $LL^T = A$. Let $L = \begin{bmatrix} L_{11} & 0 \\ l_{12}^T & l_{22} \end{bmatrix}$, $L_{11} \in \mathbb{R}^{(n-1)\times(n-1)}$, $l_{22} \in \mathbb{R}$, $l_{12} \in \mathbb{R}^{(n-1)\times 1}$,

$$\begin{bmatrix} L_{11} & 0 \\ l_{12}^T & l_{22} \end{bmatrix} \begin{bmatrix} L_{11}^T & l_{12} \\ 0 & l_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & a_{12} \\ a_{12}^T & a_{22} \end{bmatrix}$$

$L_{11}L_{11}^T = A_{11}$, $L_{11}l_{12} = a_{12}$, sparse with $|a_{12}| \approx |l_{12}| \ll n$. Call the algorithm recursively until L_{11} becomes a 1×1 matrix and it is easy solve backwards.

Every triangular solve has cost $\mathcal{O}(|a_{12}| + f)$, so it is possible to achieve $\mathcal{O}(N)$.

The elimination tree:
Suppose we solve $L_{11}l_{12} = a_{12}$ using sparse triangular solve. Let $\mathcal{L}_k$ be non-zero pattern of l_{12}, $\mathcal{L}_k = \texttt{Reach}_{G_{k-1}}(\mathcal{A}_k)$, where G_{k-1} is graph of L_{11}^T and $\mathcal{A}_k$ is non-zero pattern for kth upper column of A. We can do DFS on G_{k-1} and get a cost of $\mathcal{O}(|\mathcal{L}_k| + f)$, but we can also do it in just $\mathcal{O}(|\mathcal{L}_k|)$.

It turns out that for a Cholesky factorization $LL^T = A$, $i < j < k \wedge l_{ji} \neq 0 \wedge l_{ki} \neq 0 \Rightarrow l_{kj} \neq 0$. This implies that the graph G_{k-1} can be pruned so that each vertex has only one outgoing edge.
$|V| \approx |E|$.

Generally, the matrix we work with for elliptic PDEs are symmetric positive definite. For non-symmetric A, we have the LU-decomposition $LU = A$. For direct solvers, this only adds a constant factor to the asymptotic runtime.

3.2 Iterative Methods

Suppose we want to solve $Ax = b$, we will approximate x by x_m in the affine subspace $x_0 + K_m$ of dimension m and enforce the condition on residual $b - Ax_m \perp \mathcal{L}_m$ for another affine subspace $\mathcal{L}_m$. The methods discussed are Krylov subspace methods: Let $r_0 = Ax_0 - b$,
$K_m(A, r_0) = \text{span}\left\{r_0, Ar_0, ..., A^{m-1}r_0\right\}$.

3.2.1 Conjugate Gradient Method

Suppose A is symmetric positive definite. Let $\mathcal{L}_m = K_m$. Then x_m satifies $b - Ax_m \perp \mathcal{L}_m$ if and only if x_m minimizes the A-norm $\|x_* - x\|_A$, where $b - Ax_* = 0$ and $\|x\|_A = x^T Ax$. This is equivalent to $x_m \in x_0 + K_m$ minimizing $F(x) = x^T Ax - 2x^T b$. (Note $\nabla F(x) = 2Ax - 2b$)

Overall idea of the procedure:

1. Start with an initial guess x_0

2. Compute residual $r_k = b - Ax_k$
3. Pick a search direction $p_k = r_k - \sum_{i=0}^{k-1} \frac{\langle p_i, r_k \rangle_A}{\langle p_i, p_i \rangle_A} p_i$
4. Let $x_{k+1} = x_0 + \sum_{i=0}^{k} \alpha_i p_i$, $p_i \in K_i$. We want $V^T(b - Ax_{k+1}) = 0$, where $V = (p_0|p_1|...|p_k)$.
5. Choose α_i s.t. $\langle A(x_* - x_{k+1}), p_i \rangle = 0$, for $i = 0, 1, ..., k$

$$\begin{aligned} \langle A(x_* - x_{k+1}), p_i \rangle &= p_i^T A(x_{k+1} - x_*) = 0 \\ p_i^T A x_{k+1} &= p_i^T b \\ p_i^T A x_0 + p_i^T \sum_{i=0}^{k} \alpha_i A p_i &= p_i^T b \\ &\text{but } p_i, p_j \text{ are orthogonal} \\ \alpha_i p_i^T A p_i &= p_i^T (b - Ax_0) = p_i^T r_0, \end{aligned}$$

   so $\alpha_i = \frac{p_i^T r_0}{p_i^T A p_i}$ and we only need to update α_i at step i.
6. Set $x_{k+1} = x_k + \alpha_k p_k$
7. Residuals are orthogonal: $\langle r_j, r_i \rangle = 0$ if $i \neq j$

Proposition: 3.1:

Suppose A is PSD. Let $\mathcal{L}_m = K_m$. Then x_m minimizes $\|x_* - x\|_A$ where $b - Ax_* = 0$ if and only if $b - Ax_m \perp \mathcal{L}_m = K_m$.

Proof. Suppose $M \subset \mathbb{R}^n$ is a subspace and let $x \in \mathbb{R}^n$, then $\min_{y \in M} \|x - y\|_2 = \|x - y^*\|_2$ if and only if 1) $y^* \in M$ and $x - y^* \perp M$.

For $\tilde{x} \in x_0 + K_m$ to minimize $\|x_* - x\|_A$ over x we need $\tilde{x} \in x_0 + K_m$ and $x_* - \tilde{x} \perp_A K_m$. In other words, $(x_* - \tilde{x})^T A v = 0$ for all $v \in K_m$.

$$\Rightarrow (Ax_* - A\tilde{x})^T v = 0 \Rightarrow (b - A\tilde{x})^T v = 0 \forall v \in K_m$$

□

Proposition: 3.2:

The vector x_m satisfying $b - Ax_m \perp K_m$ if and only if it minimizes $F(x) = x^T Ax - 2x^T b$ over $x \in x_0 + K_m$.

Proof. From Proposition 3.1, we know that x_m minimizes $\|x_* - x\|_A$ over $x \in x_0 + K_m$.

$$\begin{aligned} \|x_* - x\|_A &= (x_* - x)^T A(x_* - x) = (Ax_* - Ax)^T (x_* - x) \\ &= (b - Ax)^T (x_* - x) \\ &= b^T x_* - x^T A x_* - b^T x + x^T A x = b^T x_* - x^T b - b^T x + x^T A x \\ &= C + x^T A x - 2x^T b \end{aligned}$$

For some constant C, and the constant does not affect the minimizer. □

Proposition: 3.3:

$\langle r_k, p_i \rangle_A = 0$ for all $i = 0, 1, ..., k-2$

Proof. Note we choose x_k s.t. $\langle A(x_* - x_k), p_i \rangle_A = 0$ for $i = 0, 1..., k$.

Recall that $p_i \in K_i$. Thus $Ap_i \in K_{i+1}$. There exist constants μ_j s.t. $Ap_i = \sum_{j=0}^{i+1} \mu_j p_j$.

$$\langle r_k, p_i \rangle_A = \langle r_k, Ap_i \rangle = \left\langle r_k, \sum_{j=0}^{i+1} \mu_j p_j \right\rangle_A = 0$$

if $i + 1 \leq k - 1$ or $i \leq k - 2$. □

Proposition: 3.4:

$\langle r_i, r_j \rangle = 0$ if $i \neq j$

Proof.

$$r_k = b - Ax_k = b - A\left(x_0 + \sum_{i=0}^{k-1} \alpha_i p_i\right) = b - Ax_0 - \sum_{i=0}^{k-1} \alpha_i Ap_i = \sum_{j=0}^{k} \mu_j p_j \in K_{k+1}$$

Since $\langle r_k, p_j \rangle = 0$ for $j \leq k - 1$, then

$$\langle r_k, r_i \rangle = \left\langle r_k \sum_{j=0}^{i} \mu_j p_j \right\rangle = 0 \text{ for } i \leq k - 1$$

Same for $i \geq k + 1$ □

Remark 1. p_is are A-orthogonal, r_is are orthogonal.

Algorithm 4 Conjugate Gradient

1: Start with x_0
2: Compute $r_k = b - Ax_k$, $p_0 = r_0$
3: Set $p_k = r_k - \dfrac{\langle p_{k-1}, r_k \rangle_A}{\langle p_{k-1}, p_{k-1} \rangle_A} p_{k-1}$
4: Set $x_{k+1} = x_k + \dfrac{\langle p_k, r_0 \rangle}{\langle p_k, p_k \rangle_A} p_k$
5: Repeat n times

Total runtime of Algorithm 4 is $\mathcal{O}(nN)$, where N is the size of A, $A \in \mathbb{R}^{N \times N}$. We may lose orthogonality as we progress. In the worst case, instead of converging in n steps, it will take approximately $3n$ steps.

3.2.2 Generalized Conjugate Residuals

Consider finding $x_m \in x_0 + K_m$ s.t. $b - Ax_m \perp \mathcal{L}_m$. Now, instead of setting $\mathcal{L}_m = K_m$, we relax it to $\mathcal{L}_m = AK_m$.

***Proposition:* 3.5:**

Suppose A is nonsingular. Let $\mathcal{L}_m = AK_m$. Then x_m minimizes $\|b - Ax_m\|_2$ if and only if x_m satisfies $b - Ax_m \perp \mathcal{L}_m$.

Proof. Suppose $x_m \in x_0 + K_m$ minimizing $\|b - Ax_m\|_2$.
This is equivalent to $\Delta x_m = x_m - x_0 \in K_m$ minimizing $\|b - A(x_0 + \Delta x_m)\|_2 = \|r_0 - A\Delta x_m\|_2$ over $\Delta x_m \in K_m$

$$\Leftrightarrow \min_{\Delta \tilde{x_m} \in AK_m} \|r_0 - \Delta \tilde{x}_m\|_2$$

Thus $\Delta \tilde{x}_m \in AK_m$, $r_0 - \Delta \tilde{x}_m \perp AK_m \Leftrightarrow x_m \in x_0 + K_m$, $b - Ax_m \perp AK_m$. □

Algorithm 5 Generalized Conjugate Residuals

1: Start with x_0
2: Compute $r_k = b - Ax_k$, $p_0 = r_0$
3: Choose search direction $p_k = r_k - \sum_{i=0}^{k-1} \frac{\langle Ap_i, Ar_k \rangle}{\langle Ap_i, Ap_i \rangle} p_i$ and orthogonalize $\langle p_k, p_j \rangle_{A^T A} = 0$ for $j \leq k-1$
4: Set $x_{k+1} = x_0 + \sum_{i=0}^{k} \alpha_i p_i$ s.t. $\langle A(x_{k+1} - x_*), Ap_i \rangle = 0$ for $i = 0, 1, ..., k$, $p_i \in K_{i+1}$

Algorithm 5 works for non-symmetric matrices.

***Proposition:* 3.6:**

If $A^T = A$, then $\langle r_k, p_i \rangle_{A^T A} = 0$ for $i \leq k-2$

Proof. Choose x_k s.t. $\langle A(x_* - x_k), Ap_i \rangle = 0$ for $i = 0, 1, ..., k-1$ for $b - Ax_m \perp AK_m$. Then $Ap_i = \sum_{i=0}^{i+1} \mu_j p_j$.

$$\langle r_k, p_i \rangle_{A^T A} = \langle Ar_k, Ap_i \rangle = \left\langle Ar_k, \sum_{j=0}^{i+1} \mu_j p_j \right\rangle = \left\langle r_k, \sum_{j=0}^{i+1} \mu_j Ap_j \right\rangle = 0$$

if $i + 1 \leq k - 1$ or $i \leq k - 2$. □

***Proposition:* 3.7:**

$\langle r_i, Ar_j \rangle = 0$ if $i \neq j$ (residuals are conjugate)

Proof.

$$\begin{aligned} r_k = b - Ax_k &= b - A\left(x_0 + \sum_{i=0}^{k-1} \alpha_i p_i\right) \\ &= b - Ax_0 - \sum_{i=0}^{k-1} \alpha_i Ap_i \end{aligned}$$

Since $\langle r_k, Ap_j \rangle = 0$ for $j \leq k-1$, it follows that $\langle r_i, Ar_j \rangle = 0$ for $j \leq 1$. □

Remark 2. p_is are A^TA-orthogonal, r_is are A orthogonal.

Now, we have a method that minimize $\|x_* - x_m\|_A$ for symmetric positive definite A and a method that minimize $\|Ax_m - b\|_2$ for any A. For k iterations, $\mathcal{O}(nk)$ for conjugate gradient, $\mathcal{O}(nk^2)$ for generalized conjugate residual.

3.2.3 Arnoldi's Method

Suppose we want to construct $K_m(A, v_1) = \text{span}\{v_1, Av_1, ..., A^{m-1}v_1\}$. Choose $v_1 \in \mathbb{R}^m$ s.t. $\|v_1\| = 1$. Set $w_j = Av_j$ and orthogonalize w_j to $v_1, ..., v_j$. Set $v_{j+1} = \frac{w_j}{\|w_j\|_2}$. We get $v_1, v_2, ..., v_m$ an orthonormal basis for $K_m(A, v_1)$.
Let $V_m = (v_1|v_2|\cdots|v_m)$. $AV_m = V_mH_m + w_me_m^T$ for a $m \times m$ upper Hessenberg matrix H_m, and $w_me_m^T = \|w_m\| V_{m+1}e_m^T$ is something that cannot be represented due to orthogonalization.

Full Orthogonalization Method:
Suppose we want to find $x_m \in x_0 + K_m(A, r_0)$ s.t. $b - Ax_m \perp K_m$.
Set $v_1 = \frac{r_0}{\|r_0\|_2}$ and $\beta = \|r_0\|_2$. We have $x_m = x_0 + V_my_m$, where y_m is some coefficient vector.

$$r_m = b - Ax_m = b - A(x_0 + V_my_m) = r_0 - AV_my_m$$

Thus $b - Ax_m \perp K_m$ is equivalent to $V_m^T(r_0 - AV_my_m) = 0 \Leftrightarrow V_m^Tr_0 - V_m^TAV_my_m = 0$.
Since $r_0 = \beta v_1$, we get $V_m^Tr_0 = V_m^T\beta v_1 = \beta e_1$. Also, $V_m^TAV_m = H_m$, so $\beta e_1 - H_my_m = 0$, and $y_m = H_m^{-1}\beta e_1$. If A is symmetric, then H_m is triangular.

GMRES (Generalized Minimal Residuals):
Again write $AV_m = V_mH_m + \|w_m\| V_{m+1}e_m^T = V_{m+1}\overline{H_m}$.
We want $x_m \in x_0 + K_m(A, r_0)$ s.t. $b - Ax_m \perp AK_m$.
Let $v_1 = \frac{r_0}{\|r_0\|_2}$, $\beta = \|r_0\|_2$.

$$r_m = b - Ax_m = b - A(x_0 + V_my_m) = r_0 - AV_my_m$$

The optimality condition is equivalent to minimizing r_m in $\|\cdot\|_2$-norm.
Multiplying V_{m+1}^T projects r_m onto V_{m+1}, $r_0 \in K_{m+1}$, $AV_m \in K_{m+1}$.
We get $V_{m+1}^T(\beta v_1) - V_{m+1}^TV_{m+1}\overline{H_m}y_m = \beta e_1 - \overline{H_m}y_m$.
Minimizing $\left\|\beta e_1 - \overline{H_m}y_m\right\|_2$ is a $(m+1) \times m$ least square problem.

3.2.4 Convergence of Conjugate Gradient

Recall that conjugate gradient minimizes $\|x_* - x\|_A$ over $x_m \in x_0 + K_m$. We have $x_m = x_0 + q_m(A)r_0$, where q_m is a polynomial of order $m-1$. Let $P_m = \{\text{polynomials of order } m\}$.

$$\|x_* - x_0 - q_m(A)r_0\|_A = \min_{q \in P_{m-1}} \|x_* - x_0 - q(A)r_0\|_A$$

Let $d_0 = x_* - x_0$, rewrite $r_0 = b - Ax_0 = A(A^{-1}b - x_0) = A(x_* - x_0) = Ad_0$. It is equivalent to $\min_{q \in P_{m-1}} \|d_0 - Aq(A)d_0\|_A$. Also let $r_m(A) = I - Aq_m(A)$.

$$\|x_* - x_m\|_A = \|(I - Aq_m(A))d_0\|_A = \min_{q \in P_{m-1}} \|(I - Aq(A))d_0\|_A$$

$$\|r_m(A)d_0\|_A = \min_{r \in P_m, r(0)=1} \|r(A)d_0\|_A$$

If A is symmetric p.d., then we can always diagonalize it, writing $A = UDU^*$, with U orthonormal, D diagonal. Then $r(A) = Ur(D)U^* = U\text{diag}(r(\lambda_i))U^*$ and $UA^kU^* = (UAU^*)^k$. Then $\|r(A)d_0\|_A^2 = \sum_{i=1}^n \lambda_i r(\lambda_i)^2 \xi_i^2$, where ξ_i are components of d_0 in the basis $u_1, ..., u_n$. Thus

$$\|r_m(A)d_0\|_A^2 = \min_{r\in P_m, r(0)=1} \sum_{i=1}^n \lambda_i r(\lambda_i)^2 \xi_i^2$$

Note $\sum_{i=1}^n \lambda_i \xi_i^2 = \|d_0\|_A^2$, so

$$\min_{r\in P_m, r(0)=1} \sum_{i=1}^n \lambda_i r(\lambda_i)^2 \xi_i^2 \le \min_{r\in P_m, r(0)=1} \max_{\lambda\in[\lambda_1,\lambda_n]} r(\lambda)^2 \|d_0\|_A^2$$

This provides an upper bound of $\|x_* - x_m\|_A$.

Spectrum of A^{-1} is $\frac{1}{\lambda_i}$ and we are trying to approximate A^{-1} by $p(A)$ where $p \in P_{m-1}$, whose spectrum is $p(\lambda_j)$, $\frac{1}{\lambda} - p(\lambda) = \frac{1-\lambda p(\lambda)}{\lambda} = \frac{r(\lambda)}{\lambda}$, $r(\lambda) \in P_m$ and $r(0) = 1$.

General Form: $\min_{p\in P_k, p(\gamma)=1} \max_{t\in[\alpha,\beta]} |p(t)|$, it is minimized by $\hat{T}_k = \frac{T_k\left(1+2\frac{t-\beta}{\beta-\alpha}\right)}{T_k\left(1+2\frac{\gamma-\beta}{\beta-\alpha}\right)}$, where T_k is the Chebyshev polynomial of degree k.

Minimum at T_k is thus $\frac{1}{\left|T_k\left(1+2\frac{\gamma-\beta}{\beta-\alpha}\right)\right|}$. Let $\gamma = 0$, $\alpha = \lambda_1$, $\beta = \lambda_n$, $\eta = \frac{\lambda_1}{\lambda_n - \lambda_1}$. Then

$$\|x_* - x_m\|_A \le \frac{1}{T_m(1+2\eta)} \|x_* - x_0\|_A \le 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^m \|x_* - x_0\|_A$$

Note that number of iterations m will typically be small for a good convergence. $m = \mathcal{O}(\sqrt{\kappa(A)})$.

It is possible to show that for conjugate residual $\|r_m\|_2 \le 2\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^{\lfloor \frac{m}{2} \rfloor} \|r_0\|_2$. Convergence rate is half. It will have $m = \mathcal{O}(\kappa(A))$ iterations.

3.3 Classical Iterative Methods

We use V to denote approximated solutions and u to denote exact solution. Let $e = u - v$ be error and $r = f - Au$ be the residual. Notice that $Ae = r$.

Jacobi Method:
Consider the finite difference matrix for the 1D BVP $-u'' = f$:

$$-u_{j-1} + 2u_j - u_{j+1} = h^2 f_j, 1 \le j \le n-1, u_0 = u_n = 0$$

Suppose we only solve part of this system at each step:
Take $v^{(k)}$ and solve for $v^{(k+1)}$ by the equation

$$v_j^{(k+1)} = \frac{1}{2}\left(v_{j-1}^{(k)} + v_{j+1}^{(k)} + h^2 f_j\right), \text{ for } 1 \le j \le n-1$$

If eventually $v^{(k)} \approx v^{(k+1)}$, v solves the problem.

In matrix form: let $D = \text{diag}(2)$, L and U represent lower/upper bands, then $A = D - L - U$ and we have

$$\begin{aligned}(D-L-U)u &= f\\ \Rightarrow Du &= (L+U)u + f\\ Dv^{(k+1)} &= (L+U)v^{(k)} + f\\ \Rightarrow v^{(k+1)} &= D^{-1}(L+U)v^{(k)} + D^{-1}f\end{aligned}$$

$R_J = D^{-1}(L+U)$ is the Jacobi iteration matrix.

Weighted/Damped Jacobi Method:

$$v_j^{(k+1)} = (1-w)v_j^{(k)} + \frac{w}{2}\left(v_{j-1}^{(k)} + v_{j+1}^{(k)} + h^2 f_j\right) \text{ for } 1 \le j \ne n-1, 0 \le w \le 1$$

In matrix form:

$$v^{(k+1)} = [(1-w)I + wR_J]v^{(k)} + wD^{-1}f, R_w := (1-w)I + wR_J = I - \frac{w}{2}A$$

In Jacobi methods, we have to solve for all $v^{(k)}$ before we can solve for $v^{(k+1)}$

Gauss-Seidel:
Suppose instead, we do:

$$v_j^{(k+1)} = \frac{1}{2}\left(v_{j-1}^{(k+1)} + v_{j+1}^{(k)} + h^2 f_j\right), 1 \le j \le n-1,$$

i.e. use updated value at previous step $v_{j-1}^{(k+1)}$ in the same iteration, for potentially faster process.
In matrix form:

$$(D-L)u = Uu + f \Rightarrow v^{(k+1)} = (D-L)^{-1}Uv^{(k)} + (D-L)^{(-1)}f$$

Define $R_{GS} = (D-L)^{-1}U$, the Gauss-Seidel matrix. If we switch the order back and forth (ascending to descending), then it is called the symmetric Gauss-Seidel.

Red-Black Order:

$$v_{2j}^{(k+1)} = \frac{1}{2}\left(v_{2j-1}^{(k)} + v_{2j+1}^{(k)} + h^2 f_{2j}\right); \quad v_{2j+1}^{(k+1)} = \frac{1}{2}\left(v_{2j}^{(k+1)} + v_{2j+2}^{(k+1)} + h^2 f_{2j+1}\right)$$

Once we solve $v_{2j}^{(k+1)}$, all equations in the second part are independent and can be solved in parallel.

3.3.1 Convergence

Consider $v^{(k+1)} = Rv^{(k)} + g$, where the exact solution satisfies $u = Ru + g$. Let $e^{(k)} = u - v^{(k)}$, $e^{(k+1)} = Re^{(k)}$, $e^{(m)} = R^m e^{(0)}$, so $\left\|e^{(m)}\right\| \to 0$ if $\|R^m\| \to 0$, so we need $\rho(R) < 1$ ($\max|\lambda_j| < 1$)
For the specific finite difference matrix A

$$\begin{aligned}\lambda_p(A) &= -2\left(\cos\left(\frac{p\pi}{n}\right) - 1\right) = 4\sin^2\left(\frac{p\pi}{2n}\right), v_{p,j} = \sin\left(\frac{p\pi j}{n}\right), 1 \le p \le n-1, 0 \le j \le n\\ \lambda_p(R_w) &= 1 - \frac{w}{2}\lambda_p(A) = 1 - 2w\sin^2\left(\frac{p\pi}{2n}\right),\end{aligned}$$

Small p gives $\lambda_p(R_w) \to 1$, error does not shrink much as time progresses. Large p gives $\lambda_p(R_w) \to -1$ at speed depending on w. To make $\max_{\frac{n}{2} \le p \le n-1}|\lambda_p(R_w)|$ as small as possible, we require
$\lambda_{n/2}(R_w) = \lambda_{n-1}(R_w)$, which gives $w = \frac{2}{3}$ and $|\lambda_p| \le \frac{1}{3}$ on $\left[\frac{n}{2}, n-1\right]$. Factor of 3 is scaled (smoothing factor).

4 Multigrid

Let Ω^h denote a mesh with mesh size h, $\lambda_1 \approx 1 - \frac{w\pi^2}{2}h^2$, so bigger h means error converges to zero faster everywhere. We can use coarse grids for low frequency in the original problem.

$$v^h_{p,2j} = \sin\left(\frac{p\pi 2j}{n}\right) = \sin\left(\frac{p\pi j}{n/2}\right) = v^{2h}_{p,j}, 1 \le j < \frac{n}{2}$$

We can move to coarse grid (low frequency) by adjusting the eigenvectors. However, we don't know the eigenvectors, so we cannot do projections directly.

Nested Iteration:

1. Relax $Au = f$ on Ω^{2h} to get an initial guess v^{2h}
2. Use that guess to relax $Au = f$ on Ω^h

Correction Scheme:

1. Relax $Au = f$ on Ω^h to get an approximation for v^h
2. Compute the residual $r = f - Av^h = A(u - v^h) = Ae^h$
3. Relax the residual equation $Ae = r$ on Ω^{2h} to get an approximation on the error e^{2h}
4. Correct the approximation on Ω^h by setting $v^h \leftarrow v^h + e^{2h}$ where e^{2h} is interpolated to Ω^h.

Relax always reduce errors. Adding an extra relaxation after correction, we get the *two-grid* correction scheme. This can be done recursively down and up $v^h \leftarrow V(v^h, f)$.

1. Relax $A^h u = f^h$ with initial guess
2. If Ω^h is the coarsest grid, go to step 5.
3. $f^{2h} \leftarrow f^h - A^h v^h$ (RHS becomes residuals, subsampled to Ω^{2h})
   $v^{2h} \leftarrow 0$. Then call $v^{2h} \leftarrow V^{2h}(v^{2h}, f^{2h})$
4. Correct $v^h \leftarrow v^h + v^{2h}$ where v^{2h} is interpolated to Ω^h
5. Relax $A^h u = f^h$ with initial guess v^h_i.

We can add extra cycles to get μ-cycle scheme: $v^h \leftarrow M_\mu(v^h, f^h)$.

To get a good initial guess for v^h, start with coarse grids.

Full-Multigrid: $v^h = FMG^h(f^h)$

1. If Ω^h is the coarsest grid, go to step 5
2. $f^{2h} \leftarrow f^h$ by average/subsample
3. $v^{2h} \leftarrow FMG^{2h}(f^{2h})$
4. Set $v^h \leftarrow v^{2h}$ by interpolation
5. $v^h \leftarrow V^h(v^h, f^h)$

Interpolation:
How to go from Ω^h to Ω^{2h} and vice verse?
For interpolation, we can use local polynomials.
For restriction, can set $v^{2h}_j = v^h_{2j}$ (subsampling/injection). Can also take a weighted average with weights $[1, 2, 1]$. We need to have $I^h_{2h} = c(I^{2h}_h)^T$ for some constant c.

5 Parabolic Equations

Consider the heat equation

$$\partial_t u = \partial_x^2 u, u(0,t) = u(\pi,t) = 0, u(x,0) = f(x)$$

We have the finite difference scheme:

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{1}{h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n), k = \Delta t, h = \Delta x$$

We substituted $u(x, nk) = \sin(mx)(\xi(m))^n$ into the finite difference scheme, and solve for $\xi(m)$. We showed for u_j^n to stay bounded as $n \to \infty$, we need $|\xi(m)| \leq 1$ for all m. This gives us the stability condition $\frac{2k}{h} \leq 1$.

More general BCs: $u(0,t) = g_0(t)$, $u(1,t) = g_1(t)$, IC: $u(x,0) = f(x)$. Explicit scheme is:

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{1}{h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n)$$

Let $\delta f_j = f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}$, $\delta^2 f_j = f_{j-1} - 2f_j + f_{j+1}$, we have a family of implicit schemes:

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{1}{h^2}(\theta(\delta^2 u)_j^{n+1} + (1-\theta)(\delta^2 u)_j^n), \text{ for } \theta \in (0,1]$$

For $\theta = \frac{1}{2}$, we have the Crank Nicolson method. Rewrite as:

$$-ru_{j-1}^{n+1} + (1+2r)u_j^{n+1} - ru_{j+1}^{n+1} = ru_{j-1}^n + (1-2r)u_j^n + ru_{j+1}^n, r = \frac{k}{2h^2}$$

The cost is $\mathcal{O}(M)$, where M is number of grids.

Local Truncation Error (LTE) $\tau(x,t)$ is obtained by substituting true solution $u(x,t)$ into the finite difference scheme.

Explicit method:

$$\begin{aligned}\tau(x,t) &= \frac{u(t+k) - u(x,t)}{k} - \frac{1}{h^2}(u(x-h,t) - 2u(x,t) + u(x+h,t)) \\ &= \left(\frac{1}{2}k - \frac{1}{12}h^2\right) u_{xxxx} + \mathcal{O}(k^2 + h^4) \\ &= \mathcal{O}(k + h^2)\end{aligned}$$

For Crank-Nicolson, $\tau(x,t) = \mathcal{O}(k^2 + h^4)$.

5.1 Method of Lines

Consider the explicit scheme. Apply discretization in space only, and we have an equation in t:

$$u_j'(t) = \frac{1}{h^2}(u_{j-1}(t) - 2u_j(t) + u_{j+1}(t)), j = 1, 2, ..., m$$

This is a system of ODES:

$$u'(t) = Au(t) + g(t)$$

Apply BCs:

$$u_1'(t) = \frac{1}{h^2}(g_0(t) - 2u_1(t) + u_2(t)) = \frac{1}{h^2}(-2u_1(t) + u_2(t)) + \frac{g_0(t)}{h^2}$$
$$u_m'(t) = \frac{1}{h^2}(-2u_m(t) + u_{m-1}(t)) + \frac{g_1(t)}{h^2}$$

Then we get $g(t) = \frac{1}{h^2}\begin{bmatrix} g_0(t) \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}$, $A = \frac{1}{h^2}[1, -2, 1]$ banded matrix.

This will be computationally expensive than directly solving PDEs, but we can use it to analyze stability.

Stability Analysis using MOLs:
The eigenvalues of A are $\lambda_p = \frac{2}{h^2}(\cos(p\pi h) - 1) = -p^2\pi^2 + \mathcal{O}(h^2)$.
Euler's method is $u^{n+1} = u^n + f(u^n)$. Apply it to $u'(t) = Au(t) + g(t)$. For it to be stable, we require $|1 + k\lambda| < 1$ for all eigenvalues λ of A. This is because the system can be written as $y' = \lambda y$, or $y_{n+1} = y_n + k\lambda y_n$.

The biggest eigenvalue of A is $\lambda_m = -\frac{4}{h^2}$. If $S \subset \mathbb{C}$ is the stability region, $\lambda k \in S$ gives us $\left|1 - \frac{4k}{h^2}\right| \leq 1$, so $-2 \leq -\frac{4k}{h^2} \leq 0$, which gives the stability condition we see before.

For trapezoid rule, $u^{n+1} = u^n + \frac{1}{2}\left(f(u^n) + f(u^{n+1})\right)$, $y_{n+1} = y_n + \frac{1}{2}(k\lambda y_n + k\lambda y_{n+1})$, we get Crank-Nicolson. Unconditional stability over LHP.

In general, if $u_t = \kappa u_{xx}$, we want $\kappa \propto \frac{h}{k}$. For explicit scheme, we get $k \propto h^2$.

Convergence:
Fix a point (x, t) and examine error as $k, h \to 0$.

Fix relationship between k and h, say $\frac{k}{h^2} = M$. Take $k \to 0$. We can rewrite methods as $u^{n+1} = B(k)u^n + b^n(k)$. For Crank-Nicolson, $b^n(k) = g(nk)$, we get $B(k) = \left(I - \frac{k}{2}A\right)^{-1}\left(I + \frac{k}{2}A\right)$.

Definition: 5.1: Lax-Richtmyer Stable

A linear method of the form $u^{n+1} = B(k)u^n + b^n(k)$ is Lax-Richtmyer stable if for each $T > 0$, there is a constant $C_T > 0$ s.t. $\|B(k)^n\| \leq C_T$ for all $k > 0$ and $h \geq 0$ s.t. $kh \leq T$.

Theorem: 5.1: Lax Equivalence

A consistent method of the form $u^{n+1} = B(k)u^n + b^n(k)$ is convergent if and only if it is Lax-Richtmyer stable.

Proof. Suppose we apply the iteration to the true solution $u(x, t)$: $u^{n+1} = Bu^n + b^n + k\tau^n$, where $u^n = \begin{bmatrix} u(x_1, t_n) \\ \vdots \\ u(x_m, t_n) \end{bmatrix}$ and $\tau^n = \begin{bmatrix} \tau(x_1, t_n) \\ \vdots \\ \tau(x_m, t_n) \end{bmatrix}$.
Since the numerical solution $\hat{u}$ satisfies $\hat{u}^{n+1} = B\hat{u}^n + b^n$.
Subtract to get $E^{n+1} = BE^n - k\tau^n$, where $E^n = \hat{u}^n - u^n$.

After N steps,

$$E^N = B^N E^0 - k\sum_{n=1}^{N} B^{N-n}\tau^{n-1}$$

$$\left\|E^N\right\| \le \left\|B^N\right\| \left\|E^0\right\| + k\sum_{n=1}^{N} \left\|B^{N-n}\right\| \left\|\tau^{n-1}\right\|$$

If the method is Lax-Richtmyer stable, then $\left\|B^{N-n}\right\| \le C_T$ for all $Nk \le T$, so

$$\left\|E^N\right\| \le C_T \left\|E^0\right\| + NkC_T \max_{n\in[1,N]} \left\|\tau^{n-1}\right\| \le C_T \left\|E_0\right\| + TC_T \max_{n\in[1,N]} \left\|\tau^{n-1}\right\|$$

Thus $\left\|E^N\right\| \to 0$, because $\left\|E^0\right\| = 0$ with initial condition $u(x,0) = f(x)$ known. □

5.2 Von Neumann Analysis

If $u : \mathbb{R} \to \mathbb{R}$, then the Fourier transform of u is $\hat{u}(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x} u(x)dx$. The inverse is $u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega x}\hat{u}(w)dw$. Parseval's relation: $\|u\|_2 = \|\hat{u}\|_2$.

We can also define Fourier transform on a grid function $u : \mathbb{Z} \to \mathbb{R}$:

$$\hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-im\xi} u_m, \xi \in [-\pi, \pi]$$
$$u_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{-im\xi}\hat{u}(\xi)d\xi$$

If spacing of grid is h instead of 1, change of variable with $\xi \mapsto h\xi$ to get

$$\hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-imh\xi} u_m h, \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$$
$$u_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{-imh\xi}\hat{u}(\xi)d\xi$$

Also, we have the parseval's relation:

$$\|\hat{u}\|^2 = \int_{-\pi/h}^{\pi/h} |\hat{u}(\xi)|^2 \, d\xi = \sum_{m=-\infty}^{\infty} |u_m|^2 h = \|u\|_h^2$$

Let $D_0 v_j = \frac{1}{2h}(v_{j+1} - v_{j-1})$.

$$D_0 e^{ijh\xi} = \frac{1}{2h}\left(e^{i(j+1)h\xi} - e^{i(j-1)h\xi}\right) = \frac{1}{h}\sin(h\xi)e^{ijh\xi},$$

so $e^{ijh\xi}$ is an eigenvector of D_0 with eigenvalue $\frac{1}{h}\sin(h\xi)$.
Also $\frac{\partial}{\partial x}e^{ix\xi} = i\xi e^{ix\xi}$. Thus $D_0 e^{ijh\xi} = i\xi e^{ijh\xi} + \mathcal{O}(h^2)$ by Taylor expansion of $\frac{1}{h}\sin(h\xi)$.
Consider $u^{n+1} = B(k)u^n$ with $\|B(k)^n\| \le C_T$, *i.e.* $\|B(k)\| \le 1 + \alpha k$. Then $\hat{u}^{n+1}(\xi) = g(\xi)\hat{u}^n(\xi)$.
For $u_j^{n+1} = u_j^n + \frac{k}{h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n)$. Take $u_j^n = e^{ijh\xi}$ and get $u_j^{n+1} = g(\xi)u_j^n$,
$g(\xi) = 1 + \frac{2k}{h^2}(\cos(\xi h) - 1)$.

5.3 Multidimensional Problems

Consider $u_t = u_{xx} + u_{yy}$ with 5-point stencil ∇_5^2,

$$u_{ij}^{n+1} = u_{ij}^n + \frac{k}{2}(\nabla_5^2 u_{ij}^n + \nabla_5^2 u_{ij}^{n+1})$$

Rearraging:

$$\begin{aligned}\left(1 - \frac{k}{2}\nabla_5^2\right) u_{ij}^{n+1} &= \left(1 + \frac{k}{2}\nabla_5^2\right) u_{ij}^n \\ \Rightarrow \left(I - \frac{k}{2}A_{\nabla_5^2}\right) u^{n+1} &= \left(I + \frac{k}{2}A_{\nabla_5^2}\right) u^n \\ u^{n+1} &= \left(I - \frac{k}{2}A_{\nabla_5^2}\right)^{-1} \left(I + \frac{k}{2}A_{\nabla_5^2}\right) u^n\end{aligned}$$

Let $A = \left(I - \frac{k}{2}A_{\nabla_5^2}\right)$, A has the same sparsity as $A_{\nabla_5^2}$. Eigenvalues are $\lambda_{p,q}(A) = 1 - \frac{k}{h^2}\left[(\cos(p\pi h) - 1) + (\cos(q\pi h) - 1)\right]$.

Locally One-dimensional Methods (LOD):
Consider the following LOD method:

$$\begin{aligned}u_{ij}^* &= u_{ij}^n + \frac{k}{2}((\delta_x^2 u)_{ij}^n + (\delta_x^2 u)_{ij}^*) \\ u_{ij}^{n+1} &= u_{ij}^* + \frac{k}{2}((\delta_y^2 u)_{ij}^* + (\delta_y^2 u)_{ij}^{n+1})\end{aligned}$$

In matrix form:

$$\begin{aligned}\left(I - \frac{k}{2}D_x^2\right) u^* &= \left(I + \frac{k}{2}D_x^2\right) u^n \\ \left(I - \frac{k}{2}D_y^2\right) u^{n+1} &= \left(I + \frac{k}{2}D_y^2\right) u^*\end{aligned}$$

u^* is in between u^n and u^{n+1}. How should we deal with boundary data?
We can solve for it by considering known BCs in u^{n+1} and solving the second equation backwards to get u^* on the boundary. It is a first order method, but fast.

Alternating Direction Implicit (ADI):

$$\begin{aligned}u_{ij}^* &= u_{ij}^n + \frac{k}{2}((\delta_y^2 u)_{ij}^n + (\delta_x^2 u)_{ij}^*) \\ u_{ij}^{n+1} &= u_{ij}^* + \frac{k}{2}((\delta_x^2 u)_{ij}^* + (\delta_y^2 u)_{ij}^{n+1})\end{aligned}$$

5.4 Hyperbolic Systems and Advection Equations

$u_t - \sum_{i,j} a_{ij} u_{x_i x_j} = 0$ is hyperbolic if $A = (a_{ij})$ is positive definite. Equivalently, let $d = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \\ \frac{\partial}{\partial t} \end{bmatrix}$,

$A_1 = \begin{bmatrix} A & 0 \\ 0 & -1 \end{bmatrix}$, then $dAdu = 0$.

Let $v = \begin{bmatrix} u_{x_1} \\ u_{x_2} \\ \vdots \\ u_{x_n} \\ u_t \end{bmatrix}$, B_j =matrix with 1 at (j,j)-position, zero otherwise, and $B_0 = 1$. We get:

$$B_0 v_t + \sum_{j=1}^{n} B_j v_{x_j} = 0$$

If B_0 is positive definite, B_j is symmetric, then it is symmetric hyperbolic system.

Consider $u_t + a u_x = 0$, where a is a constant. IC: $u(x,0) = \eta(x)$, then $u(x,t) = \eta(x - at)$

$$\frac{u_j^{n+1} - u_j^n}{k} = -\frac{a}{2h}(u_{j+1}^n - u_{j-1}^n)$$
$$u_j^{n+1} = u_j^n - \frac{ak}{2h}(u_{j+1}^n - u_{j-1}^n)$$

Write $u'(t) = Au(t)$, $A = -\frac{a}{2h}\begin{bmatrix} 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \\ 1 & & -1 & 0 \end{bmatrix}$. Note $A^T = -A$, so $\lambda_p(A)$ are pure imaginary.
$\lambda_p(A) = -\frac{ia}{h}\sin(2\pi p h), p = 1, 2, ..., m$, unstable for any $\frac{k}{h}$ using methods of lines.
Take $k = h^2$, $u^{n+1} = B(k)u^n$, $B(k) = I + kA(h)$.
Then $\lambda_p(B(k)) = 1 + k\lambda_p(A(h))$, $|1 + k\lambda_p|^2 \le 1 + (a\sqrt{k})^2 = 1 + a^2 k$. Then
$\|B(k)^n\| \le (1 + a^2 k)^{\frac{n}{2}} \le e^{\frac{a^2 T}{2}}$.

Leapfrog:
Let's use the midpoint method:

$$y^{n+1} = y^n + hf\left(t^{n+\frac{1}{2}}, y^{n+\frac{1}{2}}\right), u^{n+1} = u^{n-1} + 2kAu^n$$

Stability region: $\{i\alpha : \alpha \in (-1,1)\}$, *i.e.* condition is $\left|\frac{ak}{h}\right| < 1$.
Note that we cannot pertub the eigenvalues, because $k\lambda_p(A) \in \partial S$.

Lax-Friedrichs:
Suppose we take finite difference equation using Euler in time:

$$u_j^{n+1} = u_j^n - \frac{ak}{2h}(u_{j+1}^n - u_{j-1}^n)$$

Relace $u_j^n = \frac{1}{2}(u_{j-1}^n + u_{j+1}^n)$. Rewrite as

$$\frac{u_j^{n+1} - u_j^n}{k} + a\left(\frac{u_{j+1}^n - u_{j-1}^n}{2h}\right) = \frac{h^2}{2k}\left(\frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2}\right) \Rightarrow u_t + a u_x = \epsilon u_{xx}$$

$u'(t) = A_\epsilon u(t)$, where $A_\epsilon = A + \frac{\epsilon}{h^2}$ (a $[1, -2, 1]$ banded matrix with 1 at the top right and bottom left cell), $A = -\frac{a}{2h}[-1, 0, 1]$ with 1 at top right and bottom left cell. Eigenvalues are:

$$\lambda_p(A_\epsilon) = -\frac{ia}{h}\sin(2\pi p h) - \frac{2\epsilon}{h^2}(1 - 2\cos(2\pi p h)), p = 1, ..., m + 2$$

$z = k\lambda_p$ is an ellipse centered at $-\frac{2k\epsilon}{h^2}$ with width $\frac{4k\epsilon}{h^2}$ height $\frac{ak}{h}$. Choose $\epsilon = \frac{h^2}{2k}$. We rescale width to 1. If $\left|\frac{ak}{h}\right| \leq 1$, then it is stable. This is a first order method $\mathcal{O}(k+h)$.

Lax-Wendroff:
Recall $u'(t) = Au(t)$. Using Taylor series expansion for $y' = f(y)$,

$$\begin{aligned}
y^{n+1} &= y^n + kf(y^n) + \frac{k^2}{2}f'(y^n) + \mathcal{O}(k^3) \\
u^{n+1} &= u^n + kAu^n + \frac{1}{2}k^2A^2u^n \\
u_j^{n+1} - u_j^n &= -\frac{ak}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{a^2k^2}{8h^2}(u_{j-2}^n - 2u_j^n + u_{j+2}^n) \\
&= -\frac{ak}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{a^2k^2}{2h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n)
\end{aligned}$$

Write this as an Euler method discretization of $u'(t) = A_\epsilon u(t)$. Get
$k\lambda_p(A_\epsilon) = -i\frac{ak}{h}\sin(p\pi h) + \left(\frac{ak}{h}\right)^2(\cos(p\pi h) - 1)$, so we require $\left|\frac{ak}{h}\right| \leq 1$ for stability. This is a second order method $\mathcal{O}(k^2 + h^2)$.

Upwind Method:
Sometimes, we should only use information from left (previous data) j and $j-1$, but we can add symmtric term to achieve this:

$$\begin{aligned}
u_j^{n+1} &= u_j^n - \frac{ak}{h}(u_j^n - u_{j-1}^n) \\
&= u_j^n - \frac{ak}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{ak}{2h}(u_{j+1}^n - 2u_j^n + u_{j-1}^n)
\end{aligned}$$

Let $\epsilon = \frac{ah}{2}$. Like before, we require $\left|\frac{ak}{h}\right| \leq 1$ and $-2 < -\frac{2\epsilon k}{h^2} < 0$.

$\epsilon_{LF} = \frac{h^2}{2k}$, $\epsilon_{LW} = \frac{2k^2}{2}$, $\epsilon_{UP} = \frac{ah}{2}$.
For LF, we get $-2 < -1 < 0$. For LW, we get $\left|\frac{ak}{h}\right| \leq 1$. For UP, we need $a > 0$, so solution moves left to right.

CFL Condition:
Recall that $u(x,t) = \eta(x-at)$, $u(x_j - ak, t)$ at time t shifts to $u(x_j, t+k)$ at time $t+k$.

***Definition:* 5.2: Courant-Friedrichs-Lewy**

If u_j^{n+1} is computed from $u_{j+p}^n, u_{j+p+1}^n, ..., i_{j+q}^n$, then the Courant-Friedrichs-Lewy (CFL) condition says that $x_{j+p} \leq x_j - ak \leq x_{j+q}$ for the method to be stable. Since $x_j = jh$, it is equivalent to $-q \leq \frac{ak}{k} \leq -p$.

In what we have considered $p = -1$, $q = 1$.

6 Finite Element Method

Consider the BVP:

$$-u''(x) = f(x)$$
$$u(0) = 0, u'(1) = 0$$

6.1 Weak Formulation of BVPs

Define an inner product of functions $f, g \in L^2[0,1]$ by $\langle f, g\rangle = \int_0^1 f(x)g(x)dx$. Consider the BVP. Let v be some smooth function:

$$\langle f, v\rangle = \int_0^1 f(x)v(x)dx = \int_0^1 -u''(x)v(x)dx$$
$$= u'(0)v(0) - u'(1)v(1) + \int_0^1 u'(x)v'(x)dx$$

Suppose $v(x)$ satisfies $v(0) = 0$. Then with $u'(1) = 0$, we have:

$$\langle f, v\rangle = \int_0^1 u'(x)v'(x)dx$$

Define $a(u,v) = \int_0^1 = \int_0^1 u'(x)v'(x)dx$, let $V = \{v \in L^2[0,1] : a(u,v) < \infty, v(0) = 0\}$. The condition $a(u,v) < \infty$ implies $v' \in L^2[0,1]$, because $a(v,v) = \int_0^1 (v'(x))^2 dx$. Then for any $u(x)$ satisfying $-u''(x) = f(x)$ and any $v(x) \in V$, $\langle f, v\rangle = a(u,v)$.

Definition: 6.1: Variational BVP

Define a solution of the BVP to be any function satisfying $\langle f, v\rangle = a(u,v)$ for all test functions $v \in V$ and $u(0) = 0$. This is called the variational or weak formulation of the original BVP.

Dirichlet conditions $u(0) = 0$ are called essential BCs.
Neumann conditions $u'(1) = 0$ are called natural BCs, because they are encoded in weak formulation, and we don't have to enforce it.

Theorem: 6.1:

Suppose $f \in C^0[0,1]$, $u \in C^2[0,1]$ and f and u satisfies $\langle f, v\rangle = a(u,v)$ for any $v \in V$, $u(0) = 0$. Then u is a strong solution.

Proof.

$$\langle f, v\rangle = a(u,v) = \int_0^1 u'(x)v'(x)dx = -\int_0^1 u''(x)v(x)dx + u'(1)v(1)$$

Thus $\langle f + u'', v\rangle = u'(1)v'(1)$. Note that $\int_0^1 u''(x)v(x)dx = \langle u'', v\rangle$.
Since $u'(1) = 0$, then $f + u'' = 0$ on $[0,1]$.

If not, say $f + u'' > \epsilon$ on (x_0, x_1) for some $\epsilon > 0$. Since $f + u'' \in C^0[0,1]$, we can find some $v \in V$, $v(1) = 0$ s.t. $\langle f + u'', v\rangle > 0$. Take v a Schwartz function s.t. $v > 0$ on (x_0, x_1). This is a contradiction, since $\langle f + u'', v\rangle = u'(1)v(1) = 0$.

Take $v(x) = x$. Then $\langle f + u'', v\rangle = 0 = u'(1)v(1) = u'(1)$, so $u'(1) = 0$. □

Ritz-Galerkin Approximation
Let $S \subset V$ be a finite dimensional subspace. Consider the following problem: Find $u_s \in S$ s.t. $a(u_s, v) = \langle f, v \rangle$ for all $v \in S$. Then we have a discretized BVP.

***Theorem:* 6.2: Ritz-Galerkin**

Given $f \in L^2[0,1]$, $a(u_s, v) = \langle f, v \rangle$ has a unique solution.

Proof. Let $\{\phi_i\}_{i=1}^n$ be a basis for S. Then we can write $u_s = \sum_{j=1}^n u_j \phi_j$. For any $v \in S$:

$$a(u_s, v) = \sum_{j=1}^n u_j a(\phi_j, v)$$

Consider $v = \phi_i$. Then $\sum_{j=1}^n u_j a(\phi_j, \phi_i) = \langle f, \phi \rangle$. Let $K_{ij} = a(\phi_j, \phi_i)$ and $F_i = \langle f, \phi_i \rangle$. We get $KU = F$.

Suppose K is singular. Then for each j, there exists $v_i \neq 0$ s.t.

$$a(v, \phi_j) = \sum_{i=1}^n v_i K_{ij} = \sum_{i=1}^n v_i a(\phi_i, \phi_j) = 0,$$

where $v = \sum_{i=1}^n v_i \phi_i$.
But then $a(v,v) = \int_0^1 (v'(x))^2 = 0$, $v'(x) = 0$, $v(x)$ must be constant.
Recall that $v \in S \subset V$, $v(0) = 0$, then $v(x) = 0$. Contradiction, so K is non-singular $\det K \neq 0$. □

K is called *stiffness matrix*. For $\begin{cases} \nabla^4 u = f, x \in \Omega \\ \nabla^2 u = 0, u = 0, x \in \partial\Omega \end{cases}$, K is symmetric p.s.d.

Error Estimates for Weak Solutions:
How do we know that u_s is close to u?
We compare the two values $a(u_s, v) = \langle f, v \rangle$ for $v \in S$ and $a(u, v) = \langle f, v \rangle$ for $v \in V$.
Since $S \subset V$, subtracting the two in S, we get $a(u - u_s, w) = 0$ for all $w \in S$.

***Theorem:* 6.3:**

Let $\|v\|_E = \sqrt{a(v,v)}$, $\|u - u_s\|_E = \min_{v \in S} \|u - v\|_E$. Error is optimal approximation to u in S in $\|\cdot\|_E$-norm

We will show that if $S_h \subset$ is our finite difference space, then $\|u - u_s\|_E \leq h \|u''\|_{L^2}$, then we will also have $\|u - u_s\|_{L^2} \leq h^2 \|u''\|_{L^2}$.

Approximation Assumption: choose S_h s.t. $\min_{v \in S_h} \|w - v\|_E \leq h \|w''\|_{L^2}$ for all $w \in C^2[0,1] \cap V$. Also gives us $\|u - u_s\|_{L^2} \leq h^2 \|u''\|_{L^2}$ if $u \in C^2[0,1] \cap V$.

6.2 Piecewise Polynomial Spaces and Finite Elements

Let $0 = x_0 < x_1 < \cdots < x_n = 1$. Let S be space of functions s.t. $v \in C^0[0,1]$, $v(0) = 0$ and v is linear on $[x_i, x_{i+1}]$. $S \subset V$.

Let $\phi_i \in S$ be defined s.t. $\phi_i(x_j) = \delta_{ij}$. Then $\{\phi_i\}$ is a basis for S.

$$\phi_i = \begin{cases} \frac{1}{x_i - x_{i-1}}(x - x_{i-1}), x \in [x_{i-1}, x_i] \\ -\frac{1}{x_{i+1} - x_i}(x - x_i) + 1, x \in [x_i, x_{i+1}] \end{cases}$$

$\{\phi_i\}$s are called nodal basis and $x_0, x_1, ..., x_n$ are called nodes. We can define an *interpolant* $v_I \in S$ for $v \in C^0[0,1]$ by

$$v_I(x) = \sum_{i=1}^{n} v(x_i)\phi_i(x)$$

Theorem: 6.4:

$\{\phi_i\}$ spans S.

Proof. If $v \in S$ means that $v = v_I$, then we are done.

Since $v - v_I$ is piecewise linear by defintion, and vanishes at all $\{x_j\}$, it must be identically zero. □

Theorem: 6.5:

Let $h = \max_i(x_i - x_{i-1})$, then $\|u - u_I\|_E \leq Ch\|u''\|$ for all $u \in C^2[0,1] \cap V$, where C is independent of h and u.

Proof. Taylor expansion yields piecewise linear approximation to u and the error is $\mathcal{O}(h^2)$ and depends on $\|u''\|$. □

Recall that $\sum_{j=1}^{n} u_j a(\phi_j, \phi_i) = \langle f, \phi_i \rangle$, $i = 1, 2, ..., n$ becomes $KU = F$. We can show that $K_{i,i+1} = h_i^{-1} + h_{i+1}^{-1}$, $K_{i,i+1} = K_{i+1,i} = -h_{i+1}^{-1}$, $K_{nn} = h_n^{-1}$, where $h_i = x_i - x_{i-1}$, $KU = F$ becomes:

$$-\frac{2}{h_i + h_{i+1}}\left(\frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}\right) = \frac{2f_i}{h_i + h_{i+1}} = f(x_i) + \mathcal{O}(h)$$

For $h_i = h$ for all i, we get:

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f(x_i) + \mathcal{O}(h),$$

which is the same as finite difference method.

6.3 Sobolev Space

***Definition:* 6.2: L^p Spaces**

Suppose $\Omega \subset \mathbb{R}^n$ is some domain (open simply connected set). Let $f : \Omega \to \mathbb{R}$.

$$\|f\|_{L^p} = \left(\int_\Omega |f(x)|^p dx \right)^{1/p}$$
$$\|f\|_{L^\infty} = \operatorname{esssup} \{|f(x)| : x \in \Omega\}$$

$L^p(\Omega)$ defined by

$$L^p(\Omega) = \{f : \|f\|_{L^p} < \infty\}$$

Minkowski Inequality:

$$\|f + g\|_{L^p} \leq \|f\|_{L^p} + \|g\|_{L^p}$$

Holder's Inequality: For $1 \leq p, q \leq \infty$, $\frac{1}{p} + \frac{1}{q} = 1$

$$\|fg\|_{L^1} \leq \|f\|_{L^p} \|g\|_{L^q}$$

Cauchy Schwartz Inequality:

$$\int_\Omega |f(x)g(x)|dx \leq \|f\|_{L^2} \|g\|_{L^2}$$

***Definition:* 6.3: Banach Space**

Let V be a vector space, a norm is a function $V \to \mathbb{R}$ s.t.
1. $\|v\| \geq 0$ and $\|v\| = 0 \Leftrightarrow v = 0$
2. $\|cv\| = |c| \|v\|$ for c a scaler
3. $\|v + w\| \leq \|v\| + \|w\|$

A vector space equipped with a norm $\|\cdot\|$ is a normed vector psace. A Banach space is a complete normed vector space.

***Theorem:* 6.6:**

For $1 \leq p \leq \infty$, $L^p(\Omega)$ is a Banach space.

Consider $V = \left\{ v : [0,1] \to \mathbb{R} : v(0) = 0, a(v,v) = \int_0^1 (v')^2 dx < \infty \right\}$.

***Definition:* 6.4: Multi-index**

Let α be an n-tupple $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$. Define the length of α by $|\alpha| = \sum \alpha_i$. Denote $D^\alpha \phi$, $\partial^\alpha \phi$, $\phi^{(\alpha)}$ the partial derivatives

$$\left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \left(\frac{\partial}{\partial x_2} \right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_n} \right)^{\alpha_n} \phi$$

***Theorem:* 6.7: Heine Borel**

Ω is compact if and only if it is closed and bounded.

***Definition:* 6.5: Compact Support**

Let $f : \Omega \to \mathbb{R}$, the support of f is

$$\text{supp}(f) = \overline{\{x \in \Omega : f(x) \neq 0\}}$$

A function $f : \Omega \to \mathbb{R}$ is compactly supported if $\text{supp}(f)$ is compact and $\text{supp}(f) \subset \Omega$.

***Definition:* 6.6: Compactly Supported Functions**

Suppose Ω is a domain in $\mathbb{R}^n$. Denote by $D(\Omega)$ or $C_0^\infty(\Omega)$ set of all C^∞ functions with compact support in Ω.

***Definition:* 6.7: Locally Integrable Functions**

The set of locally integrable functions is

$$L^1_{loc}(\Omega) = \left\{ f : f \in L^1(K), K \subset \Omega \text{ compact} \right\}$$

***Definition:* 6.8: Weak Derivatives**

Let $f \in L^1_{loc}(\Omega)$, f has a weak derivative $D^\alpha_w f$ if there exists a function $g \in L^1_{loc}(\Omega)$ s.t. for all $\phi \in C_0^\infty(\Omega)$,

$$\int_\Omega g(x)\phi(x)dx = (-1)^\alpha \int_\Omega f(x)\phi^{(\alpha)}(x)dx$$

We call g the weak derivative of f, $D^\alpha_w f$.

***Definition:* 6.9: Sobolev Space**

Let k be a nonnegative integer, $f \in L^1_{loc}(\Omega)$. Suppose that $D^\alpha_w f$ exists for all $|\alpha| \leq k$. The Sobolev norm $\|\cdot\|_{W^k_p(\Omega)}$ is defined by

$$\|f\|_{W^k_p(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha_w f\|^p_{L^p(\Omega)} \right)^{1/p}, 1 \leq p < \infty$$
$$\|f\|_{W^k_\infty(\Omega)} = \max_{|\alpha| \leq k} \|D^\alpha_w f\|_{L^\infty(\Omega)}$$

The Sobolev spaces are defined by:

$$W^k_p(\Omega) = \left\{ f \in L^1_{loc}(\Omega) : \|f\|_{W^k_p(\Omega)} < \infty \right\}$$

***Theorem:* 6.8:**

Sobolev spaces $W^k_p(\Omega)$ are Banach spaces.

Theorem: 6.9: Traces Theorem

Suppose $\Omega \subset \mathbb{R}^n$ is bounded and has a Lipschitz continuous k-times differentiable boundary $\Gamma \in C^{k,1}$. Let $\gamma u = u|_\Gamma$ for $u \in C^\infty(\overline{\Omega})$. If $s \leq k+1$, $s - \frac{1}{p} \notin \mathbb{Z}$, and $s - \frac{1}{p} = l + \sigma$ for $l \in \mathbb{Z}$ and $0 < \sigma < 1$, then the map $u \mapsto \left(\gamma u, \gamma \frac{\partial u}{\partial v}, ..., \gamma \frac{\partial^l u}{\partial v^l}\right)$ has a unique continuous extension:

$$W_p^s(\Omega) \mapsto \prod_{j=0}^{l} W_p^{s-j-\frac{1}{p}}(\Gamma)$$

Example: $W_2^1 \mapsto W_2^{\frac{1}{2}}$, where $W_2^{\frac{1}{2}}$ is for fractionally derivatives defined by Gamma Function.

Theorem: 6.10:

1. If $k \leq m$ and $1 \leq p \leq \infty$, then $W_p^m \subset W_p^k$
2. If $1 \leq p \leq q \leq \infty$, then $W_q^k(\Omega) \subset W_p^k(\Omega)$

Theorem: 6.11: Sobolev Inequality

Let k be a positive integer and $1 \leq p < \infty$. Suppose that $k \geq n$ if $p = 1$ and $k > \frac{n}{p}$ if $p > 1$. Then there exists a constant C s.t. for all $u \in W_p^k(\Omega)$,

$$\|u\|_{L^\infty(\Omega)} \leq C \|u\|_{W_p^k(\Omega)}$$

Furthermore, u can be considered as a function in $C^0(\Omega)$, meaning that there exists a $\tilde{u} \in C^0(\Omega)$ s.t. $\|u - \tilde{u}\|_{L^\infty(\Omega)} = 0$.

Corollary 1. *Let k, m be positive integers s.t. $m < k$ and $1 \leq p < \infty$. Suppose $k \geq n + m$ if $p = 1$ and $k > m + \frac{n}{p}$ if $p > 1$. Then $\|u\|_{W_\infty^m(\Omega)} \leq C \|u\|_{W_p^k}(\Omega)$. u can be considered as a function in $C^m(\Omega)$.*

Definition: 6.10: Dual Space

Consider a Banach space $\mathcal{B}$. A linear function $L : \mathcal{B} \to \mathbb{R}$ is continuous if and only if L is bounded:

$$\|L\| = \sup_{v \in \mathcal{B}, \|v\|=1} |L(v)| < \infty$$

L is called a functional. The collection of all bounded linear functionals on $\mathcal{B}$ is also a Banach space under the operator norm. Denote as $\mathcal{B}'$ (dual space of $\mathcal{B}$)

Definition: 6.11: Hilbert Space

Let V be a vector space, $b : V \times V \to \mathbb{R}$ a bilinear and symmetric function satisfying: $b(v, v) \geq 0$ for all $v \in V$ and $b(v, v) = 0 \Leftrightarrow v = 0$. Then b is an inner product and V is an inner product space.
If an inner product space V is complete, then it is a Hilbert space. Each Hilbert space with $\|v\| = \sqrt{b(v, v)}$ is a Banach space.

- $L^2(\Omega)$ is a Hilbert space with inner product $\langle f, g \rangle = \int_\Omega f(x)g(x)dx$
- $H^k(\Omega) = W_2^k(\Omega)$ is a Hilbert space with inner product $\langle f, g \rangle_k = \sum_{|\alpha| \leq k} \langle D_w^\alpha f, D_w^\alpha g \rangle_{L^2(\Omega)}$

Theorem: 6.12: Riesz Representation

If L is a continuous linear functional on a Hilbert space $\mathcal{H}$, then there exists a unique $u \in H$ s.t. $Lv = \langle u, v \rangle$ for all $v \in \mathcal{H}$. Furthermore, $\|L\|_{H'} = \|u\|_H$.

Definition: 6.12: Continuous and Coercive Functionals

Let $a : V \times V \to \mathbb{R}$ be a symmetric bilinear form in a normed vector space $(V, \|\cdot\|)$. a is bounded/continuous if $\exists C > 0$ s.t. $|a(u, v)| \leq C \|u\| \|v\|$. a is coercive if $\exists \alpha > 0$ s.t. $\alpha \|v\|^2 < a(v, v)$.

Theorem: 6.13:

If H is a Banach space, $a : V \times V \to \mathbb{R}$ is a symmetric bilinear form that is continuous on H and coercive on a closed subspace $V \subset H$, then $(V, a(\cdot, \cdot))$ is a Hilbert space.

To define a **symmetric variational problem**, we need

1. H is a Banach space
2. $V \subset H$ is a closed subspace
3. $a : H \times H \to \mathbb{R}$ is bounded symmetric bilinear form on H and coercive on V

The problem is then: Given $F \in V'$, find $u \in V$ s.t. $a(u, v) = F(v)$ for all $v \in V$. If all three properties are satisfied, then Theorem 6.12 implies that there is a unique solution $u \in V$, where

$$V = \left\{ v : [0, 1] \to \mathbb{R} : a(v, v) = \int (v')^2 dx < \infty, v(0) = 0 \right\} = \left\{ v \in H^1[0, 1] : v(0) = 0 \right\}$$

Ritz-Galerkin Approximation Problem:
Suppose that $V_h \subset V$ is a finite dimensional subspace. Given $F \in V'$, find $u_h \in V_h$ s.t. $a(u_h, v) = F(v)$ for all $v \in V_h$. By Theorem 6.12, it has a unique solution $u_h \in V_h$. Then $a(u - u_h, v) = 0$ for all $v \in V_h$, so $\|u - u_h\|_E = \min_{v \in V_h} \|u - v\|_E$, where $\|v\|_E = \sqrt{a(u, v)}$. Coercivity and boundedness of a implies that $\|u - u_h\|_V = C \min_{v \in V_h} \|u - v\|_V$.

Theorem: 6.14: Ritz-Galerkin

Suppose u is a solution to a symmetric variational problem and u_h is the Ritz-Galerkin approximation. Then

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \min_{v \in V_h} \|u - v\|_V ,$$

where $\|\cdot\|_V$ is the norm on $V \subset H$. C is the boundedness constant and α is related to the coercivity constant.

Proof. Since a is coercive

$$\|u - u_h\|_V \leq \frac{1}{\sqrt{\alpha}} \|u - u_h\|_E = \frac{1}{\sqrt{\alpha}} \min_{v \in V_h} \|u - v\|_E$$

Since a is bounded:

$$\|u-u_h\|_V \leq \frac{1}{\sqrt{\alpha}} \min_{v\in V_h} \|u-v\|_E \leq \frac{C}{\sqrt{\alpha}} \min_{v\in V_h} \|u-v\|_V$$

□

To define a **nonsymmetric variational problem**, we need

1. $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is a Hilbert space.
2. $V \subset \mathcal{H}$ is a closed subspace
3. $a : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is bilinear, not necessarily symmetric
4. a is continuous on V
5. a is coercive on V

The last two points mean that $\exists \alpha > 0, C > 0$ s.t. $\alpha \|v\|^2 \leq a(v,v) \leq C\|v\|^2$

***Theorem:* 6.15: Lax-Milgram**

Given a Hilbert space $(V, \langle\cdot,\cdot\rangle)$, a continuous bilinear coercive $a : V \times V \to \mathbb{R}$ and $F \in V'$, there eixsts a unique solution $u \in V$ s.t. $a(u,v) = F(v)$ for all $v \in V$.

***Theorem:* 6.16: Cea**

$$\|u-u_h\|_V \leq \frac{C}{\alpha} \min_{v\in V_h} \|u-v\|_h$$

Suppose we have a variaional problem, symmetric or non-symmetric on a subset $V \subset H^1(\Omega)$ with solution $u \in V$ and Ritz-Galerkin approximation $u_h \in V_h \subset V$ in finite dimensional subspace V_h.
Let $I^h : V \cap C^k(\Omega) \to V_h$ be an interpolation operator s.t. $(I^h)^2 = I^h$. Suppose

$$\left\|u - I^h u\right\|_{H^1(\Omega)} \leq Ch^{m-1}|u|_{H^m}(\Omega), \text{ where } |u|_{H^m(\Omega)} = \left(\sum_{|\alpha|=m} \|D^\alpha_w u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}}$$

is a semi-norm on H^m. Since $I^h u \in V_h$, $\|u-u_h\|_{H^1(\Omega)} \leq Ch^{m-1}|u|_{H^m}(\Omega)$ for all $u \in V \cap C^k(\Omega)$.

6.4 Finite Element Space

For the variational problem $a(u,v) = F(v)$ for $v \in V$, we want to find $V_h \subset V$.

***Definition:* 6.13: Finite Element**

Suppose
1. $K \subset \mathbb{R}^n$ is a compact set with piecewise smooth boundary and non-empty interior, called element
2. $\mathcal{P}$ is a finite dimensional space of functions on K, called shape functions
3. $\mathcal{N} = \{N_1, ..., N_k\}$ is a basis for $\mathcal{P}'$ (The dual space with functionals) called nodal variables.

Then $(K, \mathcal{P}, \mathcal{N})$ is a finite element

Definition: 6.14: Nodal Basis

Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element. The basis $\{\phi_1, ..., \phi_k\}$ of $\mathcal{P}$, dual to $\mathcal{N}$ $(N_i(\phi_j) = \delta_{ij})$, is called nodal basis.

Lemma: 6.1:

Let $\mathcal{P}$ be a d-dimensional vector space. $\{N_1, ..., N_d\} \subset \mathcal{P}'$. Then the following are equivalent:

1. $\{N_1, ..., N_d\}$ is a basis for $\mathcal{P}'$
2. If $v \in P$, $N_i v = 0$ for all $i = 1, ..., d$, then $v = 0$

Triangular Elements:
Let K be a triangle. Let P_k denote set of polynomials in two variables upto order k.
$\dim(P_k) = \frac{1}{2}(k+1)(k+2)$.

Lagrange Elements:
$\mathcal{P} = P_1$, $\dim(P_1) = 3$. $\mathcal{N} = \{N_1, N_2, N_3\}$, $N_i(v) = v(z_i)$, where z_1, z_2, z_3 are vertices.
$\mathcal{P} = P_2$, $\dim(P_2) = 6$. z_4, z_5, z_6 are midpoints.

Suppose the variational problem is $a(u, v) = F(v)$, and there are three nodal basis ϕ_1, ϕ_2, ϕ_3 in one element, then the element matrix and the RHS are:

$$K^e = \begin{bmatrix} a(\phi_1, \phi_1) & a(\phi_1, \phi_2) & a(\phi_1, \phi_3) \\ a(\phi_2, \phi_1) & a(\phi_2, \phi_2) & a(\phi_2, \phi_3) \\ a(\phi_3, \phi_1) & a(\phi_3, \phi_2) & a(\phi_3, \phi_3) \end{bmatrix}$$

$$F^e = \begin{bmatrix} F(\phi_1) \\ F(\phi_2) \\ F(\phi_3) \end{bmatrix}$$

6.5 Interpolant

Denote the local interpolant on an element K:

$$I_K v = \sum_{i=1}^{k} N_i(v)\phi_i$$

Proposition: 6.1:

$$N_i(I_K(f)) = N_i(f)$$

Corollary 2. $I_K^2 = I_K$

Definition: 6.15: Subdivision

A subdivision of a domain Ω is a collection of element domains $\{K_i\}$ s.t.

1. $\text{int}(K_i) \cap \text{int}(K_j) = \emptyset$ for $i \neq j$
2. $\cup K_i = \overline{\Omega}$

Definition: 6.16: Global Interpolant

Suppose Ω is a domain with subdivision $\mathcal{T}$ and each $K \in \mathcal{T}$ is associated with a finite element $(K, \mathcal{P}, \mathcal{N})$. Let m be the highest derivative appearing in the nodal variables $\mathcal{N}$ of all of the elements. For $f \in C^m(\overline{\Omega}) = C^m(\mathbb{R}^n)|_\Omega$, the global interpolant is:

$$I_{\mathcal{T}} f|_{K_i} = I_{K_i} f$$

for all $K_i \in \mathcal{T}$.

Note that we don't know if $I_{\mathcal{T}} f \in C^0(\overline{\Omega})$.

Definition: 6.17: Triangulation

A triangulation of a polygon domain Ω is a subdivision consisting of triangles s.t. no vertex of any triangle is in the interior of an edge of another triangle.

Definition: 6.18: Continuity Order

An interpolant has continuity order r if $I_{\mathcal{T}} f \in C^1(\overline{\Omega})$ for all $f \in C^m(\overline{\Omega})$. Call $V_{\mathcal{T}} = \left\{ I_{\mathcal{T}} f : f \in C^m(\overline{\Omega}) \right\}$ a C^r finite element space $r \geq 0$. Also, $I_{\mathcal{T}} f \in W_\infty^{n+1}(\overline{\Omega})$.

One necessary condition for triangles is that each edge must have nodes (nodal variables) that are fixed or symmetric around the midpoints of the edge.

The Lagrange or Hermite elements are both C^0, the Argyris elements are C^1. Lagrange has $m = r = 0$, Hermite $m = 1, r = 0$, Argyris $m = 2, r = 1$.

6.6 Approximation Assumption

We want to estimate $\|I_{\mathcal{T}} f - f\|_{H^1(\Omega)}$ for $f \in C^m(\overline{\Omega})$.

Definition: 6.19: Star-Shaped

A region Ω is star-shaped w.r.t. some ball B if for all $x \in \Omega$, the closed convex hall of $\{x\} \cup B \subset \Omega$.

The Taylor polynomial of order m evaluated at y is given by

$$(T_y^m u)(x) = \sum_{|\alpha| < m} \frac{1}{\alpha!} D^\alpha u(y)(x - y)^\alpha,$$

where $\alpha = (\alpha_1, ..., \alpha_n)$ is a multi index and $\alpha! = \alpha_1! \cdots \alpha_n!$.

If $u \in W_p^{m-1}(\Omega)$, deerivatives are defined a.e., but cannot necessarily be evaluated pointwise. If $u \notin C^{m-1}(\overline{\Omega})$, then the Taylor series does not make sense pointwise. However, it is defined if we average over a ball B. Define

$$Q^m u(x) = \int_B T_y^m u(x) \phi(y) dy,$$

where $\phi \in C^\infty(\Omega)$ and $\phi = 0$ outside $\overline{B}$.

Let $\rho_{\max} = \sup \{\rho : \Omega \text{ is star-shaped w.r.t. a ball of radius } \rho\}$. Let $d = \operatorname{diam}(\Omega)$. If Ω is star-shaped, we can bound $u - Q^m u$.

***Lemma:* 6.2: Bramble-Hilbert**

Let B be a ball in Ω s.t. Ω is star-shaped w.r.t. B_1 with radius $\rho > \frac{1}{2}\rho_{\max}$. Suppose that $u \in W_p^m(\Omega)$ with $p \geq 1$. Then for $0 \leq k \leq m$, the semi-norms satisfy:

$$|u - Q^m u|_{W_p^k(\Omega)} \leq C d^{m-k} |u|_{W_p^m(\Omega)}$$

As triangles become smaller $d \to 0$, the error $\to 0$.

***Theorem:* 6.17:**

Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element s.t.

1. K is star-shaped w.r.t. some ball
2. $\mathcal{P}_{m-1} \subset \mathcal{P} \subset W_\infty^m(K)$
3. $\mathcal{N} \subset C^l(K)'$

Suppose $1 \leq p \leq \infty$ and either $m > l + \frac{n}{p}$ when $p > 1$ or $m \geq l + n$ when $p = 1$. Then for $0 \leq i \leq m$ and $v \in W_p^m(K)$, we have

$$|v - I_K v|_{W_p^i(K)} \leq C(\text{diam}K)^{m-i} |v|_{W_p^m(K)}$$

Let Ω be a domain, $\{\mathcal{T}^h\}$, $0 \leq h \leq 1$ is a *family of subdivisions* s.t.

$$\max\left\{\text{diam}\mathcal{T} : \mathcal{T} \in \mathcal{T}^h\right\} \leq h\text{diam}(\Omega)$$

***Theorem:* 6.18:**

Let $\{\mathcal{T}^h\}$ be a non-degenerate family of subdivisions of a polyhedral domain $\Omega \subset \mathbb{R}^n$. Let $(K, \mathcal{P}, \mathcal{N})$ be a reference element satisfying the conditions from Theorem 6.17. Suppose $\mathcal{T} \in \mathcal{T}^h$ is affine equivalent to the reference lement. Then $\exists C > 0$, depending only on reference element s.t. for all $0 \leq s \leq m$,

$$\left(\sum_{\mathcal{T} \in \mathcal{T}^h} \left\|v - I_{\mathcal{T}}^h v\right\|_{W_p^s(\mathcal{T})}^p\right)^{\frac{1}{p}} \leq C h^{m-s} |v|_{W_p^m(\Omega)}$$

If the gloabl interpolation $I^h v \in C^r(\overline{\Omega})$ for $r \geq 0$, then it is equivalent to

$$\left\|v - I^h v\right\|_{W_p^s(\Omega)} \leq C h^{m-s} |v|_{W_p^m(\Omega)}$$

6.7 Discontinuous Galerkin Methods

***Definition:* 6.20: Broken Sobolev Space**

The broken Sobolev space is

$$H^k(\Omega, \mathcal{T}^h) = \left\{v \in L^2(\Omega) : v|_K \in H^k(K) \text{ for all } K \in \mathcal{T}^h\right\}$$

Let $\mathcal{F}^h$ denote the set of all faces of elements $K \in \mathcal{T}^h$. Let $\mathcal{F}_B^h$ denote faces on $\partial\Omega$ and $\mathcal{F}_I^h = \mathcal{F}^h \setminus \mathcal{F}_B^h$. Suppose $\Gamma \in \mathcal{F}_I^h$. Let $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)}$ be adjacent faces.

For $v \in H^1(\Omega, \mathcal{T}^h)$, define $v_\Gamma^{(L)} = v|_{K_\Gamma^{(L)}}$, $v_\Gamma^{(R)} = v|_{K_\Gamma^{(R)}}$, the average value is $\langle v \rangle_\Gamma = \frac{1}{2}(v_\Gamma^{(L)} + v_\Gamma^{(R)})$, and the difference is $[v]_\Gamma = v_\Gamma^{(L)} - v_\Gamma^{(R)}$.

Consider $-\nabla^2 u = f$. Write

$$\int_\Omega \nabla^2 u v dx = \int_\Omega f v dx,$$

where $u \in H^2(\Omega)$ and $v \in H^1(\Omega, \mathcal{T}^h)$.

$$\begin{aligned} &\sum_{K \in \mathcal{T}^h} \int_K \nabla u \nabla v dx - \sum_{K \in \mathcal{T}^h} \int_{\partial K} (n \cdot \nabla u) v dx = \int_\Omega f v dx \\ \Leftrightarrow &\sum_{K \in \mathcal{T}^h} \int_K \nabla u \nabla v dx - \sum_{\Gamma \in \mathcal{F}_I^h} \int_\Gamma n \cdot \langle \nabla u \rangle [v] dx = \int_\Omega f v dx + \int_\Gamma n \cdot \nabla u V dx \end{aligned}$$

$a(u, v)$ is defined by LHS, and we can enforce $v = 0$ on Γ.

6.8 Isoparametric Approximations

Suppose that $(K, \mathcal{P}, \mathcal{N})$ is a fixed reference element. Suppose we have another element domain $K_e = F(K)$, for some mapping F and basis functions $\phi_j^e(x) = \phi_j(F^{-1}(x))$. *e.g.* $F = \sum_{j=1}^n \phi_j x_j$ gives an affine transformation.

Suppose that $\tilde{\Omega}$ is a polyhedral domain and let $\tilde{V}_h$ be a finite element space on $\tilde{\Omega}$. Let $\tilde{F} : \tilde{\Omega} \to \Omega$, where Ω is Lipschitz, but not necessarily polyhedral. Then

$$V_h = \left\{ v(\tilde{F}^{-1}(x)) : x \in \tilde{F}(\tilde{\Omega}), v \in \tilde{V}_h \right\}$$

is called an isoparametric equivalent finite elment space when $\tilde{F} \in \tilde{V}_h$.
i.e. to evaluate $v(x)$ for $x \in \Omega$, we transform it to $\tilde{x} = \tilde{F}^{-1}(x) \in \tilde{\Omega}$ and evaluate $v(\tilde{x})$

Let Ω be a domain with smooth boundary and Ω_h is a polyhedral approximation. It is possible to construct piecewise polynomial mapping of degree $k-1$ s.t.

1. it is equal to identity away from $\partial\Omega$
2. distance from $\partial\Omega$ and $\partial F^h(\Omega_h)$ is $\mathcal{O}(h^k)$
3. Jacobians of F are bounded

Then for $0 \le s \le 1$, $k = m - 1$,

$$\left\| v - I^h v \right\|_{W_p^s(F^h(\Omega_h))} \le C h^{m-s} |v|_{W_p^m(F^k(\Omega_h))}$$

7 Integral Equation Methods

Consider the Laplace equation:

$$\nabla^2 u = 0, x \in \Omega$$
$$u(x) = g(x), x \in \partial\Omega$$

Green's function $G(x, y)$ for Laplace equation satisfies:

$$\nabla_x^2 G(x, y) = \delta(x - y), x \in \Omega$$

In 2D:

$$G(x, y) = \frac{1}{2\pi} \log \|x - y\|$$
$$u(x) = \int_{\partial\Omega} G(x, y)\sigma(y)dy, x \in \Omega$$
$$\lim_{x\to\partial\Omega} u(x) = \int_{\partial\Omega} G(x, y)\sigma(y)dy = g(x)$$

Proof. Rewrite the Laplace equation in polar coordinates:

$$\nabla^2 f = \frac{1}{r}\left(\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right)\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2}$$

Assume $\frac{\partial^2 f}{\partial\theta^2} = 0$ (no dependence on θ), then

$$\nabla^2 G(\rho) = \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial G}{\partial\rho}\right)$$

Setting it to zero for $\rho > 0$, we have an ODE, which gives:

$$G(\rho) = c_1 \log\rho + c_2$$

With the condition: $\int_B \nabla_x^2 G(x, y)dx = 1$, we get $G(\rho) = \frac{1}{2\pi}\log\rho$ □

In 3D, $G(x, y) = -\frac{1}{4\pi}\frac{1}{\|x-y\|}$

Idea: we solve for the equation along the boundary using Green's function.
Issues:

1. Singularity of Green's function
2. The matrix for $G(x, y)$ is dense
3. Condition number for G is large. G is a compact operator, with $\lambda_i \to 0$.

Instead, we can also write:

$$u(x) = \int_{\partial\Omega}\left(\frac{\partial}{\partial n(y)}G(x, y)\right)\sigma(y)dy, x \in \Omega$$
$$\lim_{x\to\partial\Omega} u(x) = -\frac{1}{2}\sigma(x) + \int_{\partial\Omega}\left(\frac{\partial}{\partial n(y)}G(x, y)\right)\sigma(y)dy = g(x), x \in \partial\Omega$$

$\sigma(x)$ is almost identity, while the second term has $\frac{1}{N^2}$ decay. It becomes well-conditioned.

Suppose $\Omega = H = \{(x_1, x_2) : x_2 \geq 0\}$ is the upper half plane. Consider approaching $(0,0)$ from below. Write $x = (x_1, x_2)$, $y = (y_1, y_2)$, and take $x_1 = 0$, $x_2 = h \to 0$

$$\begin{aligned}
G(x,y) &= \frac{1}{2\pi} \log \|x - y\| = \frac{1}{4\pi} \log \left((x_1 - y_1)^2 + (x_2 - y_2)^2 \right) \\
\lim_{x \to 0} \int_{\partial\Omega} G(x,y)\rho(y)dy &= \lim_{x_1, x_2 \to 0} \int_{-\infty}^{\infty} \frac{1}{4\pi} \log \left((x_1 - y_1)^2 + (x_2 - y_2)^2 \right) \rho(y) dy \\
&= \lim_{h \to 0} \int_{-\infty}^{\infty} \frac{1}{4\pi} \log(y_1^2 + h^2) \rho(y_1) dy_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{4\pi} \log y_1^2 \rho(y_1) dy_1
\end{aligned}$$

Consider the second formulation:

$$\begin{aligned}
\frac{\partial}{\partial n(y)} G(x,y) &= \frac{1}{4\pi} \frac{\partial}{\partial y_2} \log \left((x_1 - y_1)^2 + (x_2 - y_2)^2 \right) \\
&= \frac{1}{4\pi} \frac{y_2 - x_2}{(x_1 - y_1)^2 + (x_2 - y_2)^2} \\
\lim_{x \to 0} \int_{\partial\Omega} \frac{\partial}{\partial n(y)} G(x,y)\sigma(y)dy &= \lim_{h \to 0} \int_{-\infty}^{\infty} \frac{1}{2\pi} \frac{-h}{y_1^2 + h^2} \sigma(y_1) dy_1. \\
\text{Define } \Phi_h(\xi) &= \int_{-\infty}^{\xi} \frac{h}{y_1^2 + h^2} dy_1 = \arctan\left(\frac{\xi}{h}\right) + \frac{\pi}{2}
\end{aligned}$$

If $\xi < 0$, then as $h \to 0$, $\lim_{h \to 0} \Phi_h(\xi) = 0$. If $\xi > 0$, we get π.

For any kernel $G(x,y)$, we get

1. First kind integral equation: $g(x) = \int_{\partial\Omega} G(x,y)\sigma(y)dy, x \in \partial\Omega$
2. Second kind integral equation: $g(x) = -\frac{1}{2}\sigma(x) + \int_{\partial\Omega} \left(\frac{\partial}{\partial u(y)} G(x,y) \right) \sigma(y)dy, x \in \partial\Omega$

The integration parts are called Fredholm integral equations.

Let $A[u] = (I + K)[u] = f$. Suppose A_n is a discretization of A. We want $A_n[u_n] = u_n + K_n u_n$. Let u be the true solution so that $A_n u = f_n + \tau$. Let $e_n = u - u_n$ be the error.
Then $A_n e_n = \tau$ or $e_n = A_n^{-1}\tau$. A_n is bounded, τ is from discretization of u and f (quadrature error, $\mathcal{O}(h^n)$)
Boundary is splitted into chunks of size h, each chunk is discretized into n points.

To solve for $u(x)$, we use the boundary condition to solve for the density function $\sigma(y)$ first, and then integrate to get $u(x)$.

7.1 Singular Quadrature

Let $w, z \in \mathbb{R}^2$. It is possible to show

$$\frac{\partial}{\partial n(z)} \log \|w - z\| = \text{Im}\left(\frac{dz}{w - z}\right)$$

Suppose we want to evaluate $\int_C \frac{\rho(z)}{w-z} dz$ along a contour C, where $\rho(z)$ is the density.

Suppose that $\rho(z) \approx \sum_{j=0}^{N} a_j z^j$. Let $p_j = \int_C \frac{z^j}{w-z} dz$. Then if $z_1 = -1$, $z_2 = 1$, $C : z_1 \sim z_2$ (a path from z_1 to z_2). Then,

$$p_0 = \log\left(\frac{w - z_2}{w - z_1}\right)$$
$$p_{j+1} = z p_j + c_j, c_j = \frac{1 - (-1)^j}{j}$$

7.2 Fast Multipole Method

Consider $G(x, y) = \frac{1}{2\pi} \log \|x - y\|$.

$$g(x_i) = -\frac{1}{2}\sigma(x_i) + \sum_{j=0}^{N-1} \left(\frac{\partial}{\partial n(x_j)} G(x_i, x_j)\right) \sigma(x_j) w_j$$

Cost of first term is $\mathcal{O}(N)$, second term is $\mathcal{O}(N^2)$.

However, with iterative method, since the matrix is well-conditioned, the cost can be reduced.

Suppose $\Omega_\sigma, \Omega_\tau \subset \mathbb{R}^2$ are the source and target set, with $|\Omega_\sigma| = N$, $|\Omega_\tau| = M$, and $\Omega_\sigma \cap \Omega_\tau = \emptyset$. $x \in \Omega_\sigma$, $y \in \Omega_\tau$. The Green's function can be approximated as:

$$G(x, y) = \sum_{p=0}^{P-1} B_p(x) C_p(y)$$

Then

$$u_i = \sum_{j=1}^{N} G(x_i, y_j) q_j = \sum_{j=1}^{N} \sum_{p=0}^{P-1} B_p(x_i) C_p(y_j) q_j = \sum_{p=0}^{P-1} B_p(x_i) \left(\sum_{j-1}^{N} C_p(y_j) q_j\right)$$

Write $\hat{q}_p = \sum_{j=1}^{N} C_p(y_j) q_j$. Cost to compute $\hat{q}_p$ is $\mathcal{O}(NP)$.
Then the cost to compute all is $\mathcal{O}(NP + MP)$ instead of $\mathcal{O}(NM)$, where N is the number of sources and M is the number of targets.
We need multi-level/multigrid evaluation to take care of the interactions among grids.

Core idea of fast multipole method is to use the compressed form:

$$G(x_i, y_j) = \sum_{p=0}^{P-1} B_p(x_i) C_p(y_j)$$

to evaluate interaction of well-separated Ω_σ and Ω_τ.

Suppose that we have many source boxes $\Omega_\sigma^{(1)}, ..., \Omega_\sigma^{(K)}$, all well-separated from Ω_τ. Rank of interactions is still P. To compute all $\hat{q}$, it costs $\mathcal{O}(KNP)$. To compute u_i, it costs $\mathcal{O}(KM)$ per box.

Using $u_i = \sum_{p=0}^{P-1} B_p(x_i) \hat{q}_p$, we can find functions $C_p(x)$ and coefficients $\hat{u}_p$ (computed from $\hat{q}_p$) s.t. $u_i = \sum_{p=0}^{P-1} C_p(x_i) \hat{u}_p$. The expansion $\hat{q}$ is called an outgoing expansion, and $\hat{u}$ is an incoming expansion.

1. $T_\sigma^{ofs} : q^\sigma \mapsto \hat{q}^\sigma$ is outgoing-from-source operator

2. $T^{ifo}_{\tau,\sigma} : \hat{q}^\sigma \mapsto \hat{u}^\tau$ is incoming-from-outgoing operator
3. $T^{tfi}_{\tau} : \hat{u}^\tau \mapsto u^\tau$ is target-from-incoming operator, evaluating the local expansion $\hat{u}^\tau$ at x_is to get u^τ.

Instead of computing $A(\Omega_\tau, \Omega_\sigma) : q^\sigma \mapsto u^\tau$ at cost $\mathcal{O}(MN)$, we break into 3 steps

1. $T^{ofs}_{\sigma} : q^\sigma \mapsto \hat{q}^\sigma$ at cost $\mathcal{O}(NP)$
2. $T^{ifo}_{\tau,\sigma} : \hat{q}^\sigma \mapsto \hat{u}^\tau$ at cost $\mathcal{O}(P^2)$
3. $T^{tfi}_{\tau} : \hat{u}^\tau \mapsto u^\tau$ at cost $\mathcal{O}(MP)$

Sketch of the algorithm:

1. Construct multi-level boxes with different scale. Parent box of τ is box on the level above containing τ. Children of τ, $\mathcal{L}^{\text{child}}_\tau$ are boxes on level below. Neighbor list $\mathcal{L}^{\text{nei}}_\tau$ are boxes on the same level touching τ. Interaction list $\mathcal{L}^{\text{int}}_\tau$ are boxes s.t.
   (a) σ, τ are on the same level
   (b) σ and τ do not touch
   (c) parents of σ and τ touches
2. start from bottom level, create $\hat{q}$ for each box
3. go above one level, combine $\hat{q}$ from $\mathcal{L}^{\text{child}}_\tau$ for the parent box
4. construct $\hat{u}$ for each grid from top down and add everything in the interaction list

Operation on each level is linear and grid size is geometric series $\frac{1}{4^n}$.