Numerical Analysis

1. Poisson equation

Problem 1.1 Solve the Poisson equation $\nabla^2 u = f$, $(x,y) \in \Omega$, Ω an open connected domain in $\mathbb{R}^2$ with a Jordan boundary with Dirichlet b.c. $u(x,y) = \phi(x,y)$, $(x,y) \in \partial\Omega$.

We cannot solve $\nabla^2 u = f$, $(x,y) \in \Omega$ directly, but we know how to solve a linear system of eqn

$$Ax = y, \quad A \in \mathbb{R}^{N\times N}, \quad x, y \in \mathbb{R}^N.$$

Consider $f : [0,10] \to \mathbb{R}$. we can approximate f with $\underline{x} = (f(0), f(h), \dots, f(10))$, where $h > 0$ is the step size.

If we approximate u by x, what should be the approximation of ∇? Consider

$$u'(a) = \lim_{h\to 0} \frac{u(a+h) - u(a)}{h}.$$

We replace $\nabla^2 u = f$, $(x,y) \in \Omega$ by a finite-difference formula. If a grid point lies in Ω, all its neighbours are in $\mathrm{cl}\,\Omega$. Since $\nabla^2 = \Delta = \partial_x^2 + \partial_y^2$, we need a finite-difference approximation of second derivatives.

Prop 1 let $g \in C^4[a,b]$, $x \in (a+h, b-h)$, then

$$\Delta_h^2 g(x) := g(x+h) - 2g(x) + g(x+h) = h^2 g''(x) + \frac{1}{12} h^4 g^{(4)}(x) + O(h^6)$$

Pf: Expanding Taylor series,

$$g(x+h) - g(x) = hg'(x) + \frac{1}{2!} h^2 g''(x) + \frac{1}{3!} h^3 g'''(x) + \dots$$
$$g(x-h) - g(x) = -hg'(x) + \frac{1}{2!} h^2 g''(x) - \frac{1}{3!} h^3 g'''(x) + \dots$$

Summing gives $\sum_{k=1}^{m} \frac{2}{(2k)!} h^{2k} g^{(2k)}(x) + O(h^{2m+2})$, and take $m=2$. □

Remark 2 In approximating g'' by $\Delta_h^+ g(x) = g(x-h) - 2g(x) + g(x+h)$, it is useful to know

$$\frac{1}{h^2}\left(g(x-h) - 2g(x) + g(x+h)\right) = g''(x) + \frac{1}{12}h^2 g^{(4)} + \frac{1}{360}h^4 g^{(6)} + O(h^6).$$

Cor 3 The approximation

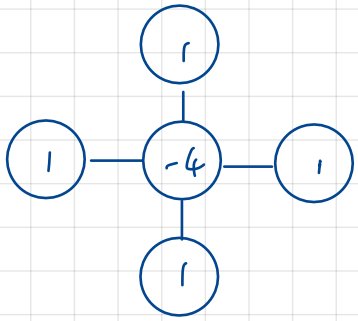
$$(\Delta_{h,x}^2 + \Delta_{h,y}^2)u(x,y) = u(x-h,y) + u(x+h,y) + u(x,y-h) + u(x,y+h) - 4u(x,y)$$
$$\approx h^2 \nabla^2 u(x,y)$$

produces a local error of $O(h^4)$.

Approximation The method justifies the five-point method

$$u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j} = h^2 f_{i,j}, \quad (ih, jh) \in \Omega,$$

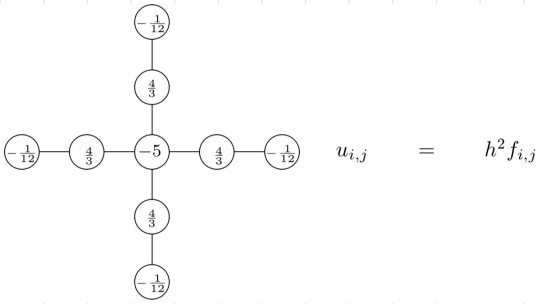
where $f_{i,j} = f(ih, jh)$ given, $u_{i,j} \approx u(ih, jh)$ approximates the exact sol^n. It is denoted by the computational stencil.



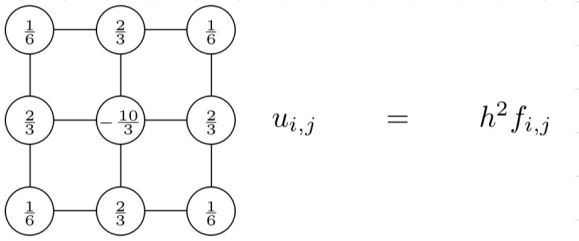
$$u_{i,j} = h^2 f_{i,j}$$

For $(ih, jh) \in \partial\Omega$, we substitute the Dirichlet b.c.

Approximation The stencil



produces a local error of $O(h^6)$. The nine-point method



produces error of $O(h^4)$, but adding $\frac{1}{12}h^4\nabla^2 f$ to RHS with 5-point approximation to $h^2\nabla^2 f$, this has order $O(h^6)$.

The boundary often fails to fit into a square grid, so we may approximate ∇^2 using non-equispaced points.

Example Grid points with spacing h αh, $\alpha \in (0,1]$.

It can be verified that

$$\frac{2}{\alpha+1}g(x-h) - \frac{2}{\alpha}g(x) + \frac{2}{\alpha(\alpha+1)}g(x+\alpha h) = g''(x)h^2 + \frac{1}{3}(\alpha-1)g'''(x)h^3 + O(h^4)$$

with error $O(h^3)$, and $O(h^4)$ if $\alpha = 1$.

If we consider h h αh,

$$\frac{\alpha-1}{\alpha+2}g(x-2h) - \frac{2(\alpha-2)}{\alpha+1}g(x-h) + \frac{\alpha-3}{\alpha}g(x) + \frac{6}{\alpha(\alpha+1)(\alpha+2)}g(x+\alpha h)$$

$$= h^2 g''(x) + O(h^4)$$

The five point formula results in

$$h^2\nabla^2 u(x,y) \approx u(x-h,y) + u(x+h,y) + u(x,y-h) + u(x,y+h) - 4u(x,y).$$

Restricting Ω to a unit square, with $h = \frac{1}{m+1}$, $m \in \mathbb{N}$. We estimate m^2 unknown f^n values $u(ih, jh)_{i,j=1}^{m}$. The RHS of eqn is $h^2 f(x,y)$, then this gives $n \times n$ system with $n = m^2$ unknown i, j.

$$u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j} = h^2 f(ih, jh) \qquad (*)$$

Having ordered grid points, we can write $(*)$ into

$$A\underline{u} = \underline{b}$$

We need to prove that as $h \to 0$, the numerical sol^n $(*)$ tends to exact sol^n of $\nabla^2 u = f$ (with appropriate Dirichlet b.c.).

Example (Natural ordering) A depends on how the grid points are arranged in the 1D array. In the natural ordering, where grids are arranged by columns,

$$A = \begin{pmatrix} B & I & & \\ I & B & \ddots & \\ & \ddots & \ddots & I \\ & & I & B \end{pmatrix}, \qquad B = \begin{pmatrix} -4 & 1 & & \\ 1 & -4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix}$$

Thm 4 (Gershgorin theorem) All evals of an $n \times n$ matrix A are contained in the union of Gershgorin discs in the complex plane.

$$\sigma(A) \subset \bigcup_{i=1}^{n} \Gamma_i, \qquad \Gamma_i := \{ z \in \mathbb{C} : |z - a_{ii}| \le r_i \}, \qquad r_i := \sum_{j \neq i} |a_{ij}|.$$

Pf: For any matrix A, if $A\underline{x} = \lambda \underline{x}$, and $|x_i| = \max |x_j|$. let $\tilde{\underline{x}} = \underline{x}/x_i$, so $|\tilde{x}_j| \le 1$ for $j \neq i$. Since $A\tilde{\underline{x}} = \lambda \tilde{\underline{x}}$,

$$\sum a_{ij} \tilde{x}_j = \lambda \tilde{x}_i = \lambda.$$

So splitting the sum and consider $\tilde{x}_i = 1$,

$$\sum_{j \neq i} a_{ij} x_j + a_{ii} = \lambda.$$

Applying triangle inequality,

$$|\lambda - a_{ii}| = \left| \sum_{j \neq i} a_{ij} x_j \right| \le \sum_{j \neq i} |a_{ij}| |x_j|$$

$$\le \sum_{j \neq i} |a_{ij}| = r_i. \qquad \square$$

Lem 5 For any ordering of grid points, the matrix A of (*) is symmetric and negative definite.

Pf: If $a_{ij} \neq 0$ for $i \neq j$, then i-th and j-th points of the grid (ph, qh) are nearest neighbours, so $a_{ij} \neq 0 \Rightarrow$ $a_{ij} = a_{ji} = 1$, so A has real eval and evec.

Let $A\underline{x} = \lambda \underline{x}$, and let $i = \operatorname{argmax}_j |x_j|$, then reordering $(A\underline{x})_i = \lambda x_i$, we have

$$\underbrace{|(\lambda - a_{ii}) x_i|}_{|\lambda + 4||x_i|} = \underbrace{\left| \sum_{j \neq i} a_{ij} x_j \right|}_{\le 4 x_i} \qquad (\dagger)$$

Here $a_{ii} = -4$, $a_{ij} \in \{0, 1\}$ for $j \neq i$, with at most 4 non-zero elt on RHS. So $\lambda > 0$ is impossible.

Assume $\lambda = 0$, then $|x_i| = |x_j|$ whenever $a_{ij} = 1$. So we can change i to any such j and repeat the argument.

Then every ent of x will have modulus $|x_i|$, but (†) at the boundary of the grid and have fewer than 4 off-diagonal terms is not true. So $\lambda = 0$ impossible, so $\lambda < 0 \Rightarrow A$ -ve def. □

Prop 6 The evals of A are

$$\lambda_{k,l} = -4\left(\sin^2\frac{k\pi h}{2} + \sin^2\frac{l\pi h}{2}\right).$$

where $h = \frac{1}{m+1}$, $k, l = 1, \dots, m$.

Pf: For each (k,l), the vectors

$$V = (V_{ij}), \quad V_{ij} = \sin ix \sin jy \quad . \; x = k\pi h, \; y = l\pi h$$

are evec of A. For $i,j = 1, \dots, m$,

$$\begin{aligned}(AV)_{ij} &= \sin(jy)\left(\sin(ix-x) - 2\sin(ix) + \sin(ix+x)\right) \\ &\quad + \sin(ix)\left(\sin(jy-y) - 2\sin(jy) + \sin(jy+y)\right) \\ &= \sin(jy)\sin(ix)\left(2\cos x - 2\right) + \sin(ix)\sin(jy)\left(2\cos y - 2\right) \\ &= \lambda V_{ij}.\end{aligned}$$

Note that $u_{i\pm1,j}$, $u_{i,j\pm1}$ do not appear for $i,j = 1$ or $i,j = m$. So we should have dropped the corresponding components, but they equals 0 since $\sin(i-1)x = 0$ for $i = 1$, $\sin(i+1)x = 0$ for $i = m$ since $x = \frac{k\pi}{m+1}$, so evals are

$$\begin{aligned}\lambda_{k,l} &= \left(2\cos x - 2\right) + \left(2\cos y - 2\right) \\ &= -4\left(\sin^2\frac{x}{2} + \sin^2\frac{y}{2}\right) = -4\left(\sin^2\frac{k\pi h}{2} + \sin^2\frac{l\pi h}{2}\right).\end{aligned}$$

□

Remark For $1 \le k, l \le m$, we have $\sin x \approx x$, so the eval for discretised laplacian ∇_h^2 are

$$\frac{\lambda_{k,l}}{h^2} \approx -\frac{4}{h^2}\left(\frac{k^2\pi^2h^2}{4} + \frac{l^2\pi^2h^2}{4}\right) = -(k^2+l^2)\pi^2.$$

Recall that the exact evals of ∇^2 are $-(k^2+l^2)\pi^2$, $k,l \in \mathbb{N}$ with ef^n $V_{k,l}(x,y) = \sin k\pi x \sin l\pi y$. So the evec for ∇_h^2 are the values of $V_{k,l}(x,y)$ on grid points, and evals for ∇_h^2 approximates the cts case.

Convergence of 5-point formula

Let $\hat{u}_{i,j} = u(ih, jh)$ and $e_{ij} = u_{i,j} - \hat{u}_{i,j}$ be pointwise error. Set $\underline{e} = (e_{i,j}) \in \mathbb{R}^n$, $n = m^2$ and for $\underline{x} \in \mathbb{R}^n$, $\|\underline{x}\| = \|\underline{x}\|_{l_2}$ be Euclidean norm.

$$\|\underline{x}\|^2 = \sum_{k=1}^{n} |x_k|^2 = \sum_{l=1}^{m}\sum_{j=1}^{m} |x_{i,j}|^2 .$$

Thm Subject to sufficient smoothness of the f^n f and of the boundary conditions, $\exists\, c>0$ indpt of $h = \frac{1}{m+1}$ s.t.

$$\|\underline{e}\| \le ch.$$

Pf: We already know

$$\hat{u}_{i-1,j} + \hat{u}_{i+1,j} + \hat{u}_{i,j-1} + \hat{u}_{i,j+1} - 4\hat{u}_{i,j} = h^2 f_{i,j} + \eta_{i,j} ,$$

where $\eta_{i,j} = O(h^4)$. Subtracting from the approximation,

$$e_{i-1,j} + e_{i+1,j} + e_{i,j-1} + e_{i,j+1} - 4e_{i,j} = \eta_{i,j} .$$

or in matrix form, $A\underline{e} = \underline{\eta}$, A symmetric, so

$$A\underline{e} = \underline{\eta} \Rightarrow \underline{e} = A^{-1}\underline{\eta} \Rightarrow \|\underline{e}\| \le \|A^{-1}\|\,\|\underline{\eta}\| .$$

Since η satisfies $|\eta_{i,j}|^2 < c^2 h^8$, $h = \frac{1}{m+1}$. $\forall i,j = 1,\dots,m$; so

$$\|\underline{\eta}\|^2 \le \sum_{i=1}^{m}\sum_{j=1}^{m} |\eta_{i,j}|^2 \le c^2 m^2 h^8 < c^2 \cdot \frac{1}{h^2} h^8 = c^2 h^6 .$$

$$\Rightarrow \|\underline{\eta}\| \le ch^3$$

Note also A sym, so A^{-1} sym, and $\|A^{-1}\| = \rho(A^{-1})$ is the spectral radius of A^{-1}. $\rho(A^{-1}) = \max_i |\lambda_i|$, λ_i evals of A^{-1}. The evals of A^{-1} are reciprocal of evals of A. So

$$\|A^{-1}\| = \frac{1}{4} \max_{k,l=1,\dots,m} \left(\sin^2 \frac{k\pi h}{2} + \sin^2 \frac{l\pi h}{2} \right)^{-1}$$

$$= \frac{1}{8 \sin^2 \left(\frac{1}{2}\pi h\right)} < \frac{1}{8h^2}.$$

Therefore, $\|e\| \le \|A^{-1}\| \|q\| \le ch$ for some $c > 0$. □

Observation The system $A\underline{u} = \underline{b}$ can be written in

$$\begin{pmatrix} B & I & & \\ I & B & \ddots & \\ & \ddots & \ddots & I \\ & & I & B \end{pmatrix} \begin{pmatrix} \underline{u}_1 \\ \vdots \\ \underline{u}_m \end{pmatrix} = \begin{pmatrix} \underline{b}_1 \\ \vdots \\ \underline{b}_m \end{pmatrix}, \qquad B = \begin{pmatrix} -4 & 1 & & \\ 1 & -4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix}$$

B is a TST-matrix (tridiagonal, symmetric, Toeplitz (const. in diag.)). The evals and orthogonal evecs are

$$B q_l = \lambda_l q_l, \quad \lambda_l = -4 + 2\cos \frac{l\pi}{m+1}, \quad q_l = \gamma_m \left(\sin \frac{jl\pi}{m+1} \right)_{j=1}^{m}$$

where $l = 1, \dots, m$. $\gamma_m = \sqrt{\frac{2}{m+1}}$ is the normalisation vector.

So $B = QDQ^{-1} = QDQ$. $D = \text{diag}(\lambda_l)$. $Q = Q^T = (q_{jl})$.

Note that all $m \times m$ TST matrices share the same full set of evecs so they commute.

Method 8 (The Hockney method) Set $\underline{v}_k = Q\underline{u}_k$, $\underline{c}_k = Q\underline{b}_k$, then

$$\begin{pmatrix} D & I & & \\ I & \ddots & \ddots & I \\ & \ddots & I & D \end{pmatrix} \begin{pmatrix} \underline{v}_1 \\ \vdots \\ \underline{v}_m \end{pmatrix} = \begin{pmatrix} \underline{c}_1 \\ \vdots \\ \underline{c}_m \end{pmatrix}$$

Reordering the grids by rows, so we permute $\underline{v} \mapsto \hat{\underline{v}} = P\underline{v}$.

$\underline{c} \mapsto \hat{\underline{c}} = P\underline{c}$. So we have the new system

$$\begin{pmatrix} \Lambda_1 & & \\ & \ddots & \\ & & \Lambda_m \end{pmatrix} \begin{pmatrix} \hat{\underline{v}}_1 \\ \vdots \\ \hat{\underline{v}}_m \end{pmatrix} = \begin{pmatrix} \hat{\underline{c}}_1 \\ \vdots \\ \hat{\underline{c}}_m \end{pmatrix}.$$

$$\Lambda_k = \begin{pmatrix} \lambda_k & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & \lambda_k \end{pmatrix}_{m \times m}, \quad k = 1, \ldots, m.$$

These are m <u>uncoupled systems</u> $\Lambda_k \hat{\underline{v}}_k = \hat{\underline{c}}_k$. Being tridiagonal, the system can be solved at cost $O(m)$. So the steps and costs are

(1)	$\underline{c}_k = Q\underline{b}_k \quad k = 1, \ldots, m$	$O(m^3)$
(2) Solve	$\Lambda_k \hat{\underline{v}}_k = \hat{\underline{c}}_k$	$O(m)$
(3)	$\underline{u}_k = Q\underline{v}_k$	$O(m^3)$.

(Permutation $\underline{c}_k \mapsto \hat{\underline{c}}_k$. $\underline{v}_k \mapsto \hat{\underline{v}}_k$ basically free).

<u>Method 9</u> (Improved Hockney algorithm) Noting $q_{jl} = \gamma_m \sin \frac{\pi jl}{m+1}$, so

$$\begin{aligned} (Qy)_l &= \sum_{j=1}^{m} \sin \frac{\pi jl}{m+1} y_j \\ &= \mathrm{Im} \sum_{j=0}^{m} \exp\left(\frac{i\pi jl}{m+1}\right) y_j \\ &= \mathrm{Im} \sum_{j=0}^{2m+1} \exp\left(\frac{2i\pi jl}{2m+2}\right) y_j \end{aligned}$$

for $l = 1, \ldots, m$.

Defn 10 (Discrete Fourier transform (DFT)). Let Π_n be the space of bi-infinite complex n-periodic sequences $\underline{x} = \{x_k\}_{k\in\mathbb{Z}}$. s.t. $x_{n+k} = x_n$. Set $\omega_n = \exp\left(\frac{2\pi i}{n}\right)$. The DFT of $\underline{x}$ is $\mathcal{F}_n : \Pi_n \to \Pi_n$ s.t. $\underline{y} = \mathcal{F}_n \underline{x}$, where

$$y_j = \frac{1}{n} \sum_{k=0}^{n-1} \omega_n^{-jk} x_k .$$

$j = 0, \dots, n-1$.

Exercise Prove that $\mathcal{F}_n$ is an iso^m of Π_m onto itself, and

$$\underline{x} = \mathcal{F}_n^{-1} \underline{y}, \quad x_k = \sum_{j=0}^{n-1} \omega_n^{jk} y_k ,$$

$k = 0, \dots, n-1$.

Observation: multiplication by Q can be reduced to calculate an inverse DFT. Since we need to evaluate DFT in a single period, we can do so by multiplying a vector by a matrix at $O(n^2)$ operations.

Fast Fourier Transform (FFT)

Assume n is power of 2, ie. $n = 2m = 2^p$, and for $\underline{y} \in \Pi_{2m}$, denote by

$$\underline{y}^{(E)} = \{y_{2j}\}_{j\in\mathbb{Z}} \qquad \underline{y}^{(O)} = \{y_{2j+1}\}_{j\in\mathbb{Z}}.$$

the even and odd portions of $\underline{y}$. Note $\underline{y}^{(E)}, \underline{y}^{(O)} \in \Pi_m$.

Suppose $\underline{x}^{(E)} = \mathcal{F}_m^{-1} \underline{y}^{(E)}$, $\underline{x}^{(O)} = \mathcal{F}_m^{-1} \underline{y}^{(O)}$, then it is possible to assemble $\underline{x} = \mathcal{F}_{2m}^{-1} \underline{y}$ in a small no. of operations.

Note that $\omega_{2m}^{2m} = 1$ and $\omega_{2m}^2 = \omega_m$, then

$$x_k = \sum_{j=0}^{2m-1} \omega_{2m}^{jk} y_j = \sum_{j=0}^{m-1} \omega_{2m}^{2jk} y_{2j} + \sum_{j=0}^{m-1} \omega_{2m}^{(2j+1)k} y_{2j+1}$$

$$= \sum_{j=0}^{m-1} \omega_m^{jl}\, y_j^{(E)} + \omega_{2m}^{l} \sum_{j=0}^{m-1} \omega_m^{jl}\, y_j^{(0)}$$

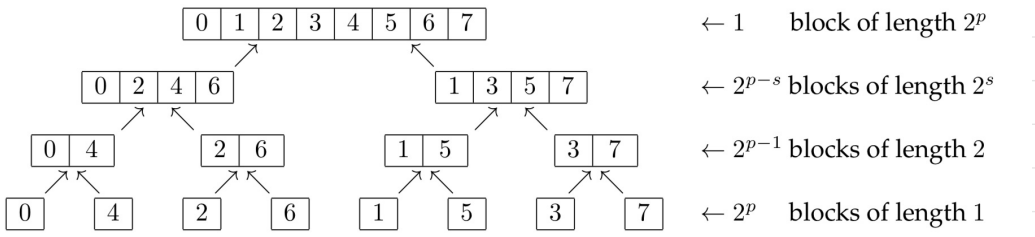
$$= x_l^{(E)} + \omega_{2m}^{l}\, x_l^{(0)},$$

where $l = 0, \dots, m-1$. Therefore, it costs m products to evaluate the first half of $\underline{x}$, provided $\underline{x}^{(E)}$, $\underline{x}^{(0)}$ known. It costs nothing to evaluate the second half, since

$$\omega_m^{j(m+l)} = \omega_m^{jl}, \quad \omega_{2m}^{m+l} = -\omega_{2m}^{l} \Rightarrow x_{m+l} = x_l^{(E)} - \omega_{2m}^{l}\, x_l^{(0)},$$

$l = 0, \dots, m-1$.

Altogether, the cost of FFT is $p 2^{p-1} = \frac{1}{2} n \log_2 n$.



2. Partial Differential Equations of Evolution

Diffusion equation

Consider the sol^n to the diffusion eqn

$$u_t = u_{xx}, \qquad x \in [0,1], \quad t \in [0, \infty).$$

with IC $u(x,0) = u_0(x)$ for $t = 0$ and Dirichlet B.C. $u(0,t) = \phi_0(t)$ at $x = 0$ and $u(1,t) = \phi_1(t)$ at $x = 1$.

Taylor gives

$$u_t = \frac{1}{k}[u(x,t+k) - u(x,t)] + O(k), \qquad k = \Delta t$$

$$u_{xx} = \frac{1}{h^2}[u(x-h,t) - 2u(x,t) + u(x+h,t)] + O(h^2), \qquad h = \Delta h$$

So for the true soln. we obtain

$$u(x,t+k) = u(x,t) + \frac{k}{h^2}\left[u(x-h,t) - 2u(x,t) + u(x+h,t)\right] + O(k^2 + kh^2).$$

This motivates the scheme with $U_m^n \approx u(x_m, t_n)$ on the rectangular mesh $(x_m, t_n) = (mh, nk)$.

$$U_m^{n+1} = U_m^n + \mu\left(U_{m-1}^n - 2U_m^n + U_{m+1}^n\right),$$

$m = 1, \ldots, M$. Here $h = \frac{1}{M+1}$, $\mu = \frac{k}{h^2} = \frac{\Delta t}{(\Delta x)^2}$ is the Courant number. With μ fixed, $k = \mu h^2$, so the local truncation error is $O(h^4)$. Substituting IC U_m^0 and BC U_0^n and U_{M+1}^n, we have enough information to advance in from

$\underline{U}^n := (U_1^n, \ldots, U_M^n)$ to $\underline{U}^{n+1} := (U_1^{n+1}, \ldots, U_M^{n+1})$

Convergence

We say that the method is convergent if, for a fixed μ, and $\forall\, T > 0$, $\lim_{h\to 0} |U_m^n - u(x_m, t_n)| = 0$ uniformly for $(x_m, t_n) \in [0,1] \times [0, T]$.

In other words, if $e_m^n := U_m^n - u(mh, nk)$ is the error of approx. and $\underline{e}^n := (e_1^n, \ldots, e_M^n)$ with $\|\underline{e}^n\| = \|e^n\|_\infty = \max_m |e_m^n|$, then the convergence is equivalent to

$$\lim_{h\to 0} \max_{1 \le n \le T/k} \|\underline{e}^n\| = 0.$$

Thm 2.1 If $\mu \le \frac{1}{2}$, then the above method converges.

Pf: Let $e_m^n := U_m^n - u(mh, nk)$. $\underline{e}^n := (e_1^n, \ldots, e_M^n)$. $\|\underline{e}^n\| = \|\underline{e}^n\|_\infty$.

Cgs equiv to $\lim_{h\to 0} \max_{1 \le n \le T/k} \|\underline{e}^n\| = 0$ for every const. $T > 0$.

Subtracting the above expressions.

$$e_m^{n+1} = e_m^n + \mu\left(e_{m-1}^n - 2e_m^n + e_{m+1}^n\right) + O(h^4).$$
$$= \mu e_{m-1}^n + (1-2\mu) e_m^n + e_{m+1}^n + O(h^4).$$

Then $\|\underline{e}^{n+1}\| = \max_m |e_m^{n+1}| \le (2\mu + |1-2\mu|)\,\|\underline{e}^n\| + ch^4$

$= \|\underline{e}^n\| + ch^4.$

Since $\mu \le \frac{1}{2}$. Since $\|\underline{e}^0\| = 0$, induction yields

$$\|\underline{e}^n\| \le cnh^4 \le \frac{cT}{k} h^4 = \frac{cT}{\mu} h^2 \to 0$$

as $h \to 0$. □

In practice, we choose h, k of comparable size. so $\mu = k/h^2$ is likely to be large.

Stability, Consistency and the Lax equivalence thm

Suppose a numerical method for a PDE can be written in

$$\underline{u}^{n+1} = A\,\underline{u}^n.$$

$\underline{u} \in \mathbb{R}^M$, $A_h = \mathbb{R}^{M\times M}$, $h = \frac{1}{M+1}$. Fix a norm on $\mathbb{R}^M$, and let

$$\|A_h\| = \sup \frac{\|A_h \underline{x}\|}{\|\underline{x}\|}.$$

If we define stability as preserving the boundedness of $\underline{u}^n$ with respect to $\|\cdot\|$, then since

$$\|\underline{u}^n\| \le \|A_h^n\,\underline{u}^0\| \le \|A_h\|^n \|\underline{u}^0\|$$

So $\|A_h\| \le 1$ as $h \to 0 \Rightarrow$ the method is stable.

If we denote the exact sol^n as $\hat{\underline{u}}^n = u(mk, nt)$, then

$$\hat{\underline{u}}^{n+1} = A\,\hat{\underline{u}}^n + \underline{\tau}^n.$$

$\underline{\tau}^n$ is the local truncation error. The error $\underline{e}^n = \underline{u}^n - \hat{\underline{u}}^n$ satisfies

$$\underline{e}^{n+1} = A_h\,\underline{e}^n + \underline{\tau}^n$$

Using $\|A_h\| \le 1$ and assume $\|\underline{e}^0\| = 0$, we get

$$\|\underline{e}_n\| \le \|\underline{\tau}^{n-1}\| + \dots + \|\underline{\tau}^0\|$$

If consistency holds, ie. $\|\underline{\tau}^n\| = O(k^2)$, then $\|\underline{e}^n\| \le nck^2$ for

some const. $c>0$. Since $n \le T/k$ so $\|\underline{e}^n\| \le cTk$, so $\|\underline{e}^n\| \to 0$ as $k \to 0$ uniformly in $n \in [1, T/k]$.

Thm 2.2 (Lax equivalence thm) "consistency + stability = convergence"

Norms

- Sup norm: $\|\underline{u}\| = \|\underline{u}\|_\infty = \max_i |u_i|$

The corresponding norm for a matrix $A \in \mathbb{R}^{m\times m}$ is

$$\|A\|_{\infty\to\infty} := \sup_{\underline{x}} \frac{\|A\underline{x}\|}{\|\underline{x}\|} = \max_i \sum_{j=1}^{M} |A_{ij}|$$

In thm 2.1.

$$A = \begin{pmatrix} 1-2\mu & \mu & & \\ \mu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \mu & 1-2\mu \end{pmatrix}, \quad \|A\|_{\infty\to\infty} = |1-2\mu| + 2\mu \le 1 \text{ if } \mu \le \tfrac{1}{2}$$

- Normalised Euclidean norm: $\|\underline{u}\| = \sqrt{\frac{1}{M}\sum_{i=1}^{M} |u_i|^2}$.

The reason for $\frac{1}{M}$ is to ensure convergence of Riemann sum

$$\|\underline{u}\| := \left(\frac{1}{M}\sum_i |u_i|^2\right)^{1/2} \to \left(\int_0^1 |u(x)|^2 dx\right)^{1/2} = \|u\|_{L_2}.$$

The induced matrix norm is the spectral norm

$$\|A\|_2 := \frac{\|A\underline{x}\|_2}{\|\underline{x}\|_2}$$

The is equal to the largest singular value of A.

Equivalently, $\|A\|_2 = |\rho(AA^T)|^{1/2}$.

$$\rho(M) := \max\{|\lambda| : \lambda \text{ eval of } M\}.$$

Proving Stability directly

Problem 2.3 We will prove stability $\Leftrightarrow \mu \le \frac{1}{2}$. Let $\underline{u}^n = (u_1^n, \dots, u_M^n)^T$

$$u_m^{n+1} = u_m^n + \mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n),$$

$m = 1, \dots, M$.

In matrix form,

$$\underline{u}_h^{n+1} = A_h \underline{u}_h^n \quad . \quad A_h = I + \mu A_* , \quad A_* = \begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}$$

Here A_* TST, $\lambda_\ell(A_*) = -4 \sin^2 \frac{\pi \ell h}{2}$. $\lambda_\ell(A_h) = 1 - 4\mu\lambda_\ell$,

so spectrum lies in $[\lambda_m, \lambda_1]$. Since A_h sym,

$$\|A_h\|_2 = \rho(A_h) = \begin{cases} |1 - 4\mu \sin^2 \frac{\pi h}{2}| < 1 & \mu \le \frac{1}{2} \\ |1 - 4\mu \cos^2 \frac{\pi h}{2}| > 1 & \mu > \frac{1}{2} . \end{cases}$$

We distinguish between

(i) $\mu \le \frac{1}{2}$: $\|\underline{u}^n\| \le \|A\| \cdot \|\underline{u}^{n-1}\| \dots \le \|A\|^n \|u^0\| \le \|\underline{u}^0\|$

as $n \to \infty$ for every $\underline{u}^0$.

(ii) Choose $\underline{u}^0$ as evec for λ s.t. $|\lambda| > 1$. Then

$\underline{u}^n = \lambda^n \underline{u}^0$, unbounded as $n \to \infty$.

Semidiscretisation

Let $u_m(t) = u(mh, t)$, $m = 1, \dots, M$, $t \ge 0$. Approximating ∂_x^2 as before, we deduce from the PDE that the semidiscretisation

$$\frac{du_m}{dt} = \frac{1}{h^2}(u_{m-1} - 2u_m + u_{m+1}) , \qquad (*)$$

$u = 1, \dots, M$. carries an error of $O(h^2)$.

This is an ODE. Euler's method yields

$$u_m^{n+1} = u_m^n + \mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n)$$

Backward Euler's method gives

$$u_m^{n+1} - \mu(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}) = u_m^n$$

This method is often known as the method of lines

Method 2.4 (Crank-Nicolson scheme) Discretising (*) and using trapezoidal rule, we get

$$U_m^{n+1} - \frac{1}{2}\mu\left(U_{m-1}^{n+1} - 2U_m^{n+1} + U_{m+1}^{n+1}\right) = U_m^n + \frac{1}{2}\mu\left(U_{m-1}^n - 2U_m^n + U_{m+1}^n\right)$$

$m = 1, \dots, M$, with error $O(k^3 + kh^2)$.

Each step requires solving a system, but the matrix is TST, so we can use Cholesky factorisation, with $O(M)$ operations.

Def^n 2.5 A is **normal** if $A = QDQ^\dagger = QDQ^*$, D diag, Q unitary.

In other words, A normal if it has a complete set of orthonormal evecs.

Note: Sym matrices and skew-sym matrices are normal.

Prop 2.6 If A normal, then $\|A\| = \rho(A)$.

Pf: Let $\underline{u} \in \mathbb{C}^n$, we expand it in the orthonormal basis of evec $\underline{u} = \sum_{i=1}^n a_i \underline{q}_i$. Then $A\underline{u} = \sum_{i=1}^n \lambda_i a_i \underline{q}_i$.

$$\|A\|_2 := \sup_{\underline{u}} \frac{\|A\underline{u}\|_2}{\|\underline{u}\|_2} = \sup_{a_i} \frac{\left(\sum_{i=1}^n |\lambda_i a_i|^2\right)^{1/2}}{\left(\sum_{i=1}^n |a_i|^2\right)^{1/2}} = |\lambda_{\max}|$$

□

Remark For any A, $\|A\|_2 = |\rho(AA^\dagger)|^{1/2}$, and the above prop. can be deduced from this.

Example (Crank-Nicolson method for diffusion eqn). Let

$$U_m^{n+1} - \frac{1}{2}\mu\left(U_{m-1}^{n+1} - 2U_m^{n+1} + U_{m+1}^{n+1}\right) = U_m^n + \frac{1}{2}\mu\left(U_{m-1}^n - 2U_m^n + U_{m+1}^n\right)$$

for $m = 1, \dots, M$. Then $B\underline{U}^{n+1} = C\underline{U}^n$, where

$$B = I - \frac{1}{2}\mu A_*, \quad C = I + \frac{1}{2}\mu A_*, \quad A_* = \begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}_{M\times M}$$

Then $\underline{U}^{n+1} = B^{-1}C\,\underline{U}^n$. All TST matrices have the same evals, so does $B^{-1}C$. The evecs are orthogonal, and $A = B^{-1}C$ is normal and the evecs are

$$\lambda_k(A) = \frac{\lambda_k(C)}{\lambda_k(B)} = \frac{1 - 2\mu \sin^2 \frac{\pi k h}{2}}{1 + 2\mu \sin^2 \frac{\pi k h}{2}} \Rightarrow |\lambda_k(A)| \le 1$$

for $k = 1, \dots, M$. So Crank-Nicolson is stable $\forall\, \mu > 0$.

Example (Convergence for Crank-Nicolson for diffusion equation). It can be verified that the local error of Crank-Nicolson is $\eta_m^n = O(k^3 + kh^2)$, where $O(k^3)$ inherited from trapezoidal rule (Compared to $O(k^2)$ for Euler's method). We also have

$$\|\underline{\eta}^n\| = \left\{ h \sum_{m=1}^{M} |\eta_m^n|^2 \right\}^{1/2} = O(k^3 + kh^2)$$

Hence, for the error vector $\underline{e}^n$, we have

$$B\,\underline{e}^{n+1} = C\,\underline{e}^n + \underline{\eta}^n \Rightarrow \|\underline{e}^{n+1}\| \le \|B^{-1}C\|\,\|\underline{e}^n\| + \|B^{-1}\|\,\|\underline{\eta}^n\|.$$

We proved $\|B^{-1}C\| \le 1$, so $\|B^{-1}\| \le 1$, since all evals of B are ≥ 1 (by Gershgorin's thm). so $\|\underline{e}^{n+1}\| \le \|\underline{e}^n\| + \|\underline{\eta}^n\|$, and

$$\|\underline{e}^n\| \le \|\underline{e}^0\| + n\|\underline{\eta}\| = n\|\underline{\eta}\| \le \frac{CT}{k}(k^3 + kh^2) = CT(k^2 + h^2).$$

Taking $k = \alpha h$ will result in $O(h^2)$ error of approx.

Consider the advection equation

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \quad . \; 0 \le x \le 1, \quad t \ge 0.$$

with IC $u(x,0) = u_0(x)$, and Dirichlet B.C $u(0,t) = \phi_0(t)$ and $u(1,t) = \phi_1(t)$.

If we discretise RHS by $\frac{\partial u}{\partial x} = \frac{1}{2h}(u(x+h,t) - u(x-h,t)) + O(h^2)$, we get

$$\frac{du_m}{dt} = \frac{1}{2h}(u_{m+1} - u_{m-1}).$$

Example (Crank-Nicolson for advection equation). Let

$$u_m^{n+1} - u_m^n = \frac{1}{4}\mu\left(u_{m+1}^{n+1} - u_{m-1}^{n+1}\right) + \frac{1}{4}\mu\left(u_{m+1}^n - u_{m-1}^n\right).$$

In this case $\underline{u}^{n+1} = B^{-1}C\,\underline{u}^n$,

$$B = \begin{pmatrix} 1 & -\frac{1}{4}\mu & & \\ \frac{1}{4}\mu & \ddots & \ddots & \\ & \ddots & \ddots & -\frac{1}{4}\mu \\ & & \frac{1}{4}\mu & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & \frac{1}{4}\mu & & \\ -\frac{1}{4}\mu & \ddots & \ddots & \\ & \ddots & \ddots & \frac{1}{4}\mu \\ & & -\frac{1}{4}\mu & 1 \end{pmatrix}$$

For $S = \begin{pmatrix} \alpha & \beta & & \\ -\beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & -\beta & \alpha \end{pmatrix}$, $\lambda_k = \alpha + 2i\beta\cos kx$, $\underline{\omega}_k = (i^m \sin kmx)_{m=1}^M$

$x = \pi h = \frac{\pi}{M+1}$, so S normal, have same evec, so does $A = B^{-1}C$,

so A normal and

$$\lambda_k(A) = \frac{\lambda_k(C)}{\lambda_k(B)} = \frac{1 + \frac{1}{2}i\mu\cos kx}{1 - \frac{1}{2}i\mu\cos kx} \Rightarrow |\lambda_k(A)| = 1, \quad k = 1, \dots, M$$

So Crank-Nicolsion stable $\forall\, \mu > 0$.

Example (Euler for advection equation). Consider

$$u_m^{n+1} - u_m^n = \mu\left(u_{m+1}^n - u_m^n\right), \quad m = 1, \dots, M$$

Then $\underline{u}^{n+1} = A\underline{u}^n$, where

$$A = \begin{pmatrix} 1-\mu & \mu & & \\ & \ddots & \ddots & \\ & & \ddots & \mu \\ & & & 1-\mu \end{pmatrix}$$

but A not normal, although evals are bounded by 1 if $\mu \le 2$,

(evals are $1-\mu$, since upper triangular).

It is easier to work with $\|A\|_{\infty\to\infty}$ given by $|1-\mu| + \mu$

and this is < 1 when $\mu \le 1$.

Technique 2.7 (Fourier analysis of stability) Assume recurrance in the form

$$\sum_{k=r}^{s} a_k u_{m+k}^{n+1} = \sum_{k=r}^{s} b_k u_{m+k}^{n}, \quad m \in \mathbb{Z} \qquad (*)$$

The coeffs a_k, b_k indpt of m, n but typically depend upon μ.

Let $\underline{v} = (v_m)_{m\in\mathbb{Z}} \in \ell_2(\mathbb{Z})$. It's Fourier transform is

$$\hat{v}(\theta) = \sum_{m\in\mathbb{Z}} e^{-im\theta} v_m, \qquad -\pi \le \theta \le \pi$$

Equip sequences and f^ns with norms.

$$\|\underline{v}\| = \left\{ \sum_{m\in\mathbb{Z}} |v_m|^2 \right\}^{1/2}, \qquad \|\hat{v}\|_* = \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{v}(\theta)|^2 \, d\theta \right\}^{1/2}.$$

Lem 2.8 (Parseval's identity) For any $\underline{v} \in \ell_2(\mathbb{Z})$, $\|\underline{v}\| = \|\hat{v}\|_*$.

证:

$$\|\hat{v}\|_*^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Big| \sum_{m\in\mathbb{Z}} e^{im\theta} v_m \Big|^2 d\theta$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{m\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} v_m \bar{v}_k e^{-i(m-k)\theta} \, d\theta$$

$$= \frac{1}{2\pi} \sum_{m\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} v_m \bar{v}_k \int_{-\pi}^{\pi} e^{-i(m-k)\theta} \, d\theta$$

$$= \sum_{m\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} v_m \bar{v}_k \delta_{m-k} = \|\underline{v}\|^2.$$

using $\int_{-\pi}^{\pi} e^{il\theta} \, d\theta = 2\pi \delta(l)$.

This implies FT is an <u>isometry</u> of the Euclidean norm.

Analysis 2.9 (Fourier Analysis of Stability) For $\theta \in [-\pi, \pi]$, let $\hat{u}^n(\theta) = \sum_{m\in\mathbb{Z}} e^{-im\theta} u_m^n$ be FT of $\underline{u}^n \in \ell_2(\mathbb{Z})$. Multiplying (*) by $e^{-im\theta}$ and sum up for $m \in \mathbb{Z}$.

$$\sum_{m=-\infty}^{\infty} e^{-im\theta} \sum_{k=r}^{s} a_k u_{m+k}^{n+1} = \sum_{k=r}^{s} a_k \sum_{m=-\infty}^{\infty} e^{-im\theta} u_{m+k}^{n+1}$$

$$= \sum_{k=r}^{s} a_k \sum_{m=-\infty}^{\infty} e^{-i(m-k)\theta} u_m^{n+1}$$

$$= \left(\sum_{k=r}^{s} a_k e^{ik\theta} \right) \hat{u}^{n+1}(\theta).$$

Similarly, on the RHS, $\hat{u}^{n+1}(\theta) = H(\theta)\, \hat{u}^n(\theta)$. (†) where

$$H(\theta) = \frac{\sum_{k=r}^{s} b_k e^{ik\theta}}{\sum_{k=r}^{s} a_k e^{ik\theta}}$$

H is called the <u>amplification factor</u> of the recurrence.

<u>Thm 2.10</u> The method (*) is stable $\Leftrightarrow |H(\theta)| \le 1 \ \forall \theta \in [-\pi, \pi]$.

<u>Pf</u>: The def^n of stability is equivalent to that $\exists c > 0$ s.t. $\|\underline{u}^n\| \le c \ \forall n \in \mathbb{Z}^+$. Note that FT is isometry, so stability is equivalent to $\|\hat{u}^n\|_* \le c \ \forall n \in \mathbb{Z}^+$. Iterating (†), we have

$$\hat{u}^n(\theta) = [H(\theta)]^n \hat{u}^0(\theta), \quad |\theta| \le \pi, \ n \in \mathbb{Z}^+.$$

Assume $|H(\theta)| \le 1 \ \forall |\theta| \le \pi$. then $|\hat{u}^n(\theta)| \le |\hat{u}^0(\theta)|$

$$\Rightarrow \|\hat{u}^n\|_*^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{u}^n(\theta)|^2 \, d\theta \le \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{u}^0(\theta)|^2 \, d\theta = \|u^0\|_*^2.$$

Hence stability.

Assume $\exists \theta_0 \in [-\pi, \pi]$ s.t. $|H(\theta_0)| = 1 + 2\varepsilon > 1$, say. Since H cts, $\exists -\pi \le \theta_1 < \theta_2 \le \pi$ s.t. $|H(\theta)| \ge 1 + \varepsilon \ \forall \theta \in [\theta_1, \theta_2]$.

Set $\eta = \theta_2 - \theta_1$, and choose IC as the f^n

$$\hat{u}^0(\theta) = \begin{cases} \sqrt{2\pi/\eta} & \theta_1 \le \theta \le \theta_2 \\ 0 & \text{o/w} \end{cases}$$

Then $\|u_n\|^2 = \frac{1}{2\pi} \int_{-\pi}^{-\pi} |H(\theta)|^{2n} |\hat{u}^0(\theta)|^2 \, d\theta$

$$\ge \frac{1}{2\pi} \int_{\theta_1}^{\theta_2} |H(\theta)|^{2n} |\hat{u}^0(\theta)|^2 \, d\theta$$

$$\geq \frac{1}{2\pi}(1+\varepsilon)^{2n}\int_{\theta_1}^{\theta_2}\frac{2\pi}{\eta}\,d\theta$$

$$= (1+\varepsilon)^{2n} \to \infty \quad \text{as } n\to\infty.$$

So the method is unstable. □

Example Consider the Cauchy problem for diffusion eqn.

(i) Euler method : $u_m^{n+1} = u_m^n + \mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n)$

$$\Rightarrow H(\theta) = 1 + \mu(e^{-i\theta} - 2 + e^{i\theta})$$

$$= 1 - 4\mu\sin^2\frac{\theta}{2} \in [1-4\mu, 1]$$

So stable iff $\mu \leq \frac{1}{2}$.

(ii) Backward Euler : $u_m^{n+1} - \mu(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}) = u_m^n$.

$$\Rightarrow H(\theta) = (1 - \mu(e^{-i\theta} - 2 + e^{i\theta}))^{-1} = (1+4\mu\sin^2\tfrac{\theta}{2})^{-1} \in (0,1]$$

So stable for all μ.

(iii) Crank - Nicolson

$$u_m^{n+1} - \tfrac{1}{2}\mu(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}) = u_m^n + \tfrac{1}{2}\mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n)$$

Then

$$H(\theta) = \frac{1+\frac{1}{2}\mu(e^{-i\theta} - 2 + e^{i\theta})}{1-\frac{1}{2}\mu(e^{-i\theta} - 2 + e^{i\theta})} = \frac{1-2\mu\sin^2\frac{\theta}{2}}{1+2\mu\sin^2\frac{\theta}{2}} \in (-1, 1]$$

So stable for all $\mu > 0$.

Advection Equation

Problem 2.11 (Advection equation)

$$u_t = u_x, \qquad x \geq 0.$$

$u = u(x,t)$. $u(x,0) = \varphi(x)$. Exact soln is $u(x,t) = \varphi(x+t)$

Example (Downwind instability) Consider

$$\frac{\partial u_m(t)}{\partial x} \approx \frac{1}{2h}\left(u_m(t) - u_{m-1}(t)\right)$$

So coming to ODE $u_m'(t) = \frac{1}{2h}\left(u_m(t) - u_{m-1}(t)\right)$. For Euler method,

$$u_m^{n+1} = u_m^n + \mu\left(u_m^n - u_{m-1}^n\right), \quad n \in \mathbb{Z}_+$$

The amplification error is

$$H(\theta) = 1 + \mu - \mu e^{-i\theta}$$

For $\theta = \pi/2$, $|H(\theta)|^2 = (1+\mu)^2 + \mu^2 > 1$, so unstable $\forall \mu$.

Example (Upwind method) Semidiscretise $\frac{\partial u_m(t)}{\partial x} \approx \frac{1}{h}\left(u_{m+1}(t) - u_m(t)\right)$, then

$$u_m^{n+1} = u_m^n + \mu\left(u_{m+1}^n - u_m^n\right), \quad n \in \mathbb{Z}_+$$

The local error is $O(k^2 + kh)$, which is $O(h^2)$ for a fixed μ, hence convergence if stable.

$$H(\theta) = 1 - \mu + \mu e^{i\theta}$$

Then $|H(\theta)| = |1 - \mu + \mu e^{i\theta}| \le |1-\mu| + \mu = 1$ for $\mu \in [0, 1]$. So stability for $\mu \le 1$. For $\mu > 1$, $|H(\pi)| = |1 - 2\mu| > 1$, so instability for $\mu > 1$.

Using Euler method for advection eqn.

$$u_m^{n+1} - u_m^n = \mu\left(u_{m+1}^n - u_m^n\right), \quad m = 1, \ldots, M,$$

we have $\underline{u}^{n+1} = A\underline{u}^n$,

$$A = \begin{pmatrix} 1-\mu & \mu & & \\ & \ddots & \ddots & \\ & & & \mu \\ & & & 1-\mu \end{pmatrix}$$

But A not normal.

Method 2.12 (leapfrog method) $\frac{\partial u_m(t)}{\partial x} \approx \frac{1}{2h}(u_{m+1}(t) - u_{m-1}(t))$.

Solve with mid-pt rule

$$y_{n+1} = y_{n-1} + 2k f(t_n, y_n)$$

Outcome is

$$u_m^{n+1} = \mu(u_{m+1}^n - u_{m-1}^n) + u_m^{n-1}$$

local error is $O(k^3 + kh^2) = O(h^3)$.

Assuming we are solving a Cauchy problem,

$$\hat{u}^{n+1}(\theta) = \mu(e^{i\theta} - e^{-i\theta})\hat{u}^n(\theta) + \hat{u}^{n-1}(\theta).$$

whence

$$\hat{u}^{n+1}(\theta) - 2i\mu\sin\theta\ \hat{u}^n(\theta) - \hat{u}^{n-1}(\theta) = 0, \quad n \in \mathbb{Z}_+$$

Our goal is to determine values of μ s.t. $|\hat{u}^n(\theta)|$ is uniformly bounded $\forall\ n, \theta$.

This is a difference eqn $w_{n+1} + bw_n + cw_{n-1} = 0$ with general sol^n $w_n = c_1\lambda_1^n + c_2\lambda_2^n$, λ_1, λ_2 roots to $\lambda^2 + b\lambda + c = 0$.

If $\lambda_1 = \lambda_2$, $w_n = (c_1 + c_2 n)\lambda^n$. Here

$$\lambda_{1,2}(\theta) = i\mu\sin\theta \pm \sqrt{1 - \mu^2\sin^2\theta}$$

Stability equivalent to $|\lambda_{1,2}(\theta)| \le 1\ \forall\theta$, and this is true iff $\mu \le 1$.

Wave Equation

Problem 2.13 (Wave eqn) $u_{tt} = u_{xx}$, $t \ge 0$, given $u(x,0)$ and $u_t(x,0)$. The usual approx look like

$$u_m^{n+1} - 2u_m^n + u_m^{n-1} = \mu(u_{m+1}^n - 2u_m^n + u_{m-1}^n)$$

with $\mu = k^2/h^2$.

Fourier analysis of Cauchy problem (infinite domain) gives

$$\hat{u}^{n+1}(\theta) - 2\hat{u}^{n}(\theta) + \hat{u}^{n-1}(\theta) = -4\mu \sin^2\frac{\theta}{2}\,\hat{u}^{n}(\theta).$$

Then char eqn is $\lambda^2 - 2(1-2\mu\sin^2\frac{\theta}{2})\lambda + 1 = 0$. Product of root is 1, so stability ($|\lambda_{1,2}| \le 1$) equivalent to roots being complex conjugate, so need

$$|1-2\mu\sin^2\frac{\theta}{2}| \le 1$$

this is true iff $\mu = k^2/h^2 \le 1$.

Diffusion eqn in 2 space dimensions

Problem 2.14 (2D diffusion eqn) Solving

$$\frac{\partial u}{\partial t} = \nabla^2 u, \qquad 0 \le x, y \le 1, \quad t \ge 0.$$

where $u = u(x,y,t)$. with IC at $t=0$ and Dirichlet B.C. at $\partial\Omega$.

let $u_{l,m}(t) \approx u(lh, mh, t)$, $h = \Delta x = \Delta y$ and $u^n_{l,m} = u_{l,m}(nk)$. the five-point formula gives

$$u'_{l,m} = \frac{1}{h^2}\left(u_{l-1,m} + u_{l+1,m} + u_{l,m-1} + u_{l,m+1} - 4u_{l,m}\right)$$

or

$$\underline{u}' = \frac{1}{h^2} A_* \underline{u}, \quad \underline{u} = (u_{l,m}) \in \mathbb{R}^N.$$

where

$$A_* = \begin{pmatrix} H & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & H \end{pmatrix}, \qquad H = \begin{pmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix}$$

Euler method yields

$$u_{l,m}^{n+1} = u_{l,m}^n + \mu\left(u_{l-1,m}^n + u_{l+1,m}^n + u_{l,m-1}^n + u_{l,m+1}^n - 4u_{l,m}^n\right)$$

or

$$\underline{u}^{n+1} = A\underline{u}^n, \qquad A = I + \mu A_*.$$

where $\mu = k/h^2$. The local error is $\eta = O(k^2 + kh) = O(h^4)$.

Note: We choose spectural analysis for bounded domain.

Fourier analysis for unbounded domain.

Note that A sym, so normal, so the evals are related to those of A_* by

$$\lambda_{k,l}(A) = 1 + \mu\,\lambda_{k,l}(A_*) = 1 - 4\mu\left(\sin^2\frac{\pi kh}{2} + \sin^2\frac{\pi lh}{2}\right).$$

Then

$$\sup_{h>0} \rho(A) = \max\{1, |1-8\mu|\}$$

So $\mu \le 1/4 \iff$ stability

Method 2.15 (Fourier Analysis) We extend the range of (x,y) from $[0,1]\times[0,1]$ to $\mathbb{R}^2$. Then the 2D FT is

$$\hat{u}(\theta,\psi) = \sum_{l,m\in\mathbb{Z}} u_{l,m}\, e^{-i(l\theta + m\psi)}$$

and all our results can generalise. In particular, FT is an isometry from $l_2[\mathbb{Z}^2]$ to $L_2([-\pi,\pi]^2)$, ie.

$$\left(\sum_{l,m\in\mathbb{Z}} |u_{l,m}|^2\right)^{1/2} =: \|\underline{u}\| = \|\hat{u}\|_* = \left(\frac{1}{4\pi^2}\int_{-\pi}^{\pi} d\theta \int_{-\pi}^{\pi} d\psi\ |\hat{u}(\theta,\psi)|^2\right)^{1/2}.$$

The method is stable iff $|H(\theta,\psi)| \le 1 \quad \forall\, \theta,\psi \in [-\pi,\pi]$

$$H(\theta,\psi) = 1 + \mu\left(e^{-i\theta} + e^{i\theta} + e^{-i\psi} + e^{i\psi} - 4\right) = 1 - 4\mu\left(\sin^2\frac{\theta}{2} + \sin^2\frac{\psi}{2}\right).$$

So again, stability $\Leftrightarrow \mu \leq 1/4$.

Method 2.16 (Crank-Nicolson for 2D). Applying trap. rule to semi-discretisation, we obtain

$$(I - \tfrac{1}{2}\mu A_*)\, \underline{y}^{n+1} = (I + \tfrac{1}{2}\mu A_+)\, \underline{y}^n$$

So we move by solving $\underline{y}^{n+1} = B^{-1}C\,\underline{y}^n$. Eval analysis shows that $A = B^{-1}C$ normal and shares same evecs with B and C, hence

$$\lambda(A) = \frac{\lambda(C)}{\lambda(B)} = \frac{1 + \frac{1}{2}\mu\lambda(A_*)}{1 - \frac{1}{2}\mu\lambda(A_*)} \Rightarrow |\lambda(A)| < 1 \text{ as } \lambda(A_*) < 0.$$

So stable for all μ.

Technique 2.17 (Splitting) In semi-discretisation, we reach ODE of the form

$$\underline{y}' = A\underline{y}, \quad \underline{u}(0) = \underline{u}_0$$

The soln is

$$\underline{u}(t) = e^{tA}\,\underline{u}_0$$

where $e^B := \sum_{k=0}^{\infty} \frac{1}{k!} B^k$. Easy to verify that $\frac{d}{dt} e^{tA} = A e^{tA}$.

If A can be diag, $A = VDV^{-1}$, then $e^{tA} = V e^{tD} V^{-1}$, where $e^{tD} = \operatorname{diag}(e^{tD_{ii}})$. But computing evals is costly.

Indeed, one step method approximates a matrix exponential.

- Euler: $\underline{u}^{n+1} = (I + kA)\,\underline{y}^n$, $e^z = 1 + z + \mathcal{O}(z^2)$.
- Implicit Euler: $\underline{u}^{n+1} = (I - kA)^{-1}\underline{y}^n$, $e^z = (1-z)^{-1} + \mathcal{O}(z^2)$
- Trapezoidal: $\underline{y}^{n+1} = (I - \frac{1}{2}kA)^{-1}(I + \frac{1}{2}kA)\,\underline{y}^n$, $e^z = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} + \mathcal{O}(z^3)$.

If A is sparse, this can be exploited when solving linear system.

In many cases, A is a sum of two matrices. say $A = B + C$.

Example 2D diffusion eqn with zero b.c., then

$$A = \frac{1}{h^2}(A_x + A_y).$$

where $\frac{1}{h^2} A_x \in \mathbb{R}^{M^2 \times M^2}$ corresponds to the 3-point discretisation of ∂_x^2.

$$A_x = \begin{pmatrix} -2I & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & -2I \end{pmatrix}. \quad A_y = \begin{pmatrix} G & & \\ & \ddots & \\ & & G \end{pmatrix}.$$

with $G = \begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{M \times M}$

Remark: We can write $A = G \otimes I$. with

$$A \otimes B = \begin{pmatrix} A_{11}B & \cdots & A_{1m_A}B \\ \vdots & \ddots & \vdots \\ A_{n_A 1}B & \cdots & A_{n_A m_A}B \end{pmatrix} \in \mathbb{R}^{n_A n_B \times m_A m_B}$$

with $A \in \mathbb{R}^{n_A \times m_A}$, $B \in \mathbb{R}^{n_B \times m_B}$.

In general, $\exp(t(B+C)) \neq \exp(tB)\exp(tC)$. Equality holds iff $[B, C] = 0$.

Prop 2.18 For any matrices B, C,

$$e^{t(B+C)} = e^{tB} e^{tC} + \frac{1}{2} t^2 (CB - BC) + \mathcal{O}(t^3).$$

If B, C commute, then $e^{B+C} = e^B e^C$.

Pf: Taylor expand.

$$e^{tB} e^{tC} = \left(I + tB + \tfrac{1}{2}t^2 B^2 + \mathcal{O}(t^3)\right)\left(I + tC + \tfrac{1}{2}t^2 C^2 + \mathcal{O}(t^3)\right)$$

$$= I + t(B+C) + \tfrac{1}{2}t^2 (B^2 + C^2 + 2BC) + \mathcal{O}(t^3).$$

$$e^{t(B+C)} = I + t(B+C) + \frac{1}{2}t^2(B^2+C^2+BC+CB) + O(t^3)$$

If B, C commute.

$$\exp(B+C) = \sum_{n=0}^{\infty} \frac{1}{n!}(B+C)^n$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} B^{n-k} C^k$$

$$= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{1}{n!} \binom{n}{k} B^{n-k} C^k$$

$$= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{1}{k!(n-k)!} B^{n-k} C^k$$

$$= \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{k!\,m!} B^m C^k = e^B e^C.$$

Technique 2.19 (Splitting for 2D diffusion eqn) Recall $u_t = \partial_x^2 u + \partial_y^2 u$ using the five-point scheme yields

$$\frac{d\underline{u}}{dt} = \frac{1}{h^2}(A_x + A_y)\underline{u}.$$

where $A_x = G \otimes I$, $A_y = I \otimes G$, $G = \begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}$.

It can be checked that $A_x A_y = A_y A_x = G \otimes G$, so

$$e^{k(A_x+A_y)/h^2} = e^{kA_x/h^2} e^{kA_y/h^2}$$

So the semi-discretised diffusion eqn in 2D with zero b.c. satisfies

$$\underline{u}^{n+1} = e^{kA_x/h^2} e^{kA_y/h^2} \underline{u}^n$$

The Split Crank-Nicolson Scheme

We approximate each exponential map by rational f^n

$$r(z) = (1 + z/2)(1 - z/2)^{-1}$$

which leads to

$$\underline{u}^{n+1} = \left(I + \frac{\mu}{2} A_x\right)\left(I - \frac{\mu}{2} A_x\right)^{-1}\left(I + \frac{\mu}{2} A_y\right)\left(I - \frac{\mu}{2} A_y\right)^{-1} \underline{u}$$

Note that computing

$$\underline{u}^{n+1/2} = \left(I + \frac{\mu}{2} A_y\right)\left(I - \frac{\mu}{2} A_y\right)^{-1} \underline{u}$$

can be done in $O(M^2)$ as A_y block diagonal.

Solving the remaining part is also $O(M^2)$ since A_x also block diag provided we permute the rows and columns so that the grid order by rows but not columns. So the scheme can be done in $O(M^2)$ and only requires tridiagonal matrix (no FFT needed).

Stability: write

$$\| r(\mu A_x)\, r(\mu A_y) \|_2 \leq \| r(\mu A_x) \|_2 \, \| r(\mu A_y) \|_2 \leq 1$$

since

$$\| r(\mu A_x) \|_2 = \| \left(I + \frac{\mu}{2} A_x\right)\left(I - \frac{\mu}{2} A_x\right)^{-1} \|_2 \leq 1$$

Since A_x sym and evals ≤ 0.

Exercise Check the consistency of the scheme

$$\underline{u}^{n+1} = r(\mu A_x)\, r(\mu A_y)\, \underline{u}^n$$

In particular, show that split Crank-Nicolson has the 'same' local error as Crank-Nicolson scheme, ie. local error is $O(k^3 + kh^2)$.

Example Consider

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a(x,y)\frac{\partial u}{\partial y}\right) + f(x,y).$$

where $a(x,y) > \alpha > 0$, $f(x,y)$ given, with I.C. on $[0,1]^2$ and Dirichlet b.c. Replace each space derivative by central differences

$$\frac{dg(\xi)}{d\xi} \approx \frac{g(\xi + \frac{1}{2}h) - g(\xi - \frac{1}{2}h)}{h}$$

resulting in

$$u'_{l,m} = \frac{1}{h^2}\Big(a_{l-1/2,m}\, u_{l-1,m} + a_{l+1/2,m}\, u_{l+1,m} + a_{l,m-1/2}\, u_{l,m-1}$$

$$+ a_{l,m+1/2}\, u_{l,m+1} - \left(a_{l-1/2,m} + a_{l+1/2,m} + a_{l,m-1/2} + a_{l,m+1/2}\right) u_{l,m}\Big) + f_{l,m}$$

Assuming zero b.c., we have $y' = Ay$, where A can be split as

$$A = \frac{1}{h^2}(A_x + A_y).$$

A_x, A_y constructed from the contributions of discretisations in x- and y-directions resp., i.e. A_x includes all the $a_{l\pm 1/2,m}$ terms, and A_y consists of $a_{l,m\pm 1/2}$ terms. A_x, A_y not necessarily commute, so

$$y^{n+1} = e^{kA_x/h^2}\, e^{kA_y/h^2}\, y^n$$

carry an error of $O(k^2)$.

Strang Splitting

One can obtain better splitting approximations of $e^{t(B+C)}$. It can be shown $e^{\frac{1}{2}tB}\, e^{tC}\, e^{\frac{1}{2}tB}$ gives a $O(t^3)$ approx. of $e^{t(B+C)}$, i.e.

$$e^{t(B+C)} = e^{\frac{1}{2}tB}\, e^{tC}\, e^{\frac{1}{2}tB} + O(t^3)$$

Technique 2.20 (Splitting of inhomogeneous systems) In general, the ODE is of the form

$$\underline{u}' = A\underline{u} + \underline{b}, \quad \underline{u}(0) = \underline{u}^0$$

where $\underline{b}$ originates in b.c. and possibly fix'g) forcing term. We should have $\underline{b} = \underline{b}(t)$, since b.c. may vary with time. The exact sol^n is provided by the variation of constants formula

$$\underline{u}(t) = e^{tA}\underline{u}(0) + \int_0^t e^{(t-s)A}\underline{b}(s)\,ds.$$

Therefore,

$$\underline{u}(t_{n+1}) = e^{kA}\underline{u}(t_n) + \underbrace{\int_{t_n}^{t_{n+1}} e^{(t_{n+1}-s)A}\underline{b}(s)\,ds}_{(*)}$$

(*) can be evaluated using quadrature, e.g. trapezoidal rule gives

$$\underline{u}(t_{n+1}) \approx e^{kA}\underline{u}(t_n) + \tfrac{1}{2}k\left(e^{kA}\underline{b}(t_n) + \underline{b}(t_{n+1})\right).$$

with a local error of $O(k^3)$. We can replace exponentials with their splittings. E.g. with Strang's splitting with $r(z) = (1+z/2)/(1-z/2)$.

$$\underline{u}^{n+1} = r(\tfrac{1}{2}kB)\,r(kC)\,r(\tfrac{1}{2}kB)\left[\underline{u}^n + \tfrac{1}{2}k\underline{b}^n\right] + \tfrac{1}{2}k\underline{b}^{n+1}.$$

Then everything reduces to inexpensive sol^n of tridiagonal systems.

3. Spectral Methods

General idea: Consider PDE in $\mathcal{L}u = f$. We can choose finite subspace of f's V spanned by $\psi_1, \dots, \psi_N$. We seek approximation by a L.C. of ψ_n, i.e. $u_N(x) = \sum_{n=1}^{N} c_n \psi_n(x)$. Then

$$\sum_{n=1}^{N} c_n \mathcal{L}\psi_n = f.$$

Assume (ψ_n) orthogonal, we require the projection of $\mathcal{L}u_N - f$ on the subspace V is zero, i.e.

$$\sum_{n=1}^{N} c_n \langle \mathcal{L}\psi_n, \psi_m \rangle = \langle f, \psi_m \rangle$$

We call $A_{mn} = \langle \mathcal{L}\psi_n, \psi_m \rangle$, then we end up with

$$Ac = \tilde{f},$$

where $\tilde{f}_m = \langle f, \psi_m \rangle$.

Problem 3.1 (Fourier approximations of f^ns). Consider the truncated Fourier approximations of a f^n f on $[-1,1]$.

$$f(x) \approx \phi_N(x) = \sum_{n=-\frac{N}{2}+1}^{N/2} \hat{f}_n e^{i\pi n x}, \quad x \in [-1,1]$$

where $N \geq 2$ and N even, and

$$\hat{f}_n = \frac{1}{2} \int_{-1}^{1} f(t) e^{-i\pi n t} dt, \quad n \in \mathbb{Z}.$$

Thm 3.2 (The de la Vallée Poussin thm) If f Riemann integrable, $\hat{f}_n = O(n^{-1})$ for $|n| \gg 1$, then $\phi_N(x) = f(x) + O(N^{-1})$ as $N \to \infty$ for every $x \in (-1,1)$, where f Lipschitz.

Remark (The Gibbs effect at end points) Note if f smoothly diff, then IBP gives

$$\hat{f}_n = \frac{(-1)^{n+1}}{2\pi i n}\,[f(1) - f(-1)] + \frac{1}{\pi i n}\,\hat{f}_n' = O(n^{-1})$$

for $|n| \gg 1$. Since f Lipschitz on $(-1,1)$. ϕ_N cgs to f with speed $O(N^{-1})$. But this is slow and we cannot guarentee cgs at ± 1. In fact, it is possible to show that

$$\phi_N(\pm 1) \longrightarrow \frac{1}{2}\,[f(-1) + f(1)] \quad \text{as } n \to \infty$$

Hence, unless f periodic, we fail to converge.

Method 3.3 Suppose f analytic f^n in $[-1,1]$, that can be extended analytically to a closed complex domain Ω. Let f be periodic with period 2. In particular, $f^{(m)}(-1) = f^{(m)}(1)$ $\forall\, m \in \mathbb{Z}_+$, then, by multiple IBP, we get

$$\hat{f}_n = \frac{1}{\pi i n}\,\hat{f}_n' = \frac{1}{(\pi i n)^2}\,\hat{f}_n'' = \cdots$$

Thus, we have

$$\hat{f}_n = \frac{1}{(\pi i n)^m}\,\hat{f}_n^{(m)},$$

for $m = 0, 1, \ldots$

But, how large is $|\hat{f}_n^{(m)}|$? Using Cauchy's thm

$$f^{(m)}(x) = \frac{m!}{2\pi i}\int_\gamma \frac{f(z)\,dz}{(z-x)^{m+1}}, \qquad x \in [-1,1]$$

where γ is +ve oriented boundary of Ω. So with $\alpha^{-1} > 0$ being the minimal distance between γ and $[-1,+1]$ and $M = \max\{|f(z)| : z \in \gamma\} < \infty$, it follows that

$$|f^{(m)}(x)| \le \frac{m!}{2\pi}\int_\gamma \frac{|f(z)|\,|dz|}{|z-x|^{m+1}} \le \frac{M\ \text{length}(\gamma)}{2\pi}\, m!\,\alpha^{m+1}$$

Hence, we can bound $|\widehat{f_n^{(m)}}| \leq c\, m!\, \alpha^{m+1}$ for some $c > 0$.

Thus,

$$|\phi_N(x) - f(x)| = \left| \sum_{n=-\frac{N}{2}+1}^{N/2} \hat{f}_n e^{i\pi n x} - \sum_{n=-\infty}^{\infty} \hat{f}_n e^{i\pi n x} \right|$$

$$\leq \sum_{|n| \geq N/2} |\hat{f}_n|$$

$$= \sum_{|n| \geq N/2} \frac{|\hat{f}_n^{(m)}|}{|\pi n|^m}$$

$$\leq \frac{c\, m!\, \alpha^{m+1}}{\pi^m} \sum_{n=N/2}^{+\infty} \frac{1}{n^m}.$$

Using $\forall\, r \in \mathbb{N},\ m > 1$.

$$\sum_{n=r+1}^{\infty} \frac{1}{n^m} \leq \int_r^{\infty} \frac{dt}{t^m} = \frac{1}{m-1} r^{-m+1}.$$

So

$$|\phi_N(x) - f(x)| \leq c'\, m! \left(\frac{\alpha}{\pi N}\right)^{m-1}, \quad m \geq 2.$$

Finally, by Stirling's formula

$$m! \sim \sqrt{2\pi}\, m^{m+1/2} e^{-m},$$

we have

$$m! \left(\frac{\alpha}{\pi N}\right)^{m-1} \approx \sqrt{2\pi m}\, \frac{m}{e} \left(\frac{\alpha m}{\pi e N}\right)^{m-1}$$

which becomes small for large N, so $|\phi_N - f| = O(N^{-p})$ for any $p \in \mathbb{N}$, and so Fourier approximation on an analytic f^n is of infinite order.

Defn 3.4 (Convergence at spectral speed) An N-term approximation ϕ_N of a f^n f cgs to f at spectral speed if $\|\phi_N - f\|$ decays faster than $O(N^{-p})$ for any $p = 1, 2, \ldots$

Remark: It is possible to prove that $\exists C_1, \omega > 0$ s.t. $\|\phi_N - f\| \le C_1 e^{-\omega N}$ $\forall N \in \mathbb{N}$ uniformly in $[-1,1]$. Thus, cgs is at least at exponential rate.

Algebra of Fourier expansions: let $\mathcal{A}$ be the set of all f^n s $f: [-1,1] \to \mathbb{C}$ analytic in $[-1,1]$. 2-periodic can be extended analytically into complex plane. Then $\mathcal{A}$ linear space.

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}_n e^{i\pi n x}, \quad g(x) = \sum_{n=-\infty}^{\infty} \hat{g}_n e^{i\pi n x}$$

We have

$$f(x) + g(x) = \sum_{n=-\infty}^{\infty} (\hat{f}_n + \hat{g}_n) e^{i\pi n x}, \quad \alpha f(x) = \sum_{n=-\infty}^{\infty} \alpha \hat{f}_n e^{i\pi n x}.$$

and

$$f(x)\, g(x) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_{n-m} \hat{g}_m e^{i\pi n x} = \sum_{n=-\infty}^{\infty} (\hat{f} * \hat{g}) e^{i\pi n x}.$$

where $\widehat{(f \cdot g)}_n = (\hat{f} * \hat{g})_n$. Moreover, if $f \in \mathcal{A}$ then $f' \in \mathcal{A}$, and

$$f'(x) = i\pi \sum_{n=-\infty}^{\infty} n \cdot \hat{f} e^{i\pi n x}.$$

Since $|\hat{f}_n|$ decays faster than $O(n^{-p})$ for any $p \in \mathbb{N}$, this provides that all derivatives of f have rapidly convergent Fourier coeff.

Example Consider $y = y(x)$, $x \in [-1,1]$.

$$y'' + a(x) y' + b(x) y = f(x), \quad y(-1) = y(1).$$

where $a, b, f \in \mathcal{A}$ and we seek a periodic solution $y \in \mathcal{A}$.

Substituting a, b, f, y by their FS.

$$-\pi^2 n^2 \hat{y}_n + i\pi \sum_{m=-\infty}^{\infty} m \hat{a}_{n-m} \hat{y}_m + \sum_{m=-\infty}^{\infty} \hat{b}_{n-m} \hat{y}_m = \hat{f}_n, \quad n \in \mathbb{Z}.$$

Since $a, b, f \in \mathcal{A}$, the Fourier coefficients decrease rapidly, like $O(n^{-p})$ for every $p \in \mathbb{N}$. Hence, we can truncate into N-dimensional system.

$$-\pi^2 n^2 \hat{y}_n + i\pi \sum_{m=-N/2+1}^{N/2} m \hat{a}_{n-m} \hat{y}_m + \sum_{m=-N/2+1}^{N/2} \hat{b}_{n-m} \hat{y}_m = \hat{f}_n \qquad (*)$$

for $n = -\frac{N}{2}+1, \dots, \frac{N}{2}$.

Remark. The matrix of (*) is in general dense, but our theory predicts fairly small values of N, hence very small matrices, are sufficient for high accuracy. E.g. choosing $a(x) = f(x) = \cos \pi x$, $b(x) = \sin(2\pi x)$, we get

$N = 16$, error 10^{-10}

$N = 22$ error 10^{-15}

Exercise The transformation matrix of DFT is

$$W = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \omega & \cdots & \omega^{N-1} \\ \vdots & \omega^2 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & & \\ 1 & \omega^{N-1} & \cdots & \omega^{(N-1)(N-1)} \end{pmatrix}$$

where $\omega_N = e^{-2i\pi/N}$. Show that $W W^* = W^* W = I$.

Method 3.5 (computations of Fourier coeff (DFT)). We want to compute

$$\hat{f}_n = \frac{1}{2} \int_{-1}^{1} f(t) e^{-i\pi nt} \, dt, \quad n \in \mathbb{Z}.$$

Suppose we wish to compute the integral on $[-1, 1]$ of a f^n $f \in \mathcal{A}$ by means of Riemann sums on the uniform partition

$$\int_{-1}^{1} h(t) \, dt \approx \frac{2}{N} \sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right)$$

This is known as a rectangle rule. We want to know how good this approximation is. Letting $\omega = e^{2\pi i/N}$, we have

$$\frac{2}{N}\sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right) = \frac{2}{N}\sum_{k=-N/2+1}^{N/2}\sum_{n=-\infty}^{\infty} \hat{h}_n e^{2\pi i nk/N}$$

$$= \frac{2}{N}\sum_{n=-\infty}^{\infty} \hat{h}_n \sum_{k=-N/2+1}^{N/2} \omega_N^{nk}.$$

Since $\omega_N^N = 1$,

$$\sum_{k=-N/2+1}^{N/2} \omega_N^{nk} = \omega_N^{-n(N/2-1)} \sum_{k=0}^{N-1} \omega_N^{nk} = \begin{cases} N, & n \equiv 0 \ (N) \\ 0, & n \not\equiv 0 \ (N). \end{cases}$$

and we deduce that

$$\frac{2}{N}\sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right) = 2\sum_{r=-\infty}^{\infty} \hat{h}_{Nr}$$

Hence, the error committed by Riemann approximation is

$$e_N(h) := \frac{2}{N}\sum_{k=-N/2+1}^{N/2} h\left(\frac{2k}{N}\right) - \int_{-1}^{1} h(t)\,dt$$

$$= 2\sum_{r=-\infty}^{\infty} \hat{h}_{Nr} - 2\hat{h}_0$$

$$= 2\sum_{r=1}^{\infty} \left(\hat{h}_{Nr} + \hat{h}_{-Nr}\right)$$

Since $h \in A$, its Fourier coeff. decays at spectral rate, ie. $\hat{h}_{Nr} = O\left((Nr)^{-p}\right)$ for any $p \in \mathbb{N}$. hence,

$$e_N(h) = O(N^{-p}) \quad \forall p \in \mathbb{N}.$$

We may compute integral of $h(x) = \frac{1}{2} f(x) e^{-i\pi nx}$ by Riemann sums.

$$\hat{f}_n \simeq \frac{1}{N}\sum_{k=-N/2+1}^{N/2} f\left(\frac{2k}{N}\right) \omega_N^{-nk}, \qquad n = -\frac{N}{2}+1, \ldots, \frac{N}{2}. \tag{†}$$

Remark The formula (†) is the DFT of $(y_k) = \left(f(\frac{2k}{N})\right)$, hence we have a spectral rate of convergence, and a fast algorithm (FFT) of computing Fourier coeff.

Problem 3.6 (Poisson eqn) consider

$$\nabla^2 u = f, \qquad -1 \le x, y \le 1.$$

f analytic and obey periodic b.c.

$$f(-1, y) = f(1, y) \qquad -1 \le y \le 1,$$
$$f(x, -1) = f(x, 1) \qquad -1 \le x \le 1$$

And add the eqn to b.c.

$$u(-1, y) = u(1, y), \quad u_x(1, y) = u_x(-1, y) \qquad -1 \le y \le 1$$
$$u(x, -1) = u(x, 1), \quad u_y(x, -1) = u_y(x, 1) \qquad -1 \le x \le 1$$

With these B.C.s, sol^n is only defined up to an additive const. So we add a normalisation condition to fix the const.

$$\int_{-1}^{1} dx \int_{-1}^{1} dy \quad u(x, y) = 0.$$

We have spectrally convergent Fourier expansions

$$f(x, y) = \sum_{k,l=-\infty}^{\infty} \hat{f}_{k,l} \, e^{-i\pi(kx+ly)}$$

and seek for

$$u(x, y) = \sum_{k,l=-\infty}^{\infty} \hat{u}_{k,l} \, e^{i\pi(kx+ly)}$$

Since

$$0 = \int_{-1}^{1} dx \int_{-1}^{1} dy \; u(x, y)$$
$$= \sum_{k,l=-\infty}^{\infty} \hat{u}_{k,l} \int_{-1}^{1} dx \int_{-1}^{1} dy \; e^{i\pi(kx+ly)} = \hat{u}_{0,0}$$

and

$$\nabla^2 u(x,y) = -\pi^2 \sum_{k,l=-\infty}^{\infty} (k^2+l^2)\, \hat{u}_{k,l}\, e^{i\pi(kx+ly)}$$

Thus,

$$\begin{cases} \hat{u}_{k,l} = -\dfrac{1}{(k^2+l^2)\pi^2} \hat{f}_{k,l} & , \ k,l \in \mathbb{Z},\ k,l \neq 0. \\ \hat{u}_{0,0} = 0 \end{cases}$$

Remark Applying a spectral method to Poisson eqn is not representative for other PDEs. In fact, $\phi_{k,l} = e^{i\pi(kx+ly)}$ are the ef^n's of ∇^2 with eval $-\pi^2(k^2+l^2)$. and they obey periodic B.C.s.

Problem 3.7 (General 2nd linear elliptic PDE). Consider

$$\nabla^T(a\nabla u) = f, \qquad -1 \le x,y \le 1,$$

with $a(x,y) > 0$, a.f periodic. Imposing periodic b.c. and normalisation condition, write

$$\nabla^T(a\nabla u) = \frac{\partial}{\partial x}(au_x) + \frac{\partial}{\partial y}(au_y) = f$$

and use Fourier expansions

$$g(x,y) = \sum_{k,l\in\mathbb{Z}} \hat{g}_{k,l}\, \phi_{k,l}(x,y) \qquad h(x,y) = \sum_{m,n\in\mathbb{Z}} \hat{h}_{m,n}\, \phi_{m,n}(x,y)$$

together with

$$\widehat{(g\cdot h)}_{k,l} = \sum_{m,n\in\mathbb{Z}} \hat{g}_{k-m,l-n}\, \hat{h}_{m,n}$$

$$(\hat{g_x})_{k,l} = i\pi k\, \hat{g}_{k,l} \qquad (\hat{g_y})_{k,l} = i\pi l\, \hat{g}_{k,l}$$

$$(\hat{h_x})_{m,n} = i\pi m\, \hat{h}_{m,n} \qquad (\hat{h_y})_{m,n} = i\pi n\, \hat{h}_{m,n}$$

This gives

$$-\pi^2 \sum_{k,l\in\mathbb{Z}} \sum_{m,n\in\mathbb{Z}} (km+ln)\, \hat{a}_{k-m,l-n}\, \hat{u}_{m,n}\, \phi_{k,l}(x,y) = \sum_{k,l\in\mathbb{Z}} \hat{f}_{k,l}\, \phi_{k,l}(x,y)$$

Truncate the expansions to $-N/2+1 \le k,l,m,n \le N/2$ and impose $\hat{u}_{0,0} = 0$, resulting in a N^2-1 linear algebraic equations in unknowns $\hat{u}_{m,n}$, $m,n = -\frac{N}{2}+1, \dots, \frac{N}{2}$, $m,n \neq 0$.

$$\sum_{m,n=-N/2+1}^{N/2} (km+ln)\, \hat{a}_{k-m,l-n}\, \hat{u}_{m,n} = -\frac{1}{\pi^2}\, \hat{f}_{k,l}.$$

for $k,l = -\frac{N}{2}+1, \dots, \frac{N}{2}$

Remark The fast convergence of spectral method rests on analyticity and periodicity. However, we can relax these two assumptions will retain advantages of FS.

- Relaxing analyticity: The smoother the f^n, the faster the truncated series converge. ie. for $f \in C^p[-1,1]$ we have $O(N^{-p})$ order of convergence. Spectral convergence can be recovered if $f \in C^\infty(-1,1)$, i.e. $f^{(m)}(x)$ exists $\forall x \in (-1,1)$, for $m = 0,1,2,\dots$.
- Relaxing periodicity: periodicity is needed for spectral convergence, since FS converges as $O(N^{-1})$ unless $f(-1) = f(1)$. However, we can set our basis f^ns, e.g. to Chebyshev. polynomials

Def^n The Chebyshev poly. of degree n is defined as

$$T_n(x) := \cos(n \cos^{-1} x).$$

or $$T_n(x) := \cos(n\theta), \quad x = \cos\theta, \quad \theta \in [0,\pi]$$

Remark · (T_n) obeys $T_0(x) = 1$, $T_1(x) = x$,

$$T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)$$

in particular T_n is a poly. of deg n, with leading coeff 2^{n-1}.

· (T_n) form a sequence of orthogonal poly w.r.t.

$$(f,g)_w := \int_{-1}^{1} f(x)g(x)\,w(x)\,dx,$$

with $w(x) := (1-x^2)^{-1/2}$. Namely, we have

$$(T_n, T_m)_w = \int_{-1}^{1} T_m(x)\,T_n(x)\,\frac{1}{\sqrt{1-x^2}}\,dx$$

$$= \int_0^{\pi} \cos m\theta \cos n\theta \, d\theta$$

$$= \begin{cases} \pi & m=n=0 \\ \pi/2 & m=n\geq 1 \\ 0 & m\neq n \end{cases}$$

Method 3.8 (Chebyshev expansion) Since $(T_n)_{n=0}^{\infty}$ form orthogonal space, a fn f s.t. $\int_{-1}^{1} |f(x)|^2 w(x)\,dx < \infty$ can be expanded in

$$f(x) = \sum_{n=0}^{\infty} \check{f}_n\,T_n(x).$$

with Chebyshev coeff $\check{f}_n$. Making inner product of both sides with T_n and using orthogonality yields

$$(f, T_n)_w = \check{f}_n\,(T_n, T_n)_w$$

$$\Rightarrow \check{f}_n = \frac{(f,T_n)_w}{(T_n,T_n)_w} = \frac{c_n}{\pi}\int_{-1}^{1} f(x)\,T_n(x)\,\frac{1}{\sqrt{1-x^2}}\,dx$$

where $c_0 = 1$, $c_n = 2$ for $n \geq 1$.

Connection to Fourier expansions

Letting $x = \cos\theta$, $g(\theta) = f(\cos\theta)$. then

$$\int_{-1}^{1} f(x)\, T_n(x) \cdot \frac{dx}{\sqrt{1-x^2}} = \int_0^{\pi} f(\cos\theta)\, T_n(\cos\theta)\, d\theta = \frac{1}{2}\int_{-\pi}^{\pi} g(\theta)\cos n\theta\, d\theta$$

Given $\cos n\theta = \frac{1}{2}(e^{in\theta} + e^{-in\theta})$, and Fourier expansion of 2π-periodic f^n g,

$$g(\theta) = \sum_{n\in\mathbb{Z}} \hat{g}_n e^{in\theta}, \quad \hat{g}_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(t) e^{-int}\, dt, \; n\in\mathbb{Z}.$$

So

$$\int_{-1}^{1} f(x)\, T_n(x) \frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2}(\hat{g}_{-n} + \hat{g}_n)$$

then

$$\check{f}_n = \begin{cases} \hat{g}_0 & n = 0 \\ \hat{g}_{-n} + \hat{g}_n & n \geq 1 \end{cases}$$

Properties of Chebyshev expansion

- For f integrable, then computing Chebyshev is equivalent to Fourier expansion of $g(\theta) = f(\cos\theta)$. The latter one is 2π-periodic, so can use DFT to compute $\check{f}_n$.
- If f can be analytically extended from $[-1,1]$ (To Bernstein ellipse), then $\check{f}_n$ decays spectrally fast for $n \gg 1$. Hence, Chebyshev expansion inherits rapid convergence of spectral methods assuming f periodic.

Algebra of Chebyshev expansions

let $\mathcal{B}$ be the set of analytic f^ns in $[-1,1]$ that can be analytically extend to the complex plane. $\mathcal{B}$ is a linear space and is closed under multiplication. In particular,

$$T_m(x)\,T_n(x) = \cos(m\theta)\cos(n\theta)$$
$$= \frac{1}{2}\left(\cos((m-n)\theta + \cos((m+n)\theta\right)$$
$$= \frac{1}{2}\left(T_{|m-n|}(x) + T_{m+n}(x)\right)$$

and hence

$$f(x)\,g(x) = \sum_{m=0}^{\infty} \check{f}_m T_m(x) \cdot \sum_{n=0}^{\infty} \check{g}_n T_n(x)$$
$$= \frac{1}{2}\sum_{m,n=0}^{\infty} \check{f}_m\left(\check{g}_{|m-n|} + \check{g}_{m+n}\right) T_n(x).$$

Lem 3.9 We can express derivatives of T_n' in terms of (T_k) as

$$T_{2n}'(x) = (2n)\cdot 2\sum_{k=1}^{n} T_{2k-1}(x)$$

$$T_{2n+1}'(x) = (2n+1)\left[T_0(x) + 2\sum_{k=1}^{n} T_{2k}(x)\right]$$

Pf: $T_m(x) = \cos m\theta \Rightarrow T_m'(x) = \dfrac{m\sin m\theta}{\sin\theta}$, $x = \cos\theta$

For $m=2n$, from the identity $\dfrac{\sin 2n\theta}{\sin\theta} = 2\sum_{k=1}^{n}\cos(2k-1)\theta$, as

$$2\sin\theta\sum_{k=1}^{n}\cos(2k-1)\theta = \sum_{k=1}^{n}\left(\sin(2k\theta) - \sin(2(k-1)\theta)\right) = \sin 2n\theta.$$

For $m=2n+1$, turns into $\dfrac{\sin(2n+1)\theta}{\sin\theta} = 1 + 2\sum_{k=1}^{n}\cos 2k\theta$, as

$$\sin\theta\left(1 + 2\sum_{k=1}^{n}\cos(2k\theta)\right) = \sin\theta + \sum_{k=1}^{n}\left[\sin(2k+1)\theta - \sin(2k-1)\theta\right]$$
$$= \sin(2n+1)\theta \qquad \square$$

Remark (Application to PDEs). All derivatives of u can be expressed in an explicit form of Chebyshev expansion. Computation of Chebyshev coeff of f^n f has to be sampled at

Chebyshev points $\cos(2\pi k/N)$, $k = -\frac{N}{2}+1, \dots, \frac{N}{2}$. This results in a grid which is denser towards the edges. For elliptic problems, this is not problematic. But for initial value PDEs, this can cause instabilities.

Rmk For analytic u,

$$\breve{u}_n^{(k)} = c_n \sum_{\substack{m=n+1 \\ n+m \text{ odd}}}^{\infty} m\, \breve{u}_m^{(k-1)}, \quad \forall k \geq 1.$$

Method 3.10 (Spectral method for evolutionary PDEs) Consider

$$\begin{cases} \dfrac{\partial u(x,t)}{\partial t} = \mathcal{L} u(x,t) & x \in [-1,1] \quad t \geq 0 \\ u(x,0) = g(x) & x \in [-1,1] \end{cases}$$

with appropriate b.c. on $\{-1,1\} \times [0,\infty)$, and $\mathcal{L}$ linear operator. We solve this using method of lines (semi-discretisation), using spectral method for approx of u and its derivatives in x.

Thus, we seek sol^n $u_N(x,t)$ with

$$u_N(x,t) = \sum_{\#\{n\}=N} c_n(t)\, \varphi_n(x).$$

where $c_n(t)$ are expansion coeff., and $\varphi_n(x)$ basis f^n chosen, e.g.

(i) Fourier expansion with $c_n(t) = \hat{u}_n(t)$, $\varphi_n(x) = e^{i\pi n x}$ for periodic b.c.s

(ii) Poly. expansion like $c_n(t) = \breve{u}_n(t)$, $\varphi_n(x) = T_n(x)$ for other b.c.s.

This results in $N \times N$ system of ODEs for the expansion coeff $\{c_n(t)\}$

$$\underline{c}' = \underline{\underline{B}}\, \underline{c}$$

Can solve with standard ODE solvers, which approximates $\underline{c}(t) = e^{t\underline{\underline{B}}}\, \underline{c}(0)$

Example (Diffusion eqn)

$$\begin{cases} u_t = u_{xx} & (x,t) \in [-1,1] \times \mathbb{R}_+ \\ u(x,0) = g(x) & x \in [-1,1] \end{cases}$$

with periodic b.c. $u(-1,t) = u(1,t)$, $u_x(-1,t) = u_x(1,t)$ and normalisation $\int_{-1}^{1} u(x,t)\,dx = 0$.

For each t, approximate $u(x,t)$ by its N-th partial Fourier sum in x.

$$u(x,t) \approx u_N(x,t) = \sum_{n \in \Gamma_N} \hat{u}_n(t)\, e^{i\pi n x}$$

where $\Gamma_N = \{-N/2+1, \ldots, N/2\}$

Then each coeff $\hat{u}_n$ fulfills

$$\hat{u}_n'(t) = -\pi^2 n^2 \hat{u}_n(t), \quad n \in \Gamma_N$$

(since diagonal)

It's exact soln is $\hat{u}_n(t) = e^{-\pi^2 n^2 t} \hat{g}_n$ for $n \neq 0$, and set $\hat{u}_0(t) = 0$ due to normalisation condition, so that

$$u_N(x,t) = \sum_{n \in \Gamma_n} \hat{g}_n\, e^{-\pi^2 n^2 t}\, e^{i\pi n x}$$

Rmk We can find exact soln due to spectral structure of Laplacian. More general ODE will need a numerical method, so issue of stability arises.

Stability analysis: The system above has the form

$$\hat{\underline{u}}' = B\,\hat{\underline{u}}, \quad B = \mathrm{diag}\{-\pi^2 n^2\}, \ n \in \Gamma_N$$

Note that (a) all evals are negative, (b) they consist of the evals $\lambda_n^{(2)}$ of the second order diff operator, with $\max |\lambda_n^{(2)}| = (N/2)^2$

If we approximate with Euler method

$$\hat{u}^{k+1} = (I + \tau B)\,\hat{u}^k \,, \qquad \tau := \Delta t$$

For stability condition $\|I + \tau B\| \leq 1$, so need to scale the time step $\tau = \Delta t \sim N^{-2}$.

For Crank-Nicolson scheme, since spectrum of B is negative, we can stability for any $\tau > 0$.

For general linear operator $\mathcal{L}$ with const. coeff, B is diag. (hence normal), provided that its spectrum is negative. For stability we need to scale $\tau \sim N^{-m}$, m is the maximal order of differentiation.

The scaling $\tau \sim N^{-2}$ may seem similar to $k \sim h^2$ which we view as a disadvantage. However, we can take N small for good approx.

Example (Diffusion eqn with non-const. coeff.) For $a(x) > 0$, $u = u(x,t)$,

$$\begin{cases} u_t = (a(x)\,u_x)_x & (x,t) \in [-1,1] \times \mathbb{R}_+ \\ u(x,0) = g(x) & x \in [-1,1] \end{cases}$$

with b.c. and norm. condition. Approx. u by its partial Fourier sum results in

$$\hat{u}_n'(t) = -\pi^2 \sum_{m \in \Gamma_N} mn\, \hat{a}_{n-m}\, \hat{u}_m(t) \qquad n \in \Gamma_N$$

We may apply Euler, giving

$$\hat{u}_n^{k+1} = \hat{u}_n^k - \tau \pi^2 \sum_{m \in \Gamma_N} mn\, \hat{a}_{n-m}\, \hat{u}_m(t) \,, \quad \tau = \Delta t$$

or in vector form

$$\hat{\underline{u}}^{k+1} = (I + \tau B)\,\hat{\underline{u}}^{k}$$

with $B = (b_{m,n}) = (-\pi^2 m n \hat{a}_{n-m})$. For stability, we need $\| I + \tau B \| \leq 1$.

Rmk In general, B.C.s for PDEs have to be implemented in Chebyshev expansion. If B.C.s imposed exactly, either the basis ϕ_n have to be slightly modified, e.g. to $T_n(x) - 1$ instead of $T_n(x)$ for B.C. $u(1) = 0$, or we get additional conditions on $\check{u}_n$.
Also, time-dependent B.C. can lead to serious stability problems.

4. Iterative Methods for Linear Systems

A general iterative method for solving $A\underline{x} = \underline{b}$ is a rule $\underline{x}^{k+1} = f_k(\underline{x}^0, \ldots, \underline{x}^k)$. Consider the one-step, stationary iterative scheme

$$\underline{x}^{k+1} = H\underline{x}^k + \underline{v}, \qquad \underline{x}^0, \underline{v} \in \mathbb{R}^n$$

Here, choose $H, \underline{v}$ s.t. $\underline{x}^*$, a sol^n to $A\underline{x} = \underline{b}$, satisfies $\underline{x}^* = H\underline{x}^* + \underline{v}$, i.e. it is a fixed point of the iteration.

Terminology
- iteration matrix H
- error $\underline{e}^k = \underline{x}^* - \underline{x}^k$
- Residual $\underline{r}^k := A\underline{e}^k = \underline{b} - A\underline{x}^k$.

For a given class of matrices A, we are interested in convergent methods, i.e. methods s.t. $\underline{x}^k \to \underline{x}^* = A^{-1}\underline{b}$ for every starting value $\underline{x}^0$.

Subtracting $x^* = Hx^* + v$, we obtain

$$e^{k+1} = He^k = \cdots = H^{k+1}e^0,$$

ie. a method is convergent if $e^k = H^k e^0 \to 0 \ \forall e^0 \in \mathbb{R}^n$.

Scheme 4.1 (Iterative refinement)

$$x^{k+1} = x^k - S(Ax^k - b)$$

If $S = A^{-1}$, then $x^{k+1} = A^{-1}b = x^*$. so it is suggestive to choose S as an approximation to A^{-1}. The iteration matrix for this scheme is $H_S = I - SA$

Scheme 4.2 (Splitting) Assume $A = B + C$ that solving the system with C is "easy", then can consider the scheme

$$Bx^k + Cx^{k+1} = b,$$

eliminating C,

$$(A-B)x^{k+1} = -Bx^k + b$$

with $H = -(A-B)B^{-1}$. Any splitting can be viewed as iterative refinement since

$$(A-B)x^{k+1} = -Bx^k + b \iff (A-B)x^{k+1} = (A-B)x^k - (Ax^k - b)$$
$$\iff x^{k+1} = x^k - (A-B)^{-1}(Ax^k - b)$$

So seek a splitting s.t. $S = (A-B)^{-1}$ approximates A^{-1}.

Thm 4.3 Let $H \in \mathbb{R}^{n\times n}$. then $\lim_{k\to\infty} H^k z = 0$ for any $z \in \mathbb{R}^n$ iff $\rho(H) < 1$.

Pf: Let λ be an eval of H. real or complex, s.t. $|\lambda| = \rho(H) \geq 1$. Let w be a corresponding evec, ie. $Hw = \lambda w$, and

$$\|H^k w\|_\infty = |\lambda|^k \|w\|_\infty \geq \|w\|_\infty =: \gamma > 0$$

If $\underline{w}$ real, choose $\underline{z} = \underline{w}$ so $\|H^k \underline{z}\| \geq \gamma$ can't tend to zero.

If $\underline{w}$ complex, then $\underline{w} = \underline{u} + i\underline{v}$. Then at least one of $(H^k \underline{u})$, $(H^k \underline{v})$ does not tend to 0. Since if both do, then $H^k \underline{w} \longrightarrow 0$, contradiction.

Now let $\rho(H) < 1$. Assume H possess n LI evec $(\underline{w}_j)$ s.t. $H\underline{w}_j = \lambda_j \underline{w}_j$. LI $\Rightarrow \forall z \in \mathbb{R}^n$, $\exists c_j \in \mathbb{C}$ s.t. $\underline{z} = \sum_{j=1}^{n} c_j \underline{w}_j$, thus

$$H^k \underline{z} = \sum_{j=1}^{n} c_j \lambda_j^k \underline{w}_j$$

since $|\lambda_j| \leq \rho(H) < 1$, $\lim_{k \to \infty} H^k \underline{z} = 0$. □

Rmk The complete proof exploits JNF of H, ie $H = S J S^{-1}$.

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix} \qquad J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & & 1 \\ & & & \lambda_i \end{pmatrix} \qquad \sum_i n_i = n$$

To prove $J_i^k \to 0$ if $|\lambda_i| < 1$. split $J_i = \lambda_i I + P$. Notice $P^m = 0$ for $m \geq n$, and

$$(\lambda I + P)^k = \sum_{m=0}^{n_i - 1} \binom{k}{m} \lambda_i^{k-m} P^m$$

Applying thm 4.3 to $\underline{e}^{k+1} = \ldots = H^{k+1} \underline{e}_0$, we get

Thm 4.4 Let $\underline{x}^*$, a soln of $A\underline{x} = \underline{b}$, satisfy $\underline{x}^* = H\underline{x}^* + \underline{v}$, and we are given the scheme

$$\underline{x}^{k+1} = H\underline{x}^k + \underline{v}, \quad \underline{x}^0, \underline{v} \in \mathbb{R}.$$

Then $\underline{x}^k \to \underline{x}^*$ for any choice of $\underline{x}^0$ iff $\rho(H) < 1$.

Method 4.5 (Jacobi and Gauss Seidel) Both of these methods are versions of splitting which can be applied to A with non-zero diag elt. Write

$$A = L_0 + D + U_0$$

L_0: strictly lower triangular; D: diag; U_0: strictly upper triangular

(1) Jacobi method: Set $A - B = D$, and we solve the diagonal system

$$D\underline{x}^{(k+1)} = -(L_0 + U_0)\underline{x}^{(k)} + \underline{b}.$$

and $H_J = -D^{-1}(L_0 + U_0)$

(2) Gauss-Seidel method: Set $A - B = L_0 + D = L$. We solve the tridiagonal system

$$(L_0 + D)\underline{x}^{(k+1)} = -U_0\underline{x}^{(k)} + \underline{b}$$

and $H_{GS} = -(L_0 + D)^{-1}U_0$.

There is no need to invert $(L_0 + D)$, we calculate $\underline{x}^{(k+1)}$ by forward substitution

$$a_{ii}x_i^{(k+1)} = -\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} + b_i$$

$i = 1, \dots, n$

The sequence $\underline{x}^{(k)}$ converges to $A\underline{x} = \underline{b}$ if the spectral radius of $H_J = -D^{-1}(L_0 + U_0)$ or $H_{GS} = -(L_0 + D)U_0$ resp, is less than 1.

We want to prove (a) diag. dominant matrices, and (b) positive def matrices satisfies the above conditions.

Rmk (Gershgorin thm) $\sigma(A) \subset \bigcup_{i=1}^{n} \Gamma_i$, $\Gamma_i := \{z \in \mathbb{C} : |z - a_{ii}| < r_i\}$, $r_i := \sum_{j \neq i} |a_{ij}|$.

Def^n 4.6 (Strictly diag dominant matrices) A is strictly diag dominant by rows if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$, $i = 1, \dots, n$

From Gershgorin thm. strictly diag. dominant matrices are nonsingular.

Thm 4.7 If A strictly diag dom., both Jacobi and Gauss-Seidel methods converge.

Pf: For G-S, evals of $H_{GS} = -(L_0 + D)^{-1} U_0$ satisfies

$$\det(H_{GS} - \lambda I) = \det(-(L_0 + D)^{-1} U_0 - \lambda I) = 0$$

$$\Rightarrow \det(A_\lambda) := \det(U_0 + \lambda D + \lambda L_0) = 0$$

Easy to see $A = L_0 + D + U_0$ strictly diag dom, then for $|\lambda| \geq 1$, $A_\lambda = \lambda L_0 + \lambda D + U_0$ strictly diag dom, so non-sing, so $\det(A_\lambda) = 0$ impossible, so $|\lambda| < 1$, hence convergence.

Similar for Jacobi. □

Thm 4.8 (Householder-John thm) If A, B real s.t. both A and $A - B - B^T$ are sym pos. def., then the spectral radius of $H = -(A-B)^{-1} B$ is strictly less than 1.

Pf: Let λ be an eval of H, so $H\underline{w} = \lambda \underline{w}$, where $\underline{w} \neq 0$ is an evec. (Note $\lambda, \underline{w}$ have non-zero imaginary parts if H not sym). Def^n of H provides the equality $-B\underline{w} = \lambda(A-B)\underline{w}$, and note $\lambda \neq 1$ since otherwise A singular. Thus, we deduce

$$\underline{w}^t B \underline{w} = \frac{\lambda}{\lambda - 1} \underline{w}^t A \underline{w}. \qquad (*)$$

Writing $\underline{w} = \underline{u} + i\underline{v}$, $\underline{u}, \underline{v}$ real, we find (for $C = C^T$).

$$\underline{w}^* C \underline{w} = \underline{u}^T C \underline{u} + \underline{v}^T C \underline{v}$$

So sym pos def. in the assumption implies $\underline{w}^* A \underline{w} > 0$ and $\underline{w}^* (A - B - B^T) \underline{w} > 0$. Use (*) and its conjugate transpose

$$\Rightarrow \quad 0 < \underline{w}^* A \underline{w} - \underline{w}^* B \underline{w} < \underline{w}^* B^T \underline{w}$$

$$= \left(1 - \frac{\lambda}{\lambda - 1} - \frac{\bar{\lambda}}{\bar{\lambda} - 1}\right) \underline{w}^* A \underline{w} = \frac{1 - |\lambda|^2}{|\lambda - 1|^2} \underline{w}^* A \underline{w}.$$

Now $\lambda \neq 1 \Rightarrow |\lambda - 1|^2 > 0$. Recall $\underline{w}^* A \underline{w} > 0$, we see $1 - |\lambda|^2 > 0$

So $|\lambda| < 1$ occurs for every eval of H. □

Cor 4.9 (i) If A sym pos. def., then G-S converges

(ii) If both A and $2D - A$ sym. pos. def., then Jacobi method converges.

Pf: (i) For G-S, B is the superdiagonal part of A, hence $A - B - B^T$ is equal to D, the diagonal part of A. If A pos. def., then D pos. def.

(ii) For Jacobi, $B = A - D$. If A sym, then $A - B - B^T = 2D - A$. (The latter matrix is the same as A except the signs of off-diagonal elts are reversed. □

Example (Poisson eqn on a square) The system $A\underline{x} = \underline{b}$, where A pos (neg) sym def matrix, frequently occur in numerical methods for solving elliptic PDEs. Typical example is Poisson eqn with 5-point formula yields an $n \times n$ system with $n = m^2$ unknowns $u_{p,q}$.

$$u_{p-1,q} + u_{p+1,q} + u_{p,q-1} + u_{p,q+1} - 4u_{p,q} = h^2 f(ph, qh). \qquad (\dagger)$$

(Note that when $p, q = 1, m$, the values in boundary are known and they should be moved to RHS).

For any ordering of grid points (ph, qh) we have shown that A is sym and neg. def.

Cor 4.10 For (†), for any ordering of the grid, both Jacobi and G-S converge.

Pf: A sym neg def, so converges for G-S. For Jacobi, we need neg. def. of $2D - A$. Recall proof of lem 1.5 operates with modulus of off-diag elt and does not depend on their signs. □

Method 4.11 (Relaxation) It is often possible to improve efficiency and the splitting method by relaxation. Specifically, instead of letting $(A-B)\,\underline{x}^{(k+1)} = -B\,\underline{x}^{(k)} + \underline{b}$, we let

$$(A-B)\,\hat{\underline{x}}^{(k+1)} = -B\,\underline{x}^{(k)} + \underline{b}$$

and then

$$\underline{x}^{(k+1)} = \omega\,\hat{\underline{x}}^{(k+1)} + (1-\omega)\,\underline{x}^{(k)}$$

for $k = 0, 1, \ldots$, where ω is a real const. called the relaxation parameter (Note $\omega = 1$ corresponds to standard "unrelaxed" iteration). Good choice of ω leads to smaller spectral radius of the iteration matrix, and the smaller the spectral radius, the faster the iteration converges. To this end, let us express H_ω in terms of $H = -(A-B)^{-1}B$.

$$\hat{\underline{x}}^{(k+1)} = H\underline{x}^{(k)} + \underline{v} \quad \Rightarrow \quad \underline{x}^{k+1} = \omega\hat{\underline{x}}^{(k+1)} + (1-\omega)\,\hat{\underline{x}}^{(k)}$$

$$= \omega H\hat{\underline{x}}^{(k)} + (1-\omega)\,\underline{x}^{(k)} + \omega\underline{v}$$

Hence,

$$H_\omega = \omega H + (1-\omega)I.$$

It follows that the spectra of H_ω and H are related by $\lambda_\omega = \omega\lambda + (1-\omega)$. therefore one can choose $\omega \in \mathbb{R}$ to minimise

$$\rho(H_\omega) = \max\{|\omega\lambda + (1-\omega)\lambda| : \lambda \in \sigma(H)\}.$$

In general, $\sigma(H)$ unknown, but we have some information to find a "good" value of ω.

For example, suppose $\sigma(H)$ is real and resides in $[\alpha, \beta]$, where $-1 < \alpha < \beta < 1$. Seek ω to minimise

$$\max\{|\omega\lambda + (1-\omega)\lambda| : \lambda \in [\alpha, \beta]\}.$$

Readily see that for $\lambda < 1$, $f(\omega) = \omega\lambda + (1-\omega)$ is decreasing. So as ω increases (decreases) from 1, the spectrum of H_ω moves to the left (right) of the spectrum of H. Clear that the optimal location of $\sigma(H_\omega)$ (or of the interval $[\alpha_\omega, \beta_\omega]$ that contains $\sigma(H_\omega)$) is the one which is centralised around the origin.

$$-[\omega\alpha + (1-\omega)] = \omega\beta + (1-\omega)$$

$$\Rightarrow \omega_{opt} = \frac{2}{2-(\alpha+\beta)} \quad , \quad -\alpha_{opt} = \beta_{opt} = \frac{\beta-\alpha}{2-(\alpha+\beta)}$$

Method 4.12 (optimisation of quadratic function) The methods we considered fit into the scheme

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + c_k \underline{d}^{(k)}$$

where we aimed at getting $\rho(H) < 1$ for iteration matrix H. Say, for Jacobi with relaxation, $c_k = \omega$, $\underline{d}^{(k)} = D^{-1}(\underline{b} - A\underline{x}^{(k)})$.

For solving $A\underline{x} = \underline{b}$, $A > 0$, we can solve by successive minimisation of the quadratic f^n

$$F(\underline{x}^{(k)}) := \|\underline{x}^* - \underline{x}^{(k)}\|_A^2 = \|\underline{e}^{(k)}\|_A^2$$

since the minimiser is the exact soln. Here

$$\|y\|_A := (Ay, y)^{1/2} := \sqrt{y^T A y}$$

is a Euclidean-type distance, which is well-defined for $A > 0$.

Approach (Minisation of quadratic fn) The method we introduced are for solving $Ax = b$, fit into

$$x^{(k+1)} = x^{(k)} + c_k d^{(k)}$$

where we aim at getting $\rho(H) < 1$. Say, for Jacobi, we set $c_k = \omega$, $d^{(k)} = D^{-1}(b - Ax^{(k)})$

For solving $Ax = b$ with $A > 0$, can construct iterative method based on successive minimisation of

$$F(x^{(k)}) := \|x^* - x\|_A^2 = \|e^{(k)}\|_A^2,$$

where $\|y\|_A := (Ay, y)^{1/2} := \sqrt{y^T A y}$ is well defined for $A > 0$.

So, at each step k, we are decreasing the A-distance between $x^{(k)}$ and x^*. Thus, for a sym pos def $A > 0$, can choose an iterative method that provides steepest descent condition

$$x^{(k+1)} = x^{(k)} + c_k d^{(k)} \Rightarrow F(x^{(k+1)}) < F(x^{(k)}) \qquad (*)$$

which is equivalent to minising

$$F_1(x) = \frac{1}{2} x^T A x - x^T b$$

which attains minimum when $\nabla F_1(x) = Ax - b$ and does not involve x^*. Easy to check $F_1(x) = \frac{1}{2} F(x) - \frac{1}{2} c$, where $c = x^{*T} A x^*$ const. indpt. of k. hence equivalent.

Example Both J and G-S satisfy (4), precisely

$$(A\underline{e}^{(k+1)}, \underline{e}^{(k+1)}) = (A\underline{e}^{(k)}, \underline{e}^{(k)}) - (C\underline{y}^{(k)}, \underline{y}^{(k)}) < (A\underline{e}^{(k)}, \underline{e}^{(k)}).$$

where G-S: $C = D > 0$. $\underline{y}^{(k)} := (L_0 + D)^{-1} A\underline{e}^{(k)}$

J : $C = D - 2A > 0$. $\underline{y}^{(k)} := D^{-1} A\underline{e}^{(k)}$.

Method 4.13 (A-orthogonal projection) We strengthen the descent condition (*), namely given $\underline{x}^{(k)}$ and $\underline{d}^{(k)}$ (search direction), we seek $\underline{x}^{(k+1)}$ from set of vectors on the line $l = \{ \underline{x}^{(k)} + \alpha \underline{d}^{(k)} \}_{\alpha \in \mathbb{R}}$ s.t. $F(\underline{x}^{(k+1)})$ is as small as possible, i.e.

$$\underline{x}^{(k+1)} := \underset{\alpha}{\operatorname{argmin}} \; F(\underline{x}^{(k)} + \alpha \underline{d}^{(k)}) \qquad (\dagger)$$

Lem 4.14 The minimiser to (†) is

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha_k \underline{d}^{(k)}, \qquad \alpha_k = \frac{(\underline{r}^{(k)}, \underline{d}^{(k)})}{(A\underline{d}^{(k)}, \underline{d}^{(k)})}$$

where $\underline{r}^{(k)} = A\underline{e}^{(k)}$

Pf: We need to choose $\underline{x}^{(k+1)} \in l$ s.t. minimise distance between $\underline{x}^*$ and $y \in l$. Clear that min when $\underline{x}^{(k+1)}$ is A-orthogonal projection of $\underline{x}$ on l, i.e.

$$\underline{x}^* - \underline{x}^{(k+1)} \perp_A \underline{d}^{(k)} \Rightarrow A(\underline{x}^* - \underline{x}^{(k+1)}) \perp \underline{d}^{(k)}$$

$$\Rightarrow \underline{r}^{(k+1)} = \underline{r}^{(k)} - \alpha_k A\underline{d}^{(k)} \perp \underline{d}^{(k)} \qquad \square$$

Method 4.15 (Steepest descent method) Taking

$$\underline{d}^{(k)} = -\nabla F_1(\underline{x}^{(k)}) = \underline{b} - A\underline{x}^{(k)}$$

for every k. The negative gradient of a quadratic fn shows the direction of the (locally) steepest descent at a given point. Thus,

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha_k (\underline{b} - A\underline{x}^{(k)}) \qquad k \geq 0 \qquad (*)$$

Can be proved that $\underline{x}^{(k)}$ cgs to $\underline{x}^*$ of $A\underline{x} = \underline{b}$, but usually the speed of convergence is slow, because (*) decreases value of $F(\underline{x}^{(k+1)})$ locally, but the global decrease, w.r.t. $F(\underline{x}^{(0)})$ is often not that large. The use of <u>conjugation directions</u> provides a method with a global minimisation property.

<u>Conjugate directions</u> For a general direction $\underline{d}$, assume $\underline{x} = \underline{x}^{(k)}$, $\underline{e}^{(k)} = \underline{x}^* - \underline{x}^{(k)}$ error, $\underline{r}^{(k)} = \underline{b} - A\underline{x}^{(k)} = A\underline{e}^{(k)}$ residual. Can write $\langle \underline{r}^{(k)}, \underline{d} \rangle = \langle \underline{e}^{(k)}, \underline{d} \rangle_A$. So with an exact line search,

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \frac{\langle \underline{e}^{(k)}, \underline{d} \rangle_A}{\langle \underline{d}, \underline{d} \rangle_A} \underline{d}$$

Subtracting $\underline{x}^*$, then

$$\underline{e}^{(k+1)} = \underline{e}^{(k)} - \frac{\langle \underline{e}^{(k)}, \underline{d} \rangle_A}{\langle \underline{d}, \underline{d} \rangle_A} \underline{d} \qquad (\dagger)$$

So $\underline{e}^{(k+1)}$ is projection of $\underline{e}^{(k)}$ on the hyperplane that is A-orth. to $\underline{d}$, ie. $\langle \underline{e}^{(k+1)}, \underline{d} \rangle_A = 0$

<u>Defn 4.16</u> (Conjugate directions) $\underline{u}, \underline{v} \in \mathbb{R}^n$ are <u>conjugate</u> w.r.t. to sym. pos. def. A if $\underline{u}, \underline{v} \neq 0$ and $\langle \underline{u}, \underline{v} \rangle_A = \langle \underline{u}, A\underline{v} \rangle = 0$.

<u>Thm 4.17</u> Let $\underline{d}^{(0)}, \ldots, \underline{d}^{(n-1)}$ be n non-zero pairwise conjugate directions, and consider the sequence of iterates

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha_k \underline{d}^{(k)} \qquad \alpha_k = \frac{\langle \underline{r}^{(k)}, \underline{d}^{(k)} \rangle}{\langle \underline{d}^{(k)}, \underline{d}^{(k)} \rangle}$$

Let $\underline{r}^{(k)} = \underline{b} - A\underline{x}^{(k)}$ be residual. Then for each $k = 1, \ldots, n$, $\underline{r}^{(k)}$ orthogonal to span $\{\underline{d}^{(0)}, \ldots, \underline{d}^{(k-1)}\}$. In particular, $\underline{r}^{(n)} = 0$.

Pf: Since $\underline{r}^{(k)} = A\underline{e}^{(k)}$, suffices to show $\underline{e}^{(k)}$ is A-orth. to $\{\underline{d}^{(0)}, \ldots, \underline{d}^{(k-1)}\}$. Induction on k.

$k=0$: nothing to prove

Assume true for $k \geq 0$, and consider (+) with $\underline{d} = \underline{d}^{(k)}$.

By induction hypothesis and the fact that $\underline{d}^{(i)}$ pairwise conjugate directions, we see that $\underline{e}^{(k+1)}$ is A-orth. to $\underline{d}^{(0)}, \ldots, \underline{d}^{(k-1)}$.

Also, $\langle \underline{e}^{(k+1)}, \underline{d}^{(k)} \rangle_A = 0$, so $\underline{e}^{(k+1)}$ A-orth. to $\underline{d}^{(0)}, \ldots, \underline{d}^{(k)}$. □

Rmk Possible to extend the method for solving $Ax = b$ with $A > 0$, $A = A^T$ to other matrices. Suppose want $Bx = c$, B non-sing. Can convert the system to sym pos def. by setting $A = B^TB$, $b = B^Tc$, then solve $Ax = b$ with conjugate gradient.

Algorithm 4.18 (Conjugate gradient method)

(A) For any initial vector $\underline{x}^{(0)}$, set $\underline{d}^{(0)} = \underline{r}^{(0)} = \underline{b} - A\underline{x}^{(0)}$.

(B) For $k \geq 0$, calculate $\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha_k \underline{d}^{(k)}$ and the residual

$$\underline{r}^{(k+1)} = \underline{r}^{(k)} - \alpha_k A\underline{d}^{(k)}, \quad \alpha_k := \{\underline{r}^{(k+1)} \perp \underline{d}^{(k)}\} = \frac{(\underline{r}^{(k)}, \underline{d}^{(k)})}{(A\underline{d}^{(k)}, \underline{d}^{(k)})}, \; k \geq 0$$

(C) For the same k, the next conjugation direction is

$$\underline{d}^{(k+1)} = \underline{r}^{(k+1)} + \beta_k \underline{d}^{(k)}, \quad \beta_k := \{\underline{d}^{(k+1)} \perp A\underline{d}^{(k)}\} = -\frac{(\underline{r}^{(k+1)}, A\underline{d}^{(k)})}{(\underline{d}^{(k)}, A\underline{d}^{(k)})}, \; k \geq 0.$$

Thm 4.19 (Prop of CGM) For every $m \geq 0$, CGM have the following properties:

(1) The linear space spanned by the residuals $\{\underline{r}^{(i)}\}$ is the same as the linear space spanned by the conjugate directions $\{\underline{d}^{(i)}\}$ and it coincides with the space spanned by $\{A^i \underline{r}^{(0)}\}$.

$$\text{span}\{\underline{r}^{(i)}\}_{i=0}^{m} = \text{span}\{\underline{d}^{(i)}\}_{i=0}^{m} = \text{span}\{A^i \underline{r}^{(0)}\}_{i=0}^{m}$$

(2) The residuals satisfy the orthogonality conditions

$$(\underline{r}^{(m)}, \underline{r}^{(i)}) = (\underline{r}^{(m)}, \underline{d}^{(i)}) = 0$$

for $i<m$.

(3) The directions are conjugate (A-orthogonal)

$$(\underline{d}^{(m)}, \underline{d}^{(i)})_A = (\underline{d}^{(m)}, A\underline{d}^{(i)}) = 0$$

for $i<m$.

Pf: Induction on $m \geq 0$.

$m=0$: trivial, since $\underline{d}^{(0)} = \underline{r}^{(0)}$.

Assume true for $m=k$

(1) We have $\underline{d}^{(k+1)} = \underline{r}^{(k+1)} + \beta_k \underline{d}^{(k)}$, implying equivalence of the space spanned by $(\underline{r}^{(i)})_0^k$ and $(\underline{d}^{(i)})_0^k$, is preserved when $k \to k+1$. Similarly, $\underline{r}^{(k+1)} = \underline{r}^{(k)} - \alpha_k A\underline{d}^{(k)}$, and from assumption, $\underline{r}^{(k)}, \underline{d}^{(k)} \in \text{span}\{A^i \underline{r}^{(0)}\}_{i=0}^{k}$, so $\underline{r}^{(k+1)} \in \text{Span}\{A^i \underline{r}^{(0)}\}_{i=0}^{k+1}$.

(2) We need $\underline{r}^{(k+1)} \perp \underline{r}^{(i)}$ for $i \leq k$. By (1), this is equivalent to

$$\underline{r}^{(k+1)} \perp \underline{d}^{(i)} \quad \forall\, i \leq k$$

We have $\underline{r}^{(k+1)} \perp \underline{d}^{(k)}$ by the def^{m} of α_k, so we need

$$\underline{r}^{(k+1)} = \underline{r}^{(k)} - \alpha_k A\underline{d}^{(k)} \perp \underline{d}^{(i)} \quad \forall\, i<k$$

This follows from assumption that $\underline{r}^{(k)} \perp \underline{d}^{(i)}$ and $A\underline{d}^{(k)} \perp \underline{d}^{(i)}$.

(3) We need $\underline{d}^{(k+1)} \perp A\underline{d}^{(i)}$ for $i \leq k$. The value β_k is defined to give $\underline{d}^{(k+1)} \perp A\underline{d}^{(k)}$ so need

$$\underline{d}^{(k+1)} = \underline{r}^{(k+1)} + \beta_k \underline{d}^{(k)} \perp A\underline{d}^{(i)} \text{ for } i<k$$

Assumption says $\underline{d}^{(k)} \perp A\underline{d}^{(i)}$. so remains to show $\underline{r}^{(k+1)} \perp A\underline{d}^{(i)}$ for $i<k$. We have

$$A\underline{d}^{(i)} = \left(\underline{r}^{(i)} - \underline{r}^{(i+1)}\right)/\alpha_i ,$$

So require $\underline{r}^{(k+1)} \perp \left(\underline{r}^{(i)} - \underline{r}^{(i+1)}\right)$ for $i<k$. and this is a consequence from (2) for $m = k+1$. □

Cor 4.20 (A termination property) If CGM is applied in exact arithmetic, then for any $\underline{x}^{(0)} \in \mathbb{R}^n$. termination occurs after at most n iteratives. More precisely, termination occurs after at most s iterations, where

$$s = \dim\left(\operatorname{span}\left\{A^i \underline{r}_0\right\}_{i=0}^{n-1}\right).$$

Rf: (2) from thm 4.19 : $(\underline{r}^{(k)})_{k\geq 0}$ form a sequence of mutually orthogonal vectors in $\mathbb{R}^n$, so at most n of them can be non-zero. Since they belong to the space $\operatorname{span}\{A^i \underline{r}_0\}_{i=0}^{n-1}$, the number is bounded by dim. of space. □

Defn 4.21 (The Krylov subspaces) Let A be $n\times n$ matrix. $\underline{v} \in \mathbb{R}^n \setminus \{0\}$, $m \in \mathbb{N}$. The linear space

$$K_m(A, \underline{v}) := \operatorname{span}\left\{A^i \underline{v}\right\}_{i=0}^{m-1}$$

is called the m-th Krylov subspace in $\mathbb{R}^n$.

Thm 4.22 (No. of iterations in CGM). Let $A>0$, and s be the no. of its distinct evals. then for any $\underline{v}$,

$$\dim K_m(A, \underline{v}) \leq s \quad \forall m \qquad (*)$$

Hence, $\forall A>0$, the no. of iterations of CGM for solving $A\underline{x} = \underline{b}$ is bounded by the number of distinct evals of A.

Pf: (*) is true not just for pos. def $A>0$, but for any A with n linearly independent evecs $(\underline{u}_i)$. Indeed, expand $\underline{v} = \sum_{i=1}^{n} a_i \underline{u}_i$, and group together evecs with same eval: for each λ_ν, set

$$\underline{w}_\nu = \sum_{k=1}^{m_\nu} a_{i_k} \underline{u}_{i_k} \quad \text{if} \quad A\underline{u}_{i_k} = \lambda_\nu \underline{u}_{i_k}, \text{ then}$$

$$\underline{v} = \sum_{\nu=1}^{s} c_\nu \underline{w}_\nu, \qquad c_\nu \in \{0,1\}$$

Hence $A^i \underline{v} = \sum_{\nu=1}^{s} c_\nu \lambda_\nu^i \underline{w}_\nu$. thus for any m we get

$$K_m(A, \underline{v}) \subseteq \text{span}\{\underline{w}_1, \dots \underline{w}_s\}.$$

By Cor 4.20, no. of iteration in CGM is bounded by $\dim K_m(A, \underline{r}^{(0)})$ □

Rmk The thm shows that, unlike other iterative schemes, CGM is both iterative and direct: each iteration produces a reasonable approximation to the exact sol^n, and the exact sol^n will be recovered after n iterations at most.

Simplify and reformulate CGM:

- Rewrite α_k and β_k:

$$\alpha_k = \frac{(\underline{r}^{(k)}, \underline{d}^{(k)})}{(\underline{d}^{(k)}, A\underline{d}^{(k)})} \overset{(c)}{=} \frac{\|\underline{r}^{(k)}\|^2}{(\underline{d}^{(k)}, A\underline{d}^{(k)})} > 0$$

$$\beta_k = -\frac{(\underline{r}^{(k+1)}, A\underline{d}^{(k)})}{(\underline{d}^{(k)}, A\underline{d}^{(k)})} \overset{(a)}{=} -\frac{(\underline{r}^{(k+1)}, \underline{r}^{(k+1)} - \underline{r}^{(k)})}{(\underline{d}^{(k)}, \underline{r}^{(k+1)} - \underline{r}^{(k)})}$$

$$\overset{(b)}{=} \frac{\|\underline{r}^{(k+1)}\|^2}{(\underline{d}^{(k)}, \underline{r}^{(k)})} \overset{(c)}{=} \frac{\|\underline{r}^{(k+1)}\|^2}{\|\underline{r}^{(k)}\|^2} > 0.$$

(a): $A\underline{d}^{(k)}$ is a multiple of $\underline{r}^{(k+1)} - \underline{r}^{(k)}$

(b): $\underline{r}^{(k+1)} \perp \underline{r}^{(k)}, \underline{d}^{(k)}$

(c): $(\underline{d}^{(k)}, \underline{r}^{(k)}) = \|\underline{r}^{(k)}\|^2$, follows from $\underline{d}^{(k+1)} = \underline{r}^{(k+1)} + \beta_k \underline{d}^{(k)}$,

take inner product with $\underline{r}^{(k+1)}$ and use $\underline{r}^{(k+1)} \perp \underline{d}^{(k)}$.

- Let $\underline{x}^{(0)}$ be zero vector

Algo 4.23 (Standard form of CGM).

(1) Set $k=0$, $\underline{x}^{(0)} = 0$, $\underline{r}^{(0)} = b$, $\underline{d}^{(0)} = \underline{r}^{(0)}$.

(2) Calculate $\underline{v}^{(k)} = A\,\underline{d}^{(k)}$. $\alpha_k = \|\underline{r}^{(k)}\|^2 / (\underline{d}^{(k)}, \underline{v}^{(k)}) > 0$

(3) Apply $\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha_k \underline{d}^{(k)}$, $\underline{r}^{(k+1)} = \underline{r}^{(k)} - \alpha_k \underline{v}^{(k)}$.

(4) Stop if $\|\underline{r}^{(k+1)}\|$ small

(5) Set $\underline{d}^{(k+1)} = \underline{r}^{(k+1)} + \beta_k \underline{d}^{(k)}$. where $\beta_k = \|\underline{r}^{(k+1)}\|^2 / \|\underline{r}^{(k)}\|^2 > 0$

(6) Increase $k \to k+1$, go back to (2).

The total work is dominated by number of iterations, multiplied by the time it takes to compute $\underline{v}^{(k)} = A\,\underline{d}^{(k)}$. So CGM highly suitable when most ent of A are zero, i.e. A is sparse.

Technique 4.24 (Preconditioning) In $A\underline{x} = \underline{b}$, change variables $\underline{x} = P^T \hat{\underline{x}}$, P non-sing, and multiply by P

$$PAP^T \hat{\underline{x}} = P\underline{b} \Leftrightarrow \hat{A}\,\hat{\underline{x}} = \hat{\underline{b}}$$

Note that A sym pos def. $\Rightarrow \hat{A} = PAP^T$ sym pos. def. since $(\hat{A}y, y) = (PAP^T y, y) = (AP^T y, P^T y) > 0$. So can apply CGM to the new system. Then we have sol^n $\hat{\underline{x}}$, hence $\underline{x} = P^T \hat{\underline{x}}$. This is called preconditioned CGM and P is preconditioner.

The condition number of A is $\kappa(A) := \|A\| \cdot \|A^{-1}\|$, so for sym pos def A is

$$\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$$

The closer is this number to 1, the faster the convergence of CGM. More precisely, for the rate of convergence of CGM, we have the upper estimate

$$\|e^{(k)}\|_A \leq 2\rho^k \|e^{(0)}\|_A, \quad \rho = \rho_A = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} < 1$$

The main idea of preconditioning is to pick P s.t. $\kappa(\hat{A})$ is much smaller than $\kappa(A)$, thus accelerating convergence.

*Thm 4.25 Given $A \in \mathbb{R}^{n\times n}$, $A > 0$, let $\{d^{(k)}\}_{k=0}^{m-1}$ be a set of conjugate directions, i.e. $(Ad^{(k)}, d^{(i)}) = 0$ for $i<k$. and consider

$$F(x^{(k)}) := \|x^* - x^{(k)}\|_A^2 = \|e^{(k)}\|_A^2.$$

Then the value of $F(x^{(m+1)})$ obtained through CGM coincides with the min value of $F(y)$ taken over all $y = x^{(0)} + \sum_{k=0}^{m} c_k d^{(k)}$, i.e.

$$\arg\min_{c_0,\dots,c_m} F(y) = x^{(m+1)} = x^{(0)} + \sum_{k=0}^{m} \alpha_k d^{(k)}$$

Pf: Every $d^{(k)}$ in CGM is a linear combination of $(A^s r^{(0)})_{s=0}^{i}$, so any vector in the form $x^{(k)} = x^{(0)} + \sum_{i=0}^{k-1} a_i d_i$ can be written as

$$x^{(k)} = x^{(0)} + \sum_{i=0}^{k-1} c_i A^i r^{(0)}$$

$$\Rightarrow \hat{e}^{(k)} = e^{(0)} - \sum_{i=0}^{k-1} c_i A^i r^{(0)}$$

Since $r^{(0)} = Ae^{(0)}$,

$$\hat{e}^{(k)} = x^* - \hat{x}^{(k)} = \left(I - \sum_{i=1}^{k} c_i A^i\right) e^{(0)} = P_k(A)\, e^{(0)},$$

where P_k poly of deg $\leq k$, which satisfies $P_k(0) = 1$.

Recall, at the k-th stage, CGM produces the vector $x^{(k)}$ that minimises

$$F(\hat{x}^{(k)}) = \|\hat{e}^{(k)}\|_A^2 = (A\hat{e}^{(k)}, \hat{e}^{(k)})$$

over all vectors $\hat{x}^{(k)}$ of the form $\hat{x}^{(k)} = x^{(0)} + \sum_{i=0}^{k-1} a_i d^{(i)}$, hence overall all $\hat{e}^{(k)}$. Expressing $e^{(0)} = \sum \gamma_i w_i$. ($w_i$) orthonormal evecs of A, we find $\hat{e}^{(k)} = \sum_i \gamma_i P_k(\lambda_i) w_i$, and $A\hat{e}^{(k)} = \sum_i \gamma_i P_k(\lambda_i)\lambda_i w_i$. and

$$\|\hat{e}^{(k)}\|^2 = \sum_i (P_k(\lambda_i))^2 \lambda_i \gamma_i^2 \leq \max_{\lambda \in \sigma(A)} (P_k(\lambda))^2 \|e^{(0)}\|_A^2.$$

Hence due to minimisation prop. of CGM,

$$\|e^{(k)}\|_A = \min_{P_k} \|\hat{e}^{(k)}\|_A \leq \min_{P_k} \max_{\lambda \in \sigma(A)} |P_k(\lambda)| \, \|e^{(0)}\|_A.$$

Now assume for the spectrum of A, we know the largest and the smallest eval, or some upper and lower bounds, say $0 < m \leq \lambda \leq M$. Then the following minimisation problem, on the class of poly. of deg k arises:

$$P_k(0) = 1, \quad \max_{x \in [m,M]} |P_k(x)| \to \min$$

This has sol^n $P_k^* = T_k^*$, where T_k^* is the Chebyshev poly on the interval $[m,M]$, obtained by dilation and translation of T_k on $[-1,1]$

$$T_k(x) = \cos k\theta, \quad x = \cos\theta, \quad \theta \in [0,\pi]$$

One can show that $|T_k^*(x)| \leq 2\rho^k$ on $[m,M]$, hence a rate of convergence of CGM admits

$$\|e^{(k)}\|_A \leq 2\rho^k \|e^{(0)}\|_A, \quad \rho = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} < 1, \quad \sigma(A) \in [m,M] \qquad \square$$

Note that similarity transform $B \to C^{-1}BC$ preserves spectrum, so

$$\kappa(\hat{A}) = \kappa(PAP^T) = \kappa\left(P^{-1}(PAP^T)P\right) = \kappa(AP^TP)$$

So choose S as an approximation to A which is easy to Cholesky-factorise, i.e. $S = QQ^T$, then take $P = Q^{-1}$, then $AP^TP = AS^{-1}$ is close to identity, hence

$$\kappa(\hat{A}) = \kappa(AP^TP) \approx \kappa(I) = 1 \quad \Rightarrow \kappa(\hat{A}) \ll \kappa(A)$$

then the preconditioned system will be solved much faster.

Each step in CGM for solving $Ax = y$ requires one matrix-vector product Ay, so with $P = Q^{-1}$, additional expense in each step of CGM for preconditioned system while computing $\hat{A}y = PAP^Ty$ is two additional computations

$$u = P^Ty = Q^{-T}y,$$

$$v = Pz = Q^{-1}z.$$

for some $y, z \in \mathbb{R}^n$. Note that computing $Q^{-1}z$ is same as solving $Qv = z$, which is cheap as Q is a lower triangular matrix.

Example (1) Choose S to be $D = \operatorname{diag} A \Rightarrow P = D^{-1/2}$

(2) Choose S as a band matrix with small bandwidth.

e.g. Solving Poisson eqn with 5-point formula

(3) Take $P = L^{-1}$.

Example For $Ax = b$,

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}, \quad S = QQ^T = \begin{pmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$$

S coincides with A except at $(1,1)$-entry. The matrix $\hat{A} = Q^{-1}AQ^T$ for the preconditioned CGM has just 2 evals $\Rightarrow$ recover in two steps.

Note $\hat{A} \sim Q^TQ^{-1}A = S^{-1}A$, hence same spectrum.

Since $A = S + e_1 e_1^T$, $S^{-1}A = I + ue_1^T$, a rank-1 perturbation of the identity matrix, with all evals but one equal to 1 (remaining one is $1+u_1$).

5. Eigenvalues and eigenvectors

Rmk Write $p(\lambda) = \det(A - \lambda I)$. $\lambda \in \mathbb{C}$. We assume $Aw_i = \lambda w_i$, are satisfied by LI w_i, which can be achieved by making an arbitrary small change to A if necessary.

Method 5.1 (The power method) The iterative algorithms that will be studied are closely related to the power method.

Pick a non-zero $x^{(0)} \in \mathbb{R}^n$, then for $k = 0, 1, 2, \ldots$, let $x^{(k+1)}$ be a non-zero multiple of $Ax^{(k)}$, typically to satisfy $\|x^{(k+1)}\| = 1$, so that

$$x^{(k+1)} = Ax^{(k)} / \|Ax^{(k)}\|, \quad k = 0, 1, 2, \ldots$$

This is oriented on finding an evec corresponding to the largest eval as the following thm. shows.

Thm 5.2 Let $A\underline{w}_i = \lambda_i \underline{w}_i$, where evals of A satisfy $|\lambda_1| \le \dots \le |\lambda_{n-1}| < |\lambda_n|$ and the evecs are of unit length $\|\underline{w}_i\| = 1$. Assume $\underline{x}^{(0)} = \sum_{i=1}^{n} c_i \underline{w}_i$, $c_n \neq 0$, then $\underline{x}^{(k)} \to \pm \underline{w}_n$ as $k \to \infty$.

Pf: Given $\underline{x}^{(0)}$ in assumption, $\underline{x}^{(k)}$ is a multiple of

$$A^k \underline{x}^{(0)} = \sum_{i=1}^{n} c_i \lambda_i^k \underline{w}_i = c_n \lambda_n^k \left(\underline{w}_n + \sum_{i=1}^{n-1} \frac{c_i}{c_n} \left(\frac{\lambda_i}{\lambda_n} \right)^k \underline{w}_i \right)$$

Since $\|\underline{x}^{(k)}\| = \|\underline{w}_n\| = 1$, we conclude that $\underline{x}^{(k)} = \pm \underline{w}_n + O(\rho^k)$, where the sign is that of $c_n \lambda_n^k$ and $\rho = \frac{|\lambda_{n-1}|}{|\lambda_n|} < 1$ characterises the rate of convergence. □

Implementation of the procedure:

(0) Pick $\underline{x}^{(0)} \in \mathbb{R}^n$ satisfying $\|\underline{x}^{(0)}\| = 1$. Let ε be a small positive tolerance. Set $k = 0$

(1) Calculate $\tilde{\underline{x}}^{(k+1)} = A\underline{x}^{(k)}$. Set $\lambda = \dfrac{\underline{x}^{(k)T} A \underline{x}^{(k)}}{\underline{x}^{(k)T} \underline{x}^{(k)}}$.

(λ is Raleigh quotient and it minimises $f(\mu) = \|\tilde{\underline{x}}^{(k+1)} - \mu \underline{x}^{(k)}\|$ over all μ)

(2) If $f(\lambda) \le \varepsilon$, accept λ as an eval and $\underline{x}^{(k)}$ as corresponding evec.

(3) O/W, let $\underline{x}^{(k+1)} = \tilde{\underline{x}}^{(k+1)} / \|\tilde{\underline{x}}^{(k+1)}\|$. Increase k by 1. Return to (1).

The termination occurs because, by prev. thm.,

$$\begin{aligned} \|\tilde{\underline{x}}^{(k+1)} - \lambda \underline{x}^{(k)}\| &= \min_{\mu} \|\tilde{\underline{x}}^{(k+1)} - \mu \underline{x}^{(k)}\| \\ &\le \|\tilde{\underline{x}}^{(k+1)} - \lambda_n \underline{x}^{(k)}\| \\ &= \|A\underline{x}^{(k)} - \lambda_n \underline{x}^{(k)}\| = \|A\underline{w}_n - \lambda_n \underline{w}_n\| + O(\rho^k) = O(\rho^k) \to 0. \end{aligned}$$

Rmk (Deficiencies of power method) The method may perform adequately if $c_n \neq 0$, $|\lambda_{n-1}| < |\lambda_n|$, but it is often unacceptably slow. Difficulty of $c_n = 0$ is that the method should find an even w_n with the largest m s.t. $c_m \neq 0$ but practically rounding error introduce a small non-zero component of w_n into sequence $x^{(k)}$, and w_n may be found, but need to wait until the small component to grow.

Method 5.3 (Inverse iteration) We choose

$$(A - sI)\, x^{(k+1)} = x^{(k)}, \quad k = 0, 1, \ldots$$

where s is a scaler that may depend on k and $\|x^{(k)}\| = 1$. So calculation of $x^{(k+1)}$ from $x^{(k)}$ requires the s/l of an $n \times n$ system of linear eqn with matrix $(A - sI)$.

If s const., $\det(A - sI) \neq 0$, then $x^{(k)}$ is a multiple of $(A - sI)^{-k} x^{(0)}$.

Let $x^{(0)} = \sum_{i=1}^{n} c_i w_i$, assuming w_i, $i = 1, \ldots, n$, LI evec of A that satisfy $A w_i = \lambda w_i$. Note that the eval eqn implies

$$(A - sI) w_i = (\lambda_i - s) w_i$$

$$\Rightarrow (A - sI)^{-1} w_i = (\lambda_i - s)^{-1} w_i$$

So $x^{(k)}$ is a multiple of

$$(A - sI)^{-k} x^{(0)} = \sum_{i=1}^{n} c_i (A - sI)^{-k} w_i = \sum_{i=1}^{n} c_i (\lambda_i - s)^{-k} w_i$$

Thus, if m-th number in the set $\{|\lambda_i - s|\}$ is smallest,

and if $c_m \neq 0$, then $x^{(k)}$ tends to be a multiple of w_m as $k \to \infty$. We see that the speed of convergence can be excellent if s close to λ_m. Further, it can be even faster by adjusting s in the calculation.

Algorithm 5.4

(0) Set s to an estimate of an eval of A. Prescribe $x^{(0)} \neq 0$, let $0 < \varepsilon \ll 1$. Set $k = 0$

(1) Calculate (with pivoting if necessary) the LU fact. of $A - sI$.

(2) Stop if U singular. Since s eval is an eval of A. evec in the null space of U. Can be found easily since U upper triangular.

(3) Calculate $x^{(k+1)}$ by solving $(A - sI)\, x^{(k+1)} = LU x^{(k+1)} = x^{(k)}$ using LU from (1).

(4) Set η to the no. that minimises $f(\mu) = \| x^{(k)} - \mu x^{(k+1)} \|$.

(5). Stop if $f(\eta) \leq \varepsilon \| x^{(k+1)} \|$. Since $f(\eta) = \| A x^{(k+1)} - (s+\eta) x^{(k+1)} \|$ we let $s+\eta$ to be the calculated eval of A and $x^{(k+1)} / \| x^{(k+1)} \|$ its evec.

(6) o/w replace $x^{(k+1)}$ by $x^{(k+1)} / \| x^{(k+1)} \|$. Increase k by 1, return to (3) w/o changing s, or to (1) after replacing s by $s+\eta$.

Rmk The algo. is efficient if A upper Hessenberg matrix ($a_{ij} = 0$ for $j < i-1$). The LU requires $O(n^2)$ ($A \neq A^T$), or $O(n)$ ($A = A^T$). So replacement of s by $s+\eta$ need not be expensive, so fast convergence achieved easily.

Thm 5.5 Let A, S be $n \times n$ matrices, S non-sing. Then $\underline{w}$ is an evec of A with eval λ iff $\hat{\underline{w}} = S\underline{w}$ is an evec of $\hat{A} = SAS^{-1}$ with same eval.

Pf: $A\underline{w} = \lambda \underline{w} \Leftrightarrow AS^{-1}(S\underline{w}) = \lambda \underline{w} \Leftrightarrow (SAS^{-1})(S\underline{w}) = \lambda(S\underline{w})$. □

Defn 5.6 (Deflation) Suppose we have found one sol^n of $A\underline{w} = \lambda \underline{w}$, $A \in M^{n \times n}$. Then deflation is the task of constructing an $(n-1) \times (n-1)$ matrix, B say, whose evals are other evals of A.

Specifically, we apply similarity transformation S to A s.t. the first col of $\hat{A} = SAS^{-1}$ is λ times the first coordinate vector $\underline{e}_1$, since it follows from the char. eqn. for eval and we can let B be the bottom submatrix of $\hat{A} = SAS^{-1}$.

Write the condition on S as $(SAS^{-1})\underline{e}_1 = \lambda \underline{e}_1$, the thm shows sufficient if S has $S\underline{w} = c\underline{e}_1$, c any non-zero scaler.

Technique 5.7 (Alg^o of deflation for sym. A) Suppose A sym, $\underline{w} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$ given so that $A\underline{w} = \lambda \underline{w}$. Seek non-sing S s.t. $S\underline{w} = c\underline{e}_1$ and SAS^{-1} also sym. The last condition holds if S orthogonal, since then $S^{-1} = S^T$. It is suitable to pick a Householder reflection, ie. S has the form

$$H_{\underline{u}} = I - 2\underline{u}\underline{u}^T / \|\underline{u}\|^2, \quad \underline{u} \in \mathbb{R}^n.$$

Because Householders reflections are orthogonal, since $H_{\underline{u}}\underline{u} = -\underline{u}$, $H_{\underline{u}}\underline{v} = \underline{v}$ if $\underline{u}^T\underline{v} = 0$, they reflect any vector in $\mathbb{R}^n$ w.r.t. $(n-1)$-dim hyperplane $\perp \underline{u}$.

So for x, y of equal length,

$$H_u x = y, \qquad u = x - y.$$

Hence

$$\left(I - 2\frac{uu^T}{\|u\|^2}\right) w = \pm \|w\| e_1, \qquad u = w \mp \|w\| e_1$$

Since the bottom $n-1$ components coincide, the calculation of u requires only $O(n)$ operations. Further, calculation of SAS^{-1} can be done in $O(n^2)$ operations, taking advantage of the form $S = I - 2uu^T/\|u\|^2$, even if all elts of A are non-zero.

After deflation, we may find evec $\hat{w}$ of SAS^{-1}. The new evec to A is $S^{-1}\hat{w} = S\hat{w}$ since $S^2 = I$.

Method 5.8 (Transformation to upper Hessenberg form) Replace A by $\hat{A} = SAS^{-1}$, where S is a product of Given's rotation $\Omega^{[i,j]}$ chosen to annihilate subdiagonal elts $a_{j,i-1}$ in the $(i-1)$-th col:

$$\begin{bmatrix}****\\****\\****\\****\end{bmatrix} \overset{\Omega^{[2,3]}\times}{\to} \begin{bmatrix}****\\\bullet\bullet\bullet\bullet\\0\bullet\bullet\bullet\\****\end{bmatrix} \overset{\times\Omega^{[2,3]T}}{\to} \begin{bmatrix}*\bullet\bullet*\\*\bullet\bullet*\\0\bullet\bullet*\\*\bullet\bullet*\end{bmatrix} \overset{\Omega^{[2,4]}\times}{\to} \begin{bmatrix}****\\\bullet\bullet\bullet\bullet\\0***\\0\bullet\bullet\bullet\end{bmatrix} \overset{\times\Omega^{[2,4]T}}{\to} \begin{bmatrix}*\bullet\bullet*\\*\bullet\bullet*\\0\bullet\bullet*\\0\bullet\bullet*\end{bmatrix} \overset{\Omega^{[3,4]}\times}{\to} \begin{bmatrix}****\\****\\0\bullet\bullet\bullet\\00\bullet\bullet\end{bmatrix} \overset{\times\Omega^{[3,4]T}}{\to} \begin{bmatrix}**\bullet\bullet\\**\bullet\bullet\\0*\bullet\bullet\\00\bullet\bullet\end{bmatrix}$$

The $\bullet$-elements have changed through a single transformation while the $*$-elements remained the same.

Alternatively, can use Householder transformation. Choose u s.t. with $H_u = I - 2uu^T/\|u\|^2$, the i-th col of $\tilde{B} = H_u B$ is consistent with upper Hessenberg form. Such u has its first i coords vanishing, so $\hat{B} = \tilde{B} H_u^T$ has first i cols unchanged, and all new and old zeros stay untouched.

Algorithm 5.9 (QR algo) Set $A_0 = A$. For $k=0,1,\dots$, the QR fact. $A_k = Q_k R_k$ (Q_k orthogonal, R_k upper triangular) and set $A_{k+1} = R_k Q_k$

The evals of A_{k+1} are the same as the evals of A_k, since

$$A_{k+1} = R_k Q_k = Q_k^{-1}(Q_k R_k) Q_k = Q_k^{-1} A_k Q_k.$$

a similarity transformation. Moreover, $Q_k^{-1} = Q_k^T$, so if A_k sym, so is A_{k+1}.

If for some $k \geq 0$, A_{k+1} can be regarded as 'deflated', ie. it has block form

$$A_{k+1} = \begin{pmatrix} B & C \\ D & E \end{pmatrix}.$$

where B, E square, $D \approx 0$. We calculate evals of B and E separately (with QR, except there is nothing to calculate for 1×1 and 2×2 blocks).

Technique 5.10 (QR for upper Hessenberg matrices) If A_k upper Hessenberg, then its QR fact by Given's rotation produces

$$R_k = Q_k^T A_k = \Omega^{[n-1,n]} \cdots \Omega^{[2,3]} \Omega^{[1,2]} A_k.$$

QR sets $A_{k+1} = R_k Q_k = R_k \Omega^{[1,2]T} \Omega^{[2,3]T} \cdots \Omega^{[n-1,n]T}$, and it follows that A_{k+1} also upper Hessenberg, since

$$\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix} \xrightarrow{\times\Omega^{[1,2]T}} \begin{bmatrix} \bullet & \bullet & * & * \\ \bullet & \bullet & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix} \xrightarrow{\times\Omega^{[2,3]T}} \begin{bmatrix} * & \bullet & \bullet & * \\ * & \bullet & \bullet & * \\ 0 & \bullet & \bullet & * \\ 0 & 0 & 0 & * \end{bmatrix} \xrightarrow{\times\Omega^{[3,4]T}} \begin{bmatrix} * & * & \bullet & \bullet \\ * & * & \bullet & \bullet \\ 0 & * & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \end{bmatrix}$$

Then in QR, Q_k is a product of $n-1$ Givens rot^n, so each iteration is $O(n^2)$.

Technique 5.11 (QR for sym matrices) Bring A to upper Hessenberg form, so QR commences from a sym. tridiag. matrix A_0, then above technique applied for every k. Since Hessenberg and sym both conserved, A_{k+1} also sym tridiag. Each QR iteration is $O(n)$.

Notation Write $\bar{Q}_k = Q_0 \dots Q_k$, $\bar{R}_k = R_k \dots R_0$, $k=0,1,\dots$

Note that $\bar{Q}_k$ orthogonal, $\bar{R}_k$ upper triangular.

Lem 5.12 A_{k+1} is related to original matrix A by the similarity transformation $A_{k+1} = \bar{Q}_k^T A \bar{Q}_k$. Further, $\bar{Q}_k \bar{R}_k$ is the QR fact. of A^{k+1}.

Pf: Prove the first assertion by induction. First,

$$A_1 = Q_0^T A_0 Q_0 = \bar{Q}_0^T A \bar{Q}_0$$

Assuming $A_k = \bar{Q}_{k-1}^T A \bar{Q}_{k-1}$,

$$\begin{aligned} A_{k+1} &= Q_k^T A_k Q_k \\ &= Q_k^T (\bar{Q}_{k-1}^T A \bar{Q}_{k-1}) Q_k = \bar{Q}_k^T A \bar{Q}_k \end{aligned}$$

Second assertion is true for $k=0$, since $\bar{Q}_0 \bar{R}_0 = \bar{Q}_0 \bar{R}_0 = A_0 = A$.

Assuming $\bar{Q}_{k-1} \bar{R}_{k-1} = A^k$, then

$$\begin{aligned} \bar{Q}_k \bar{R}_k &= (\bar{Q}_{k-1} Q_k)(R_k \bar{R}_{k-1}) \\ &= \bar{Q}_{k-1} A_k \bar{R}_{k-1} \\ &= \bar{Q}_{k-1} (\bar{Q}_{k-1}^T A \bar{Q}_{k-1}) \bar{R}_{k-1} \\ &= A \bar{Q}_{k-1} \bar{R}_{k-1} \\ &= A \cdot A^{k-1} = A^k . \end{aligned}$$

□

Rmk Assume eval of A has diff. magnitude. $|\lambda_1| < |\lambda_2| < \cdots < |\lambda_n|$.
Let $\underline{e}_1 = \sum_{i=1}^{n} c_i \underline{w}_i = \sum_{i=1}^{m} c_i \underline{w}_i$ be the expansion of the first coord. vector in terms of normalised evec of A, where m is the largest integer s.t. $c_m \neq 0$.

Consider the first col. of both sides of the matrix eqn $A^{k+1} = \bar{Q}_k \bar{R}_k$.
By the power method arguments, $A^{k+1} \underline{e}_1$ is a multiple of $\sum_{i=1}^{m} c_i (\lambda_i/\lambda_m)^{k+1} \underline{w}_i$, so first col. of A^{k+1} tends to a multiple of $\underline{w}_m$ for $k \gg 1$.
On the other hand, if $\underline{q}_k$ is the first col. of $\bar{Q}_k$, since $\bar{R}_k$ upper triangular, first col. of $\bar{Q}_k \bar{R}_k$ is a multiple of $\underline{q}_k$.

So $\underline{q}_k$ tends to be a multiple of $\underline{w}_m$. Further, since $\underline{q}_k$, $\underline{w}_m$ has unit length, $\underline{q}_k = \pm \underline{w}_m + \underline{h}_k$, where $\underline{h}_k \to 0$ as $k \to \infty$. So

$$A \underline{q}_k = \lambda_m \underline{q}_k + O(\underline{1}) \quad , k \to \infty.$$

Thm 5.13 Let $|\lambda_1| < \cdots < |\lambda_n|$. $\underline{e}_1 = \sum_{i=1}^{n} c_i \underline{w}_i = \sum_{i=1}^{m} c_i \underline{w}_i$, then as $k \to \infty$, the first col. of A_k tends to $\lambda_m \underline{e}_1$, making A_k suitable for deflation.

Pf: The first col. of A_{k+1} is $\bar{Q}_k^T A \bar{Q}_k \underline{e}_1$, and

$$\begin{aligned} A_{k+1} \underline{e}_1 = \bar{Q}_k^T A \bar{Q}_k \underline{e}_1 &= \bar{Q}_k^T A \underline{q}_k \\ &= \bar{Q}_k^T (\lambda_m \underline{q}_k + O(\underline{1})) \quad \left.\right\} \bar{Q}_k^T \underline{q}_k = \underline{e}_1 \\ &= \lambda_m \underline{e}_1 + O(\underline{1}) \quad \because \text{orthogonal} \end{aligned}$$

and $\bar{Q}_k \underline{x} = O(\underline{x})$ since orthogonal mapping is isometry. □

Rmk In practise, as $k \to \infty$, the off-diag elt in the bottom row of $A_{k+1} \to 0$ must faster than the off-diag elt in first col.

Let $|\lambda_1| < \dots < |\lambda_n|$. $\underline{e}_n^T = \sum_{i=1}^{n} c_i \underline{v}_i^T = \sum_{i=s}^{n} c_i \underline{v}_i^T$ be the expansion of last coord row vector $\underline{e}_n^T$ is the basis of <u>left eigenvectors</u> of A. ie. $\underline{v}_i^T A = \lambda_i \underline{v}_i^T$. where s is least integer s.t. $c_s \neq 0$.

Assume $\det A \neq 0$, can write $A^{k+1} = \bar{Q}_k \bar{R}_k$ in the form $A^{-(k+1)} = \bar{R}_k^{-1} \bar{Q}_k^T$.

Consider bottom rows of both sides

$$\underline{e}_n^T A^{-(k+1)} = (\underline{e}_n^T \bar{R}_k^{-1}) \bar{Q}_k^T.$$

By inverse iteration arguments. $\underline{e}_n^T A^{-(k+1)}$ is a multiple of $\sum_{i=s}^{n} c_i (\lambda_s/\lambda_i)^{k+1} \underline{v}_i^T$ so bottom row of $A^{-(k+1)}$ tends to a multiple of $\underline{v}_s^T$. Let $\underline{f}_k^T$ be bottom row of $\bar{Q}_k^T$, since $\bar{R}_k$ upper tri. $\bar{R}_k^{-1}$ is upper tri, and so bottom row of $\bar{R}_k^{-1} \bar{Q}_k^T$ is a multiple of $\underline{f}_k$.

So $\underline{f}_k$ tends to a multiple of $\underline{v}_s^T$, and since unit length, $\underline{f}_k = \pm \underline{v}_s^T + \underline{h}_k^T$, $\underline{h}_k \to 0$, ie.

$$\underline{f}_k^T A = \lambda_s \underline{f}_k^T + o(1), \quad k \to \infty.$$

<u>Thm 5.14</u> Assume above condition, then as $k \to \infty$, the bottom row of A_k tends to $\lambda_s \underline{e}_n^T$, making A_k suitable for deflation.

<u>Pf</u>:

$$\begin{aligned} \underline{e}_n^T A_{k+1} &= \underline{e}_n^T \bar{Q}_k^T A \bar{Q}_k = \underline{f}_k^T A \bar{Q}_k \\ &= (\lambda_s \underline{f}_k^T + o(1)) \bar{Q}_k \\ &= \lambda_s \underline{e}_n^T + o(1) \end{aligned}$$

□

Technique 5.15 (Single shifts) The better we can estimate $s_k \approx \lambda_k$, the more we can accomplish by a step of inverse iteration with shifted matrix $A_k - s_k I$. Thm 5.14 shows $(A_k)_{nn}$ is a good estimate of λ_s. So replace A_k by $A_k - s_k I$, $s_k = (A_k)_{nn}$, before QR fact:

$$A_k - s_k I = Q_k R_k$$

$$A_{k+1} = R_k Q_k + s_k I.$$

A good approx $s_k = (A_k)_{nn}$ to λ_s generates even better approx. of $s_{k+1} = (A_{k+1})_{nn}$ and convergence is accelerating at a higher rate. Note that

$$A_{k+1} = Q_k^T (Q_k R_k + s_k I) Q_k = Q_k^T A_k Q_k$$

hence $A_{k+1} = \bar{Q}_k^T A \bar{Q}_k$. but note $\bar{Q}_k \bar{R}_k \neq A_{k+1}$. but we have

$$\bar{Q}_k \bar{R}_k = \prod_{m=0}^{k} (A - s_m I)$$

Method 5.16 (Multigrid method) Consider $Ay = b$ from 3-point formula on m-grid $\Omega_m = \{ih : 1 \le i \le m\}$, $h = 1/(m+1)$.

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

$D = 2I$, so the weighted Jacobi looks like

$$\underline{u}^{(\nu+1)} = H_\omega \underline{u}^{\nu} + (\omega/2)\underline{b},$$

$\nu = 0, 1, \ldots$. $H_1 = I - D^{-1}A = I - \frac{1}{2}A$, $H_\omega = \omega H_1 + (1-\omega) I = I - \frac{\omega}{2} A$.

The error decay in terms of H_ω are

$$\underline{e}^{(\nu)} = H_\omega^{\nu} \underline{e}^{(0)}$$

and evals of H_ω are

$$\underline{\omega}^k = \left(\sin i \frac{k\pi}{m+1} \right)_{i=1,\ldots,m}, \quad \lambda_k(\omega) = 1 - 2\omega \sin^2 \frac{k\pi}{2(m+1)}, \quad k = 1, \ldots, m$$

Consider choice $\omega = 1/2$, evals of H_ω are $\lambda_k = \cos^2 \frac{k\pi}{2(m+1)}$.
In particular $\rho(H_\omega) = \lambda_1 \simeq 1 - \frac{\pi^2}{4m^2} < 1$. So convergence despite a slow one when m large.

However, expanding the error w.r.t. evecs, we obtain

$$\underline{e}^{(\nu)} \sum_{k=1}^{m} a_k^{(\nu)} \underline{w}^k, \quad \underline{e}^{(\nu)} = H_\omega^\nu \underline{e}^{(0)} \Rightarrow |a_k^{(\nu)}| = |\lambda_k|^\nu |a_k^{(0)}|$$

So components decay at a different rate for diff. freq $k = 1, \dots, m$. For high freq. where k close to m. will decay faster than low freq. We say that $k \in (0, \frac{1}{h})$ is <u>high freq</u> w.r.t. Ω_h if $kh \geq \frac{1}{2}$, then decay rate is at least

$$\mu_* = |\lambda_{(m+1)/2}| = 1 - \sin^2(\pi/4) = \frac{1}{2}.$$

So for coeff. of HF components of $\underline{e}^{(\nu)}$. we obtain

$$|a_k^{(\nu)}| \leq |\mu_*|^\nu |a_k^{(0)}| = \left(\frac{1}{2}\right)^\nu |a_k^{(0)}| \ll |a_k^{(0)}|$$

For low freq $k \in \left(\frac{1}{4h}, \frac{1}{2h}\right)$ w.r.t. Ω_h become high freq for the coarser grid Ω_{2h} with step $2h$. For such k, $k(2h) \geq \frac{1}{2}$.

Algorithm 5.17 (MGV)

1. If A small enough, use direct method to solve $A\underline{u} = \underline{b}$.
2. Presmoothing: perform a small number (≤ 5) of J or GS on $A\underline{u} = \underline{b}$ starting from $\underline{u}^0$.
3. Let $\underline{r} = \underline{b} - A\underline{u}$ be residual, with $\underline{u}$ from the previous step.
4. Let $I_{2h}^{h} : \mathbb{R}^{\frac{m+1}{2}-1} \to \mathbb{R}^{m}$ be interpolation that interpolates vectors on coarse grid Ω_{2h} to vectors in fine grid Ω_h. Let $R_h^{2h} : \mathbb{R}^m \to \mathbb{R}^{\frac{m+1}{2}-1}$ be restriction operator that restricts vectors on Ω_h to Ω_{2h}.
5. Let $\tilde{A} = R_h^{2h} A I_{2h}^{h}$ of size $\approx m/2 \times m/2$
6. Recurse: let $\tilde{\underline{\delta}} = MGV(\tilde{A}, R_h^{2h}\underline{r}, 0)$.
7. Let $\underline{u} = \underline{u} + I_{2h}^{h}\tilde{\underline{\delta}}$.
8. Post smoothing: apply a few J or GS starting from $\underline{u}$ on $A_h\underline{u} = \underline{b}$
9. Return $\underline{u}$.