ADVANCED NUMERICAL ANALISIS

Part 1

FDM for the solution of boundary value problem for the second order ODE.

$$y'' = p(x)y + q(x) \quad \text{on } (a,b) \qquad (1)$$

$$y(a)=\alpha\,,\ y(b)=\beta \qquad (2)$$

or

$$y'(a)-\alpha_0 y(a)=\alpha_1\,,\ y'(b)+\beta_0 y(b)=\beta_1 \qquad (3)$$

Assume that $p(x), q(x) \in C^2[a,b]$, and

$p(x)\geq 0$; $\alpha_0, \beta_0 \geq 0$, $\alpha_0^2+\beta_0^2>0$.

The solution (1),(2) and (1),(3) $\in C^4[a,b]$.

Let us divide $[a,b]$ into n subintervals by the equally spaced points

$$x_i = a+ih \quad \left(i=0,1,2,\dots,n;\ h=\frac{b-a}{n}\right).$$

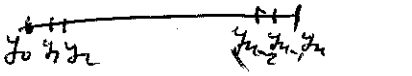
we call the points x_i the grids.

If we replace in (1) the second order derivative y'' at x_i by

$$y''(x_i) \approx \frac{y_{i+1}-2y_i+y_{i-1}}{h^2},$$

and $y'(a)$ and $y'(b)$ in (3) by

$$\frac{-y_2+4y_1-3y_0}{2h} \quad \text{and} \quad \frac{3y_n-4y_{n-1}+y_{n-2}}{2h}$$



Show that these are the second order formulae

For (1),(2), we have

For the numerical solution

$$\frac{y_{i+1}-2y_i+y_{i-1}}{h^2} - p_i y_i = q_i \qquad (i=1,2,\dots,n-1) \qquad (1')$$

$$y_0=\alpha\,,\ y_n=\beta. \qquad (2')$$

and for (1),(3)

$$\frac{y_{i+1}-2y_i+y_{i-1}}{h^2} - p_i y_i = q_i \qquad (i=1,2,\dots,n-1)\,, \qquad (1')$$

$$\frac{-y_2+4y_1-3y_0}{2h} - \alpha_0 y_0 = \alpha_1\,,\quad \frac{3y_n-4y_{n-1}+y_{n-2}}{2h} + \beta_0 y_n = \beta_1\,. \qquad (3')$$

We prove solvability of these systems. To do this we consider the following Lemma.

Lemma 1. If for the system of numbers $y_0, y_1, \to y_n$ $(y_i \neq C)$

$$L(y_i) \equiv \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - p_i y_i \geq 0 \quad (i = 1, 2, \to n-1) \quad (4)$$

then the maximum positive number among y_i can be only y_0 or y_n.

Proof. If $y_k = M$ is the maximal positive number from $y_0, y_1, \to y_n$, such that at least one of the numbers y_{k-1} or y_{k+1} less that M. By assumption

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - p_k y_k \geq 0.$$

If we replace y_{k-1} and y_{k+1} by M, then we increase the left hand side. Therefore,

$$\frac{M - 2M + M}{h^2} - p_k M = -p_k M > 0$$

But it impossible, since $p_k \geq 0$ and $M > 0$

Similarly we prove the next

Lemma 2. If the system of numbers $y_0, y_1, \to y_n$ $(y_i \neq C)$, $L(y_i) \leq 0$ for all $i = 1, 2, \to n-1$, then the minimum negative number among them can be only y_0 or y_n.

Let us prove the solvability of (1') (2'). For this it is enough to show that the corresponding homogeneous system

$$\left.\begin{aligned} &\frac{z_{i+1} - 2z_i + z_{i-1}}{h^2} - p_i z_i = 0 \quad (i = 1, 2, \to n-1), \\ &z_0 = 0, \quad z_n = 0 \end{aligned}\right\} \quad (5)$$

has only trivial solution.

If the system (5) has nontrivial solution, then ~~it should be~~ among $z_1, z_2, \dots, z_{n-1}$ must exist maximal positive or minimal negative number.
But this contradicts to lemmas 1 or 2, i.e., $l(z_i)=0$ for all $i=1,2,\dots,n-1$.

Now we prove the solvablity of the system (1'),(3').

We consider the nonhomogeneous system

$$\frac{z_{i+1}-2z_i+z_{i-1}}{h^2} - p_i z_i = 0 \quad (i=1,2,\dots,n-1), \qquad (6)$$

$$-\frac{z_2+4z_1-3z_0}{2h} - \alpha_0 z_0 = 0, \quad \frac{3z_n-4z_{n-1}+z_{n-2}}{2h} + \beta_0 z_n = 0.$$

We eliminate z_2 from eq.s

$$\frac{z_2-2z_1+z_0}{h^2} - p_1 z_1 = 0 \quad \text{and} \quad \frac{-z_2+4z_1-3z_0}{2h} - \alpha_0 z_0 = 0$$

we have

$$\left.\begin{aligned} z_2 &= h^2 p_1 z_1 + 2z_1 - z_0 \\ z_2 &= -2h\alpha_0 z_0 + 4z_1 - 3z_0 \end{aligned}\right\} \quad h^2 p_1 z_1 + 2z_1 - z_0 = -2h\alpha_0 z_0 + 4z_1 - 3z_0$$

$$z_1 = \frac{-2h\alpha_0 - 2}{h^2 p_1 - 2} z_0 = \frac{2+2h\alpha_0}{2-h^2 p_1} z_0$$

$$\boxed{z_1 = \frac{1+\alpha_0 h}{1-\frac{1}{2}p_1 h^2} z_0} \qquad (7)$$

Similarly,

$$\boxed{z_{n-1} = \frac{1+\beta_0 h}{1-\frac{1}{2}p_{n-1} h^2} z_n} \qquad (8)$$

If $z_i \not\equiv 0$ of (6). Since $l(z_i)=0$ $(i=1,2,\dots,n-1)$, then max. (positive) or min. (negative) value can be only z_0 or z_n. Let z_0 be max. positive value. Assume that h is so small that $\frac{1}{2}p_i h^2 < 1$. From (7) follows that $z_1 \geq z_0$. Since z_1 can not be greater than z_0, then $z_1 = z_0 \Rightarrow \alpha_0 = 0$ and $p_1 = 0 \Rightarrow$ all z_i are the same, and $z_{n-1} = z_n$. From (8) follows that $\beta_0 = 0$, $p_{n-1} = 0$, but $\alpha_0^2 + \beta_0^2 > 0$.

There is no
By contradiction ~~a~~ nontrivial solution which takes the maximal positive value on the boundary. Similarly there is no $\equiv$ solution which takes the negative value on the boundary. These mean that the system (6) has only trivial solution $\Rightarrow$ the system (1'), (3') for any q_i, α_i, β_i has a unique solution.

We estimate an error of the approximate solution of (1), (2) obtained by this method.

Lemma 3. If we have two systems $y_0, y_1, \dots, y_n$ and $Y_0, Y_1, \dots, Y_n$ such that

$$l(Y_i) \le -|l(y_i)| \quad (i=1,2,\dots,n-1), \tag{7}$$

$$|y_0| \le Y_0, \ |y_n| \le Y_n, \tag{8}$$

then for all i

$$|y_i| \le Y_i. \tag{9}$$

Proof. From (7) follows that

$$l(Y_i) + |l(y_i)| \le 0 \Rightarrow l(Y_i) \pm l(y_i) \le 0$$

$$\Rightarrow l(Y_i \pm y_i) \le 0, \ Y_0 \pm y_0 \ge 0, \ Y_n \pm y_n \ge 0.$$

On the basis of Lemma 2 follows that $Y_i \pm y_i$ cannot have negative values, i.e., $Y_i \pm y_i \ge 0$

$$\Rightarrow |y_i| \le Y_i.$$

We denote by y_i the exact solution of the BVP (1), (2) at the grid x_i, i.e., $y_i = y(x_i)$, by $\tilde{y}_i$ the approximate solution obtained as the solution of the system (1'), (2'), and by ε_i an error of this solution, i.e., $\varepsilon_i = y_i - \tilde{y}_i$.

It is easy to show that

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - p_i y_i = \varphi_i + R_i,$$

where R_i is the error of approximation, when the second order derivative is approximated by the central difference formula, and it is known that

$$|R_i| \le \frac{h^2}{12} M_4, \quad M_4 = \max_{x \in (a,b)} |y^{(4)}(x)|.$$

Then ε_i will satisfy the system of equations

$$\frac{\varepsilon_{i+1} - 2\varepsilon_i + \varepsilon_{i-1}}{h^2} - p_i \varepsilon_i = R_i \quad (i = 1, 2, \dots, n-1),$$

$$\varepsilon_0 = \varepsilon_n = 0.$$

We consider a system of numbers η_i, which is a solution of the system

$$\frac{\eta_{i+1} - 2\eta_i + \eta_{i-1}}{h^2} - p_i \eta_i = -\frac{h^2}{12} M_4 \quad (i = 1, 2, \dots, n-1),$$

$$\eta_0 = \eta_n = 0.$$

On the basis of Lemma 3, we obtain that (show!)

$$|\varepsilon_i| \le \eta_i.$$

If we consider the solution ζ_i of the system

$$\frac{\zeta_{i+1} - 2\zeta_i + \zeta_{i-1}}{h^2} = -\frac{h^2}{12} M_4 \quad (i = 1, 2, \dots, n-1),$$

$$\zeta_0 = \zeta_n = 0,$$

then again by Lemma 3, we have $\eta_i \le \rho_i$, therefore

$$|\varepsilon_i| \le \rho_i$$

Show! (At Home done)

The solution ρ_i can be found easily. Indeed, since the second order finite difference of ρ_i is constant, then ρ_i can be consider as a value of some polynomial of second order at x_i. This polynomial is

$$P_2(x) = \frac{h^2}{24} M_4 (x-a)(b-x), \quad P_2(a) = P_2(b) = 0$$

and $\rho_i = P_2(x_i)$

$$\tilde{P}_2(x) = xb - x^2 - ab + ax = -x^2 + (b+a)x - ab$$

$$\tilde{P}_2'(x) = -2x + (b+a)$$

$$\tilde{P}_2''(x) = -2$$

$$\frac{\tilde{P}_{2,i+1} - 2\tilde{P}_{2,i} + \tilde{P}_{2,i-1}}{h^2} = \tilde{P}_2'' + O = -2, \quad P^{IV}(x) = 0$$

$$M_4 \frac{h^2}{24}(-2) = -\frac{h^2}{12} M_4$$

$$\max_{x\in[a,b]} P_2(x) = ?$$

$$P_2'(x) = \frac{h^2}{24} M_4 (-2x + (b+a)) = 0$$

$$x = \frac{a+b}{2}$$

$$P_2''(x) = -\frac{h^2 M_4}{12} < 0$$

So

$$\max_{x\in[a,b]} P_2(x) = P_2\left(\frac{a+b}{2}\right) = \frac{h^2}{24} M_4 \left(\frac{a+b}{2} - a\right)\left(b - \frac{a+b}{2}\right)$$

$$= \frac{h^2}{24} M_4 \frac{b-a}{2} \cdot \frac{b-a}{2} = \frac{h^2 (b-a)^2}{96} M_4$$

So

$$|\varepsilon_i| \le \frac{h^2 (b-a)^2}{96} M_4$$

$$\Rightarrow \varepsilon_i \to 0 \text{ as } h \to 0.$$

$\ell(\xi_i) = R_i \quad \xi_0 = \xi_n = 0$

$\ell(\eta_i) = -\frac{h^2}{12} M_4 \quad \eta_0 = \eta_n = 0$

$|\ell(\xi_i)| = |R_i|$

$|R_i| \le |\ell(\xi_i)| \le \frac{h^2}{12} M_4$

$\ell(\eta_i) \le -\frac{h^2}{12} M \le -|\ell(\xi_i)| = -|R_i|$

$-\frac{h^2}{12} M_4 \le |\ell(\xi_i)| \le \frac{h^2}{12} M_4$

$|R_i| \le \frac{h^2}{12} M_4$

$-\frac{h^2}{12} M_4 \le -|R_i| = -|\ell(\xi_i)| \quad -|R_i| \ge -\frac{h^2}{12} M_4$

$\ell(\eta_i) = -\frac{h^2}{12} M \le -|\ell(\xi_i)|$

$\ell(\eta_i) \le -|\ell(\xi_i)|$

$\ell(\eta_i \pm \xi_i) \le 0, \quad \eta_0 \pm \xi_0 = 0$

$\eta_i \pm \xi_i \ge 0$

$|\xi_i| \le \eta_i$

$\frac{\rho_{i+1} - 2\rho_i + \rho_{i-1}}{h^2} - p_i \rho_i = -\frac{h^2}{12} M_4 - p_i \rho_i = -\left(\frac{h^2}{12} M_4 + p_i \rho_i\right)$

$\ell(\rho_i \pm \eta_i) = \ell(\rho_i) \pm \ell(\eta_i) = -\left(\frac{h^2}{12} M_4 + p_i \rho_i\right) \pm \left(-\frac{h^2}{12} M_4\right)$

$-\frac{h^2}{6} - p_i \rho_i \le 0$

$-\left(\frac{h^2}{12} M_4 + p_i \rho_i\right) - \left(-\frac{h^2}{12} M_4\right)$

$= -p_i \rho_i \le 0$

$$y'' - p(x)y = q(x) \quad \text{on } (a,b) \tag{1}$$

$$y'(a) - \alpha_0 y(a) = \alpha, \quad y'(b) + \beta_0 y(b) = \beta_1 \tag{2}$$

$p(x) \geq 0$; $p(x), q(x) \in C^2[a,b]$, $\boxed{\alpha_0^2 + \beta_0^2 > 0}$

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - p_i y_i = q_i \quad , \quad i = 1,2,\dots,n-1 \tag{1'}$$

~~$y_0 = \alpha$, $y_n = \beta$~~

$$-\frac{y_2 + 4y_1 - 3y_0}{2h} - \alpha_0 y_0 = \alpha_1 \quad , \quad \frac{3y_n - 4y_{n-1} + y_{n-2}}{2h} + \beta_0 y_n = \beta_1 \tag{2'}$$

$$\left.\begin{aligned} &\frac{z_{i+1} - 2z_i + z_{i-1}}{h^2} - p_i z_i = 0 \quad , \quad i = 1,2,\dots,n-1. \\ &-\frac{z_2 + 4z_1 - 3z_0}{2h} - \alpha_0 z_0 = 0 \quad , \quad \frac{3z_n - 4z_{n-1} + z_{n-2}}{2h} + \beta_0 z_n = 0 \end{aligned}\right\} (*)$$

$i=1$: $\dfrac{z_2 - 2z_1 + z_0}{h^2} - p_1 z_1 = 0$, $\dfrac{-z_2 + 4z_1 - 3z_0}{2h} - \alpha_0 z_0 = 0$

$$\Rightarrow \left.\begin{aligned} z_2 &= h^2 p_1 z_1 + 2z_1 - z_0 \\ z_2 &= -2h\alpha_0 z_0 + 4z_1 - 3z_0 \end{aligned}\right\} \Rightarrow \begin{aligned} &h^2 p_1 z_1 + 2z_1 - z_0 = -2h\alpha_0 z_0 + 4z_1 - 3z_0 \\ &(h^2 p_1 - 2) z_1 = (-2 - 2h\alpha_0) z_0 \end{aligned}$$

$$z_1 = \frac{2 + 2h\alpha_0}{2 - p_1 h^2} z_0$$

Let $z_i \not\equiv 0$ of (*). Since, $l(z_i) = 0$ $i = 1,2,\dots,n-1$, then max. positive or min. negative can be only z_0 or z_n.

Let z_0 be max. positive value. Assume that h is so small that $\frac{1}{2} p_i h^2 < 1$. From (3) follows that $z_1 \geq z_0$. Since z_1 can not be greater than z_0, then $z_1 = z_0 \Rightarrow \alpha_0 = 0$ and $p_1 = 0 \Rightarrow$ all z_i are the same, and $z_{n-1} = z_0$. From (2) follows that $\beta_0 = 0$, $p_{n-1} = 0$, but $\alpha_0^2 + \beta_0^2 > 0$.

$$\boxed{z_1 = \frac{1 + \alpha_0 h}{1 - \frac{1}{2} p_1 h^2} z_0} \tag{3}$$

$$\boxed{z_{n-1} = \frac{1 + \beta_0 h}{1 - \frac{1}{2} p_{n-1} h^2} z_0} \tag{4}$$

By contradiction there is no nontrivial solution which takes the maximal positive value on the boundary. Similarly, there is no nontrivial solution which takes the negative value on the boundary. These mean that the system (*) has only trivial solution. So, the system (1'), (2') for any q_i, α_i, β_i has a unique solution.

Elliptic equations

Finite difference method

$a\frac{\partial^2\Phi}{\partial x^2}+b\frac{\partial^2\Phi}{\partial x\partial y}+c\frac{\partial^2\Phi}{\partial y^2}+d\frac{\partial\Phi}{\partial x}+e\frac{\partial\Phi}{\partial y}+f\Phi+g=0$

$b^2-4ac<0$ - elliptic

The best known elliptic equations are Poisson's equation

$$\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}=f(x,y), \quad \left(\nabla^2\Phi=f(x,y)\right) \text{ or } \Delta\Phi=f(x,y) \qquad (1)$$

and Laplace's equation

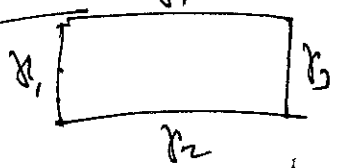
$$\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}=0 \Rightarrow \nabla^2\Phi=0. \quad (\Delta\Phi=0) \qquad (2)$$

1) Dirichlet problem on rectangles

Let R be open rectangle $\{a<x<b, c<y<d\}$ and γ be the boundary of R, $\bar{R}=R\cup\gamma$.

$\frac{a}{b}$ is a rational γ_4

γ_k, $k=1,2,3,4$ be sides of R without the end points



Dirichlet problem:

$$\Delta\Phi=f(x,y) \text{ on } R \qquad (3)$$

$$\Phi=\varphi_k \text{ on } \gamma_k, \ k=1,2,3,4 \qquad (4)$$

It is known that if $\varphi\in C(\gamma)$, $u\in C^4(\bar{R})$, then

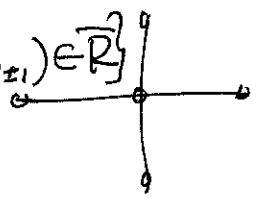
For each (i,j) $x=x_i=ih$, $y=y_j=jh$, $i=1,2,\dots,M$, $j=1,2,\dots,N$

$$\left(\frac{\partial^2\Phi}{\partial x^2}\right)_{(i,j)}\approx\frac{\Phi_{i+1,j}-2\Phi_{ij}+\Phi_{i-1,j}}{h^2}$$

$$\left(\frac{\partial^2\Phi}{\partial y^2}\right)_{(i,j)}\approx\frac{\Phi_{i,j+1}-2\Phi_{ij}+\Phi_{i,j-1}}{h^2}$$

$R_h=\{(x_i,y_j):(x_{i\pm1},y_{j\pm1})\in\bar{R}\}$

$\gamma_h=\bigcup_{k=1}^{4}\gamma_{kh}$



Then

$$\Delta_h\Phi_{ij}=\frac{1}{h^2}\left[\Phi_{i+1,j}+\Phi_{i-1,j}+\Phi_{i,j+1}+\Phi_{i,j-1}-4\Phi_{ij}\right]=f_{ij} \text{ on } R_h \qquad (5)$$

is called a 5-point approximation of (1)

$$\Phi|_{\gamma_{kh}}=\varphi_k, \quad k=1,2,3,4 \qquad (6)$$

A Solution of (5), (6) exists and unique.

Since

Lemma 1. If $W_{ij} \not\equiv \text{Const}$, and

$\Delta_h W_{ij} \geq 0 \quad (\leq 0)$ on R_h, then

positive max. (negative min) all values W_{ij} can be (face)

only

From Lemma follows:

$\Rightarrow$ The homogeneous system

$$\Delta_h \Phi_{ij} = 0 \quad \text{on } R_h,$$
$$\Phi_{ij} = 0 \quad \text{on } \gamma_h$$

has only trivial solution

Analysis of the discretization error.

Differential problem:

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f(x,y), \quad (x,y) \in R$$
$$U(x,y) = g(x,y), \quad (x,y) \in \gamma.$$

5-point approximation (FD) - problem

$$\Delta_h U_{ij} = f_{ij}, \quad (i,j) \in R_h,$$
$$-U_{ij} = \varphi_{ij}, \quad (i,j) \in \gamma_h.$$

Let $e_{ij} = U_{ij} - u_{ij}$ be discretization error at the (i,j)-th mesh point R_h.

Then $u_{ij} = U_{ij} - e_{ij}$,

and

$$\left.\begin{aligned} \Delta_h(U_{ij} - e_{ij}) &= f_{ij} \\ U_{ij} - e_{ij} &= \varphi_{ij} \end{aligned}\right\} \Rightarrow \begin{aligned} \Delta_h e_{ij} &= \Delta_h U_{ij} - f_{ij} \\ e_{ij} &= U_{ij} - \varphi_{ij} = 0 \end{aligned}$$

$$\Rightarrow \quad \Delta_h e_{ij} = T_{ij}, \qquad (7)$$

where T_{ij} is the local truncation error of the difference approximation $\Delta u - f = 0$ at the point (i,j).

It can be shown (show!) that

$$T_{ij} = \frac{1}{12} h^2 \left\{ \left(\frac{\partial^4 u}{\partial x^4} \right)_{(\xi, y_j)} + \left(\frac{\partial^4 u}{\partial y^4} \right)_{(x_i, \eta)} \right\},$$

where $x_i - h < \xi < x_i + h$, $y_j - h < \eta < y_j + h$.

If

$$M_4 = \max \left\{ \max_R \left| \frac{\partial^4 u}{\partial x^4} \right|, \max_R \left| \frac{\partial^4 u}{\partial y^4} \right| \right\},$$

then

$$\max_{R_h} |T_{ij}| \le \frac{1}{6} h^2 M_4$$

From (7), we have

$$\max_{R_h} |\Delta_h e_{ij}| = \max_{R_h} |T_{ij}| \le \frac{1}{6} h^2 M_4 . \qquad (8)$$

Lemma If v is any function defined on the set of mesh points $\overline{R}_h = R_h \cup \gamma_h$ in the rectangular region $0 \le x \le a$, $0 \le y \le b$, then

$$\max_{R_h} |v| \le \max_{\gamma_h} |v| + \frac{1}{4}(a^2 + b^2) \max_{R_h} |\Delta_h v| ,$$

where

$$\Delta_h v_{ij} = \frac{1}{h^2} \left(v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4 v_{ij} \right).$$

Proof.

Define the function Φ_{ij} by the equation

$$\Phi_{ij} = \frac{1}{4}(x_i^2 + y_j^2) = \frac{1}{4}(i^2+j^2)h^2, \quad (i,j) \in R_h.$$

Clearly $$0 \le \Phi_{ij} \le \frac{1}{4}(a^2+b^2) \text{ for all } (i,j) \in R_h. \quad (9)$$

It also follows that $\forall (i,j) \in R_h$,

$$\Delta_h \Phi_{ij} = \frac{1}{4}\{(i+1)^2+j^2+(i-1)^2+j^2+i^2+(j+1)^2+i^2+(j-1)^2-4i^2-4j^2\}$$

$$= \frac{1}{4}\{i^2+2i+1+j^2+i^2-2i+1+j^2+i^2+j^2+2j+1+i^2+j^2-2j+1$$

$$-4i^2-4j^2\} = 1 \quad (10)$$

Now define the functions w^+ and w^- by

$$w^+ = v + N\Phi \text{ and } w^- = -v + N\Phi, \quad (11)$$

where $$N = \max_{R_h} |\Delta_h v_{ij}|.$$

$$\Delta_h w^{\pm} = \pm \Delta_h v + N \quad (i,j) \in R_h$$

$$\Rightarrow \Delta_h w_{ij}^{\pm} \ge 0$$

(If for any function $\Delta_h w \ge 0$ on R_h, then $\max_{R_h} w_{ij} \le \max_{\gamma_h} w_{ij}$)

Then

$$\max_{R_h} w_{ij}^{\pm} \le \max_{\gamma_h} w_{ij}^{\pm} = \max_{\gamma_h} (\pm v_{ij} + N\Phi_{ij}) \quad (\text{by } (11))$$

$$\le \max_{\gamma_h} (\pm v_{ij}) + \frac{1}{4}(a^2+b^2) N \quad (\text{by } (9))$$

Since $w_{ij}^{\pm} = \pm v_{ij} + N\Phi_{ij}$ and $N\Phi_{ij} \ge 0$

$$\Rightarrow \pm v_{ij} \le w_{ij}^{\pm} \text{ for all } (i,j) \in R_h.$$

Hence $$\max_{R_h} (\pm v_{ij}) \le \max_{\gamma_h} (\pm v_{ij}) + \frac{1}{4}(a^2+b^2) N,$$

i.e., $$\max_{R_h} |v_{ij}| \le \max_{\gamma_h} |v_{ij}| + \frac{1}{4}(a^2+b^2) \max_{R_h} |\Delta_h v_{ij}|.$$

Applying this Lemma to the discretization error ℓ_{ij} gives that

$$\max_{R_h} |\ell_{ij}| \le \max_{\gamma_h} |\ell_{ij}| + \frac{1}{4}(a^2+b^2) \max_{R_h} |\Delta_4 \ell_{ij}| \qquad (12)$$

But $\ell_{ij} = v$ on γ_h, because $U_{ij} = u_{ij} = \varphi_{ij}$, $(i,j) \in \gamma_h$.

Then from (12) and (8) we obtain

$$\max_{R_h} |\ell_{ij}| \le \frac{1}{24}(a^2+b^2) h^2 M_4 .$$

$O(h^2)$

The Variational difference methods (the Ritz method and Bubnov-Galerkin methods) and the finite element method (FEM)

Let A be a self-adjoint positive definite linear operator in Hilbert space H equipped with an inner product $(,)$ and let f be a given element of the space H.

The problem of minimizing the functional

$$I[u] = (Au,u) - 2(u,f) \tag{1}$$

is equivalent to the problem of solving the equation

$$Au = f \tag{2}$$

The element $u_0 \in H$ satisfying the equation $Au_0 = f$ and realizing

$$\min I[u] = I[u_0]$$

is unique.

The main idea behind the Ritz method is to take into consideration a sequence of finite-dimensional spaces V_n with basis functions $\varphi_i^{(n)}$, $i=1,2,\dots,n$, and look for an element $u_n \in V_n$, minimizing the functional $I[u]$ in the space V_n.

We take an approximate solution u_n in the form

$$u_n = \sum_{j=1}^{n} y_j \varphi_j \tag{3}$$

with unknown coefficients $y_1, y_2, \dots, y_n$. By inserting this expression in the formula for $I[u]$ we find that

$$I[u_n] = \sum_{i,j=1}^{n} \alpha_{ij} y_i y_j - 2\sum_{j=1}^{n} \beta_j y_j , \tag{4}$$

$$\left(I[u_n] = \left(A\sum_{j=1}^{n} y_j \varphi_j , \sum_{j=1}^{n} y_j \varphi_j\right) - 2\left(\sum_{j=1}^{n} y_j \varphi_j , f\right) \right)$$

where

$$\alpha_{ij}=(A\varphi_i,\varphi_j)\ ,\ \beta_i=(f,\varphi_i)\ . \qquad (5)$$

Since $A=A^*$ is a self-adjoint operator, we have $\alpha_{ij}=\alpha_{ji}$. The functional $I[u_n]$ is a function of n coefficients $y_1, y_2, \dots, y_n$. By equating the derivatives $\frac{\partial I[u_n]}{\partial y_i}$ to zero and using the symmetry of coefficients $\alpha_{ij}=\alpha_{ji}$, we obtain n eq.s for determination of y_i:

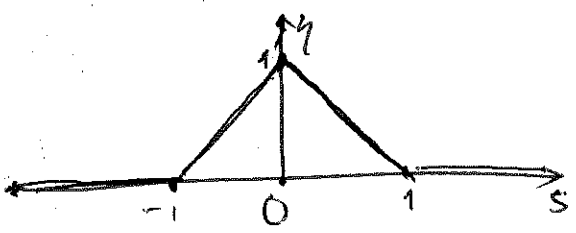
$$\sum_{j=1}^{n}\alpha_{ij}y_j-\beta_i=0\ ,\quad i=1,2,\dots,n\ . \qquad (6)$$

Now let us apply the Ritz method to the problem

$$(ku')'-qu=-f(x)\ ,\quad 0<x<1,\ u(0)=0,\ u(1)=0 \qquad (7)$$

~~We define~~ Let

$$\eta(s)=\begin{cases}0\ , & s<-1,\ s>1\\ 1+s\ , & -1<s<0\\ 1-s\ , & 0<s<1\end{cases} \qquad (7)$$



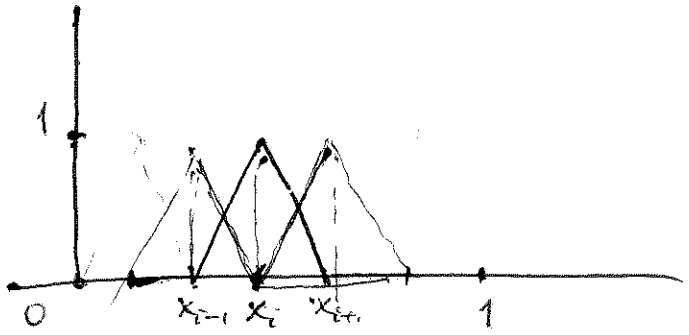
We define

$$\varphi_i(x)=\eta\left(\frac{x-x_i}{h}\right)=\eta_i(x)\ , \qquad (8)$$

where $x_i = ih$, $i = 1, 2, \dots, N-1$, is a node of the grid $\overline{\omega_h} = \{x_i = ih,\ i = 0, 1, \dots, N,\ hN = 1\}$.

From (7) and (8) follows

$$\eta_i(x) = \begin{cases} 0 & \text{for } x < x_{i-1} \text{ and } x > x_{i+1} \\ \dfrac{x - x_{i-1}}{h} & \text{for } x_{i-1} < x < x_i, \\ \dfrac{x_{i+1} - x}{h} & \text{for } x_i < x < x_{i+1}, \end{cases} \qquad (7')$$



and, hence

$$\frac{d\eta_i}{dx} = \begin{cases} 0 & \text{for } x < x_{i-1} \text{ and } x > x_{i+1}, \\ \dfrac{1}{h} & \text{for } x_{i-1} < x < x_i, \\ -\dfrac{1}{h} & \text{for } x_i < x < x_{i+1}. \end{cases} \qquad (9)$$

Upon substituting

$$Au = -\frac{d}{dx}\left(k \frac{du}{dx}\right) + qu = f(x)$$

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

into (5) we get

$$\alpha_{ij} = \int_0^1 \left(k \frac{d\eta_i}{dx} \frac{d\eta_j}{dx} + q\eta_i\eta_j\right) dx, \quad \beta_i = \int_0^1 f(x)\eta_i(x)\,dx$$

$$(u,v)=\int_0^1 uv\,dx$$

$$I[u]=(Au,u)-2(u,f)$$

$$=\int_0^1\left(-\frac{d}{dx}\left(k\frac{du}{dx}\right)+qu\right)\cdot u\,dx-2\int_0^1 uf\,dx$$

$$\int_0^1\left[-\frac{d}{dx}\left(k\frac{du}{dx}\right)\right]u\,dx+\int_0^1 qu^2dx-2\int_0^1 uf\,dx$$

$$=-k\frac{du}{dx}\cdot u\Big|_0^1+\int_0^1\left[k\frac{du}{dx}\cdot\frac{du}{dx}+qu^2\right]dx$$

$$-2\int_0^1 uf(x)dx$$

$$dv=-\frac{d}{dx}\left(k\frac{du}{dx}\right)dx$$

$$v=-k\frac{du}{dx}$$

$$u=u$$

$$du=u'dx$$

$$I[u]=\int_0^1\left[k\left(\frac{du}{dx}\right)^2+qu^2\right]dx-2\int_0^1 uf\,dx$$

$$u_n=\sum_{j=1}^{n}y_j\varphi_j\,,\quad u_n=\sum y_j\varphi_j(x)$$

$$I[y_1,y_2,\dots,y_n]=$$

$$=\int_0^1\left[k\left(\sum_{j=1}^{n}y_j\frac{d\varphi_j}{dx}(x)\right)^2+q\left(\sum_{j=1}^{n}y_j\varphi_j(x)\right)^2\right]dx-2\int_0^1\left(\sum_{j=1}^{n}y_j\varphi_j(x)\right)f\,dx$$

$$\frac{\partial I}{\partial y_i}=0\Rightarrow\int_0^1\left[k\cdot 2\left(\sum_{j=1}^{n}y_j\frac{d\varphi_j}{dx}\right)\frac{d\varphi_i}{dx}+q2\left(\sum_{j=1}^{n}y_j\varphi_j(x)\right)\varphi_i(x)\right]dx$$

$$-2\int_0^1\varphi_i(x)f(x)dx=0$$

$$\sum_{j=1}^{n}\left[\int_0^1 \left(k\frac{d\eta_i}{dx}\frac{d\eta_j}{dx} + q\eta_i(x)\eta_j(x)\right)dx\right]y_j - \int_0^1 f(x)\eta_i(x)\,dx = 0$$

(α_{ij} : the first integral; β_i : the second integral)

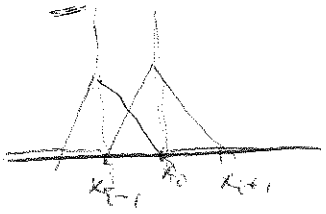
$$\Rightarrow \boxed{\sum_{j=1}^{n}\alpha_{ij}y_j - \beta_i = 0\,, \quad i=1,2,\dots,n} \qquad (10)$$

In light of the properties of the function $\eta_i(x)$ and its derivatives the matrix $\{\alpha_{ij}\}$ is tridiagonal, because only the elements with $j=i-1$, $j=i$ and $j=i+1$ are nonzero. Therefore from (10), we have

$$\alpha_{i,i-1}y_{i-1} + \alpha_{ii}y_i + \alpha_{i,i+1}y_{i+1} = \beta_i\,, \quad i=1,2,\dots,n \qquad (11)$$

where

$$\alpha_{i,i-1} = \int_0^1 \left(k\frac{d\eta_i}{dx}\frac{d\eta_{i-1}}{dx} + q\eta_i(x)\eta_{i-1}(x)\right)dx$$



$$= \int_{x_{i-1}}^{x_i}\left[k(x)\frac{1}{h}\left(-\frac{1}{h}\right) + q(x)\frac{x-x_{i-1}}{h}\cdot\frac{x_i-x}{h}\right]dx$$

$$\alpha_{i,i-1} = -\frac{1}{h^2}\int_{x_{i-1}}^{x_i}k(x)\,dx + \frac{1}{h^2}\int_{x_{i-1}}^{x_i}q(x)(x-x_{i-1})(x_i-x)\,dx$$

$$\alpha_{i,i} = \int_{x_{i-1}}^{x_{i+1}}\left(k(x)\frac{1}{h^2}\right)dx + \frac{1}{h^2}\int_{x_{i-1}}^{x_i}q(x)(x-x_{i-1})^2dx + \frac{1}{h^2}\int_{x_i}^{x_{i+1}}q(x)(x-x_{i+1})^2dx$$

$$\alpha_{i,i} = \frac{1}{h^2}\left[\int_{x_{i-1}}^{x_{i+1}}k(x)dx + \int_{x_{i-1}}^{x_i}q(x)(x-x_{i-1})^2dx + \int_{x_i}^{x_{i+1}}q(x)(x-x_{i+1})^2dx\right]$$

$$\alpha_{i,i+1} = \int_0^1 \left(k(x) \frac{d\eta_i}{dx} \frac{d\eta_{i+1}}{dx} + q(x) \eta_i(x) \eta_{i+1}(x) \right) dx$$

$$= \int_{x_i}^{x_{i+1}} \left[k(x) \left(-\frac{1}{h}\right)\left(\frac{1}{h}\right) + q(x) \frac{(x_{i+1}-x)}{h} \frac{(x-x_i)}{h} \right] dx$$

$$\alpha_{i,i+1} = \alpha_{i+1,i} = \frac{1}{h^2} \left[-\int_{x_i}^{x_{i+1}} k(x)\,dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)(x-x_i)\,dx \right]$$

$$\beta_i = \int_0^1 f(x) \eta_i(x)\,dx = \int_{x_{i-1}}^{x_i} f(x) \frac{x-x_{i-1}}{h}\,dx + \int_{x_i}^{x_{i+1}} f(x) \frac{x_{i+1}-x}{h}\,dx$$

$$\beta_i = \frac{1}{h} \left[\int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1})\,dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1}-x)\,dx \right]$$

We denote by

$$a_i = -h\alpha_{i,i-1}, \quad h^2 d_i = h\alpha_{i,i} + h(\alpha_{i,i-1} + \alpha_{i,i+1})$$

we obtain

$$a_i = \frac{1}{h} \int_{x_{i-1}}^{x_i} k(x)\,dx - \frac{1}{h} \int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})(x_i-x)\,dx$$

$$h^2 d_i = h\alpha_{ii} + h(\alpha_{i,i-1} + \alpha_{i,i+1}) = \frac{1}{h} \int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})^2\,dx + \frac{1}{h} \int_{x_i}^{x_{i+1}} q(x)(x-x_{i+1})^2\,dx$$

$$+ \frac{1}{h} \int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})(x_i-x)\,dx + \frac{1}{h} \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)(x-x_i)\,dx \qquad (12)$$

$$= \int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})\,dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)\,dx$$

$$d_i = \frac{1}{h^2} \left(\int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})\,dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)\,dx \right).$$

Then the system of eq.s

$$\alpha_{i,i-1} y_{i-1} + \alpha_{i,i} y_i + \alpha_{i,i+1} y_{i+1} - \beta_i = 0$$

can be rewritten as

$$a_i (y_{i-1} - (a_i + a_{i+1} + h^2 d_i) y_i + a_{i+1} y_{i+1} + h^2 \varphi_i = 0$$

or

$$(a y_{\bar{x}})_x - d y = -\varphi, \qquad (13')$$

where

$$\varphi_i = \frac{1}{h^2} \left(\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})\,dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x)\,dx \right). \qquad (14')$$

So, the three-point scheme (12')–(14') constructed by the Ritz method is identical with scheme obtained by FIM.

Bubnov-Galerkin Method

In contrast to the Ritz method the Bubnov-Galerkin method applies equally well to problems, where there no fixed sign and non-self-adjoint. The coefficients y_i of the approximate solution

$$u_n = \sum_{j=1}^{n} y_j \varphi_j$$

are to be determined from the orthogonality conditions for the residual $Au_n - f$ with respect to all of the basis functions $\eta_i(x)$:

$$(Au_n - f, \eta_i) = 0, \quad i = 1, 2, \dots, n \qquad (12)$$

We apply this method for a non-self-adjoint boundary-value problem in the form:

$$\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + r(x)\frac{du}{dx} - q(x)u = -f(x), \quad 0 < x < 1$$

$$u(0) = u(1) = 0 \qquad (13)$$

$$k(x) > 0, \quad q(x) \geq 0.$$

We introduce the grid $\overline{\omega}_h = \{x_i = ih, \ i = 0, 1, \dots, N, \ hN = 1\}$. Then the dimension n of the space V_n equals $N-1$. The functions

$$\eta_i(x) = \eta\left(\frac{x - x_i}{h}\right), \quad i = 1, 2, \dots, N-1,$$

where the function $\eta(s)$ was specified by (7).

$$\eta(s) = \begin{cases} 0, & s < -1, \ s > 0 \\ 1+s, & -1 < s < 0, \\ 1-s, & 0 < s < 1 \end{cases}$$

$$\int_0^1 \left[\frac{d}{dx}\left(k(x)\frac{dy}{dx}\right)\eta_i(x) + r(x)\frac{dy}{dx}\cdot\eta_i(x) - q(x)u\,\eta_i(x) + f(x)\eta_i(x) \right] dx = 0$$

$$\int_0^1 \frac{d}{dx}\left(k(x)\frac{dy}{dx}\right)\eta_i(x)\,dx =$$

$$dv = \frac{d}{dx}\left(k\frac{dy}{dx}\right)dx$$

$$= k\frac{dy}{dx}\eta_i(x)\Big|_0^1 - \int_0^1 k(x)\frac{dy}{dx}\frac{d\eta_i}{dx}\,dx$$

$$v = k\frac{dy}{dx}$$

$$u = \eta_i(x)$$

$$du = \eta_i'(x)\,dx$$

$\eta_i(0) = \eta_i(1) = 0$

$$\sum_{j=1}^{N-1} \int_0^1 \left(k(x)\frac{d\eta_j}{dx}\cdot\frac{d\eta_i}{dx}y_j - r(x)\frac{d\eta_j}{dx}\eta_i(x)y_j + q(x)y_j\eta_j(x)\eta_i(x) - f(x)\eta_i(x) \right) dx$$

$$\alpha_{ij}\cdot y_j$$

$$\sum_{j=1}^{N-1} \alpha_{ij}y_j - \beta_i = 0 \quad , \quad i = 1, 2, \dots, N-1$$

?

where

$$\alpha_{ij} = \int_0^1 \left(k(x)\frac{dy_j}{dx}\frac{d\eta_i}{dx} - r(x)\frac{d\eta_j}{dx}\eta_i(x) + q(x)\eta_i(x)\eta_j(x) \right) dx,$$

$$\beta_i = \int_0^1 f(x)\eta_i(x)\,dx \quad , \quad i, j = 1, 2, \dots, N-1 .$$

In this context, condition (12) becomes

$$\sum_{j=1}^{N-1} \alpha_{ij} y_j - \beta_i = 0 \quad , \quad i = 1, 2, \ldots, N-1 , \qquad (14)$$

where

$$\alpha_{ij} = \int_0^1 \left(k(x) \frac{d\eta_i}{dx} \frac{d\eta_j}{dx} - r(x) \frac{d\eta_i}{dx} \eta_j(x) + q(x) \eta_i(x) \eta_j(x) \right) dx , \qquad (15)$$

$$\beta_i = \int_0^1 f(x) \eta_i(x) \, dx \quad , \quad i, j = 1, 2, \ldots, N-1$$

By definition (12) $\eta(s)$ and $\eta_i(x)$ of the function $\eta_i(x)$, the coefficients α_{ij} are nonzero only for $j = i-1, i, i+1$.

At Home!

1) Show that, when $r(x) \equiv 0$ $(B-G) \equiv Ritz$.
2) Derive (obtain) the finite-difference approximation by the $(B-G)$ method, where $k(x)$, $r(x)$ and $q(x)$ are contents.

Using the same notations (12) for a_i and d_i and (14') for φ_i and denoting by

$$b_i^- = \frac{1}{h^2}\int_{x_{i-1}}^{x_i} r(x)(x-x_{i-1})dx =$$

$x = x_i + sh$, $dx = h\,ds$, $x = x_{i-1} + h + sh$, $x - x_{i-1} = h(1+s)$

$$= \frac{1}{h^2}\int_{-1}^{0} r(x_i+sh)(1+s)h\,ds$$

$$b_i^- = \int_{-1}^{0} r(x_i+sh)(1+s)ds$$

$x = x_{i-1} - h + sh$, $x_{i-1} - x = h(1-s)$

$$b_i^+ = \frac{1}{h^2}\int_{x_i}^{x_{i+1}} r(x)(x_{i+1}-x)dx = \int_0^1 r(x_i+sh)(1-s)ds \qquad (16)$$

we reduce the system of eq.s (14) with the members b_i^+ and b_i^- to

$$\frac{1}{h^2}\left[a_{i+1}(y_{i+1}-y_i) - a_i(y_i-y_{i-1})\right] + \frac{b_i^-(y_i-y_{i-1})}{h}$$

$$+ \frac{b_i^+(y_{i+1}-y_i)}{h} - d_i y_i = -\varphi_i, \quad i = 1, 2, \dots, N-1$$

Thus, we arrive at the difference scheme

$$(ay_{\bar{x}})_x + b^- y_{\bar{x}} + b^+ y_x - dy = -\varphi(x), \quad 0 < x = ih < 1,$$

$$y_0 = 0, \; y_N = 0.$$

whose coefficients can be find from (12), (13') and (16)

For $r(x) \equiv 0$ this scheme is identical with scheme obtained by means of the Ritz method. In the case of constant coefficients $k(x)$, $r(x)$ and $q(x)$

$$a_i = k - \frac{h^2}{2}q, \quad d_i = d = q, \quad b_i^- = b_i^+ = \frac{r}{2}, \quad b_i^- y_{\bar{x}} + b_i^+ y_x = r y_{\mathring{x}}.$$

Where the coordinate functions $\varphi_i(x) = \eta\left(\frac{x-x_i}{h}\right)$ are chosen by an approach rule as suggested before, the Ritz and Bubnov-Galerkin methods coincide with the FEM.

The elimination method

The problems we must solve hence now the form

$$A_i y_{i-1} - C_i y_i + B_i y_{i+1} = -F_i, \quad i = 1, 2, \dots, N-1; \tag{1}$$
$$y_0 = \varkappa_1 y_1 + \mu_1, \quad y_N = \varkappa_2 y_{N-1} + \mu_2,$$

where $A_i \neq 0$ and $B_i \neq 0$ for all $i = 1, 2, \dots, N-1$.

Let

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \tag{2}$$

where α_i and β_i unknown parameters.

From (2), we have $y_{i-1} = \alpha_i y_i + \beta_i$ and from (1), we obtain

$$A_i(\alpha_i y_i + \beta_i) - C_i y_i + B_i y_{i+1} = -F_i$$

$$\Rightarrow (A_i \alpha_i - C_i) y_i + A_i \beta_i + B_i y_{i+1} = -F_i$$

which leads, because of (2), to

$$(A_i \alpha_i - C_i)(\alpha_{i+1} y_{i+1} + \beta_{i+1}) + A_i \beta_i + B_i y_{i+1} = -F_i$$

$$[(A_i \alpha_i - C_i)\alpha_{i+1} + B_i] y_{i+1} + A_i \beta_i + (A_i \alpha_i - C_i)\beta_{i+1} = -F_i,$$

If the conditions

$$(A_i \alpha_i - C_i)\alpha_{i+1} + B_i = 0, \quad A_i \beta_i + (A_i \alpha_i - C_i)\beta_{i+1} + F_i = 0$$

are fulfilled simultaneously, then the equation in view holds true for any y_i. Thus, assuming $C_i - \alpha_i A_i \neq 0$, we establish the recurrence formulae for determination of both α_{i+1} and β_{i+1}:

$$\alpha_{i+1} = \frac{B_i}{C_i - \alpha_i A_i}, \quad i = 1, 2, \dots, N-1, \tag{3}$$

$$\beta_{i+1} = \frac{A_i \beta_i + F_i}{C_i - \alpha_i A_i}, \quad i = 1, 2, \dots, N-1, \tag{4}$$

Having substituted $i=0$ into (2), we get

$$y_0 = \alpha_1 y_1 + \beta_1 .$$

On the other hand,

$$y_0 = \varkappa_1 y_1 + \mu_1 ,$$

giving

$$\alpha_1 = \varkappa_1 , \qquad (5)$$

$$\beta_1 = \mu_1 . \qquad (6)$$

Therefore, using (3), (4), (5), (6), we calculate

$$\alpha_1 , \alpha_2 , \dots , \alpha_N ;$$
$$\beta_1 , \beta_2 , \dots , \beta_N .$$

At the second stage, knowing α_i and β_i, the boundary value y_N is recovered from the system of the equations

$$\left.\begin{aligned} y_N &= \varkappa_2 y_{N-1} + \mu_2 , \\ y_{N-1} &= \alpha_N y_N + \beta_N \end{aligned}\right\} \Rightarrow \begin{aligned} y_N &= \varkappa_2 (\alpha_N y_N + \beta_N) + \mu_2 \\ (1 - \alpha_N \varkappa_2) y_N &= \mu_2 + \varkappa_2 \beta_N \end{aligned}$$

When $1 - \alpha_N \varkappa_2 \neq 0$, we have

$$\Rightarrow \quad y_N = \frac{\mu_2 + \varkappa_2 \beta_N}{1 - \alpha_N \varkappa_2} \qquad (7)$$

The computational formulae (2) and (7) constitute what is called the backward elimination path.

Therefore, we have

$$(\rightarrow) \quad \alpha_{i+1} = \frac{B_i}{C_i - \alpha_i A_i} , \quad i = 1, 2, \dots, N-1 , \quad \alpha_1 = \varkappa_1$$

$$(\rightarrow) \quad \beta_{i+1} = \frac{A_i \beta_i + F_i}{C_i - \alpha_i A_i} , \quad i = 1, 2, \dots, N-1 , \quad \beta_1 = \mu_1$$

$$y_N = \frac{\mu_2 + \varkappa_2 \beta_N}{1 - \alpha_N \varkappa_2} ,$$

$$(\leftarrow) \quad y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1} , \quad i = N-1, N-2, \dots, 1, 0 .$$

Now, when will satisfy the assumptions

$$C_i - \alpha_i A_i \neq 0 \quad \text{and} \quad 1 - \alpha_N \varkappa_2 \neq 0 .$$

Sufficient conditions

$$|C_i| \geq |A_i| + |B_i| \quad , \quad i = 1, 2, \ldots N-1$$

$$\varkappa_\alpha \leq 1 \quad , \alpha = 1, 2 \quad , \quad |\varkappa_1| + |\varkappa_2| < 2 ,$$

yielding $|\alpha_i| \leq 1$ for all $i = 1, 2, \ldots, N$.

The proof is carried out by induction.

Assuming $|\alpha_i| \leq 1$ we will show that $|\alpha_{i+1}| \leq 1$.

Since $|\alpha_1| = |\varkappa_1| \leq 1$, for all $i = 2, 3, \ldots, N$, we have

$$|C_i - \alpha_i A_i| - |B_i| \geq |C_i| - |\alpha_i| |A_i| - |B_i| \geq$$

$$\geq |A_i| + |B_i| - |\alpha_i| |A_i| - |B_i| = |A_i| (1 - |\alpha_i|) \geq 0$$

from which it follows that $|C_i - \alpha_i A_i| > 0$, hence $B_i \neq 0$.

$\Rightarrow$ when $|\alpha_i| \leq 1$,

$$|\alpha_{i+1}| = \frac{|B_i|}{|C_i - A_i \alpha_i|} \leq 1$$

Also $|\alpha_{i+1}| < 1$ when $|\alpha_i| < 1$.

The assumption $|\alpha_1| = |\varkappa_1| < 1$ implies $|\alpha_i| < 1$ for all $i = 1, 2, \ldots, N$.

For the denominator in formula (7):

$$|1 - \alpha_N \varkappa_2| \geq 1 - |\alpha_N| |\varkappa_2| \geq 1 - |\varkappa_2| > 0$$

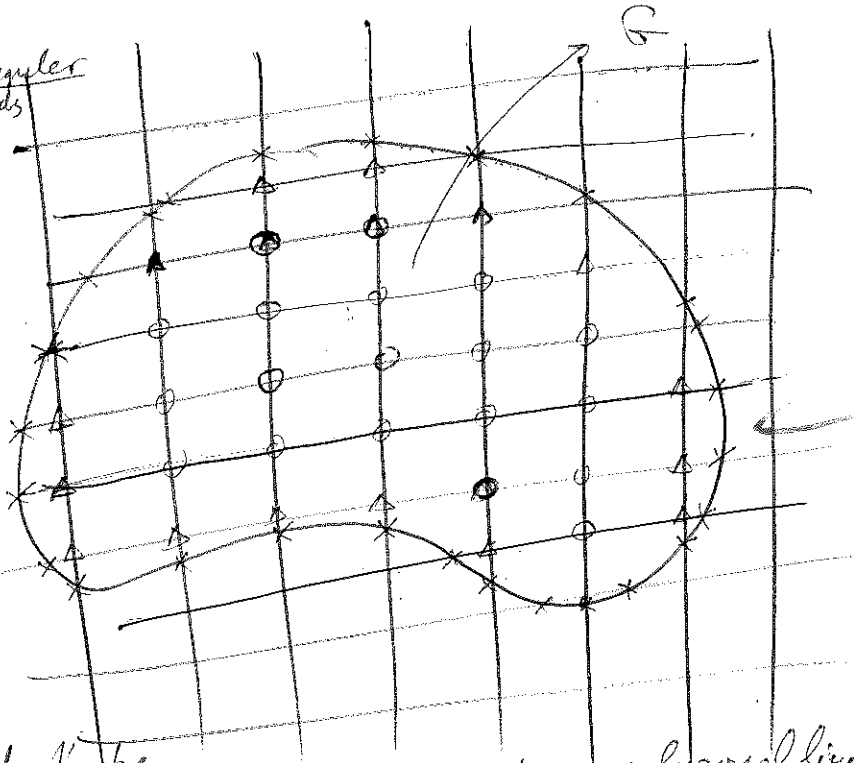
because either $|\varkappa_2| < 1$ or $|\alpha_N| < 1$.

The Dirichlet difference problem in a domain of rather complicated configuration.

set of
○ - nodes
are called strictly inner nodes or regular nods
we denote set of such points as $\overset{\circ}{\omega}_h$

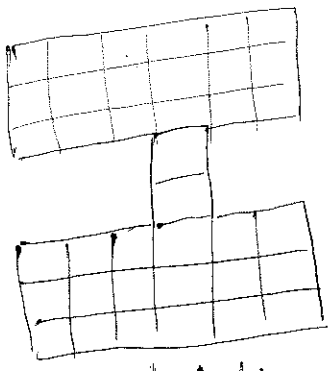
▲ - irregular nodes
if at least one of 4 neighbor nodes is outside of $\overline{G}$
The set of such nodes we denote by ω^*

× intersection of the grid lines with the boundary of G
we denote these points as γ_h

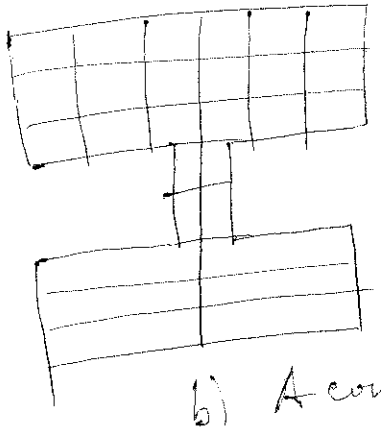


Then
$$\overline{\omega}_h = \overset{\circ}{\omega} + \omega^* + \gamma_h$$

The grid $\overline{\omega}_h$ is said to be connected if any two inner nodes can be joined by a polygonal line, the ports of which are parallel to the coordinate axes and vertices coincide with inner nodes of the grid. Then at least one of the four nodes $x^{(\pm\alpha)}$, $\alpha = 1,2$, of the five-point pattern $(x^{(\pm 1_1)}, x, x^{(\pm 1_2)})$ (regular or irregular) falls within the collection of inner nodes



a) A disconnected grid



b) A connected grid

The procedure of constructing a grid in the plane domain we have described above can easily be generalized to the case of an arbitrary P-dimentional domain. A grid to ownstructed is a result of the intersection of hyperplanes (planes for $P=3$ or straight lines for $P=2$)

$$x_\alpha^{i_\alpha} = i_\alpha h_\alpha, \quad i_\alpha = 0, \pm 1, \dots, \quad \alpha = 1, 2, \dots, P,$$

where $h_\alpha > 0$. The preceding classification of nodes remains unchanged here.

Our purpose here is to construct a difference scheme for solving the Dirichlet problem in the domain $\overline{G} = G + \Gamma$, the complete posing of which is to find an unknown solution to the equation

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in G,$$

which is continuous in the closed domain $\overline{G} = G + \Gamma$ and satisfies the b. c. $u|_\Gamma = \mu(x)$.

At each of the inner nodes $x \in \omega_h$ we approximate the differential operator

$$L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2}$$

by the three-point difference operator Λ_α

If a node $x \in \omega_h$ is regular with respect to x_α, then the difference operator Λ_α on the regular pattern $(x^{(-h_\alpha)}, x, x^{(+h_\alpha)})$ is

$$\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha} = \frac{y^{(+h_\alpha)} - 2y + y^{(-h_\alpha)}}{h_\alpha^2} \qquad (1)$$

But if a node $x \in \omega_{h,\alpha}^*$, that is, a node is irregular with respect to x_α on the irregular pattern, the difference operator Λ_α can be rewritten as

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+h_\alpha)} - y}{h_\alpha} - \frac{y - y^{(-h_\alpha)}}{h_\alpha^*}\right) \text{ for } x^{(-h_\alpha)} \in \gamma_{h,\alpha}, \qquad (2)$$

where h_α^* is the distance between the nodes x and $x^{(-k)}$, or

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+k)}-y}{h_\alpha^*} - \frac{y-y^{(-k)}}{h_\alpha}\right) \quad \text{for } x^{(+\frac{1}{2}k)} \in \gamma_{h,\alpha}, \tag{3}$$

where h_α^* is the distance between the nodes x and $x^{(+k)}$.

If $x^{(-k)} \in \gamma_{h,\alpha}$ and $x^{(+k)} \in \gamma_{h,\alpha}$, then

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+k)}-y}{h_{\alpha+}^*} - \frac{y-y^{(-k)}}{h_{\alpha-}^*}\right) \quad \text{for } x^{(\pm k)} \in \gamma_{h,\alpha} \tag{4}$$

where $h_{\alpha\pm}^* \neq h_\alpha$ is the distance between x and $x^{(\pm k)}$.

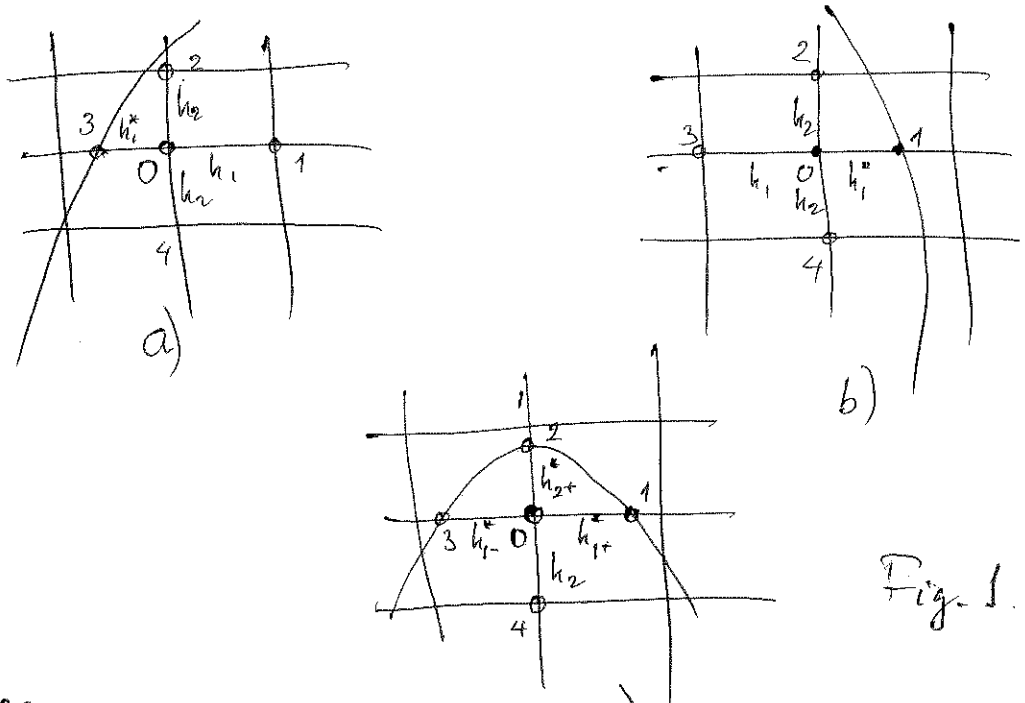


Fig. 1

Finally, we arrive at the Dirichlet difference problem of determining a grid function $y(x)$ defined for $x \in \bar{\omega}_h = \omega_h + \gamma_h$, satisfying at the inner nodes the equation

$$\begin{aligned} &\Lambda y + \varphi(x) = 0 && \text{at the regular nodes,} \\ &\Lambda^* y + \varphi(x) = 0 && \text{at the irregular nodes} \\ &y = \mu(x), && x \in \gamma_h. \end{aligned} \tag{5}$$

where ψ is the approximation error equal for $\varphi(x) = f(x)$ to

a) $\Lambda_1^* y = \frac{1}{h_1}\left(\frac{y_1 - y_0}{h_1} - \frac{y_0 - y_3}{h_1^*}\right)$, $\Lambda_2 y = y_{\bar{x}_2 x_2}$, $\Lambda^* = \Lambda_1^* + \Lambda_2$.

b) $\Lambda_1^* y = \frac{1}{h_1}\left(\frac{y_1 - y_0}{h_1^*} - \frac{y_0 - y_3}{h_1}\right)$, $\Lambda_2 y = y_{\bar{x}_2 x_2}$, $\Lambda^* = \Lambda_1^* + \Lambda_2$

c) $\Lambda_1^* y = \frac{1}{h_1}\left(\frac{y_1 - y_0}{h_{1+}^*} - \frac{y_0 - y_3}{h_{1-}^*}\right)$, $\Lambda_2 y = \frac{1}{h_2}\left(\frac{y_2 - y_0}{h_2^*} - \frac{y_0 - y_4}{h_2}\right)$,

$\Lambda^* y = \Lambda_1^* + \Lambda_2^*$,

$\psi = \Lambda u + \varphi = \Lambda u - Lu$ at the regular nodes,

$\psi^* = \Lambda^* u - Lu$ at the irregular nodes. (6)

Let $u \in C^{(4)}(\bar{G})$, ~~there $C^{(4)}$ is the~~ As stated ~~before~~ we have

$$|\psi| \le M_4 \frac{|h|^2}{12}, \quad |h|^2 = h_1^2 + h_2^2 + \dots + h_p^2. \quad (7)$$

at the irregular nodes

$$\psi^* = \sum_{\alpha=1}^{p} \psi_\alpha^*, \quad \psi_\alpha^* = \Lambda_\alpha u - L_\alpha u,$$

$$\psi_\alpha^* = \frac{h_{\alpha+} \pm h_{\alpha-}}{2h_\alpha} \frac{\partial^2 u}{\partial x_\alpha^2} + O(h_\alpha) = O(1), \quad \psi^* = O(1),$$

meaning that at the irregular nodes the scheme does not approximate the equation $\Delta u + f(x) = 0$.

Thus, in the p-dimensional case a difference scheme such as

$$\Lambda y = \sum_{\alpha=1}^{p} \Lambda_\alpha y = -f(x) \text{ at the regular nodes,}$$

$$\Lambda^* y = \sum_{\alpha=1}^{p} \Lambda_\alpha^* y = -f(x) \text{ at the irregular nodes,}$$

where $\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}$ and Λ_α^* is specified by the formula

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+1_\alpha)} - y}{h_{\alpha+}} - \frac{y - y^{(-1_\alpha)}}{h_{\alpha-}}\right),$$

is associated with problem ~~(4)~~

$$\Delta u = \sum_{\alpha=1}^{p} \frac{\partial^2 u}{\partial x_\alpha^2} = -f(x), \quad x \in G,$$

$$u|_\Gamma = \mu(x)$$

where $x = (x_1, x_2, \ldots, x_p)$, G is a p-dimensional finite domain with the boundary Γ.

Remark. Quite often, the Dirichlet problem is approximated by the method based on the difference approximation of the near-boundary nodes of the Laplace operator on an irregular pattern, with the use of formula with $\frac{1}{h_\alpha}(\quad)$ instead of $\frac{1}{h_\alpha}(\quad)$. However, in this case ~~several~~ the difference operator important properties does not possess, such as the self-adjointness and the property of having fixed sign.

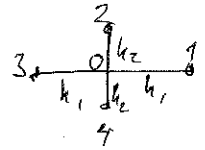
5. The canonical form of a difference equation

We now consider the $(2p+1)$-point scheme $\Lambda y = -f$ at a regular node

$$\sum_{\alpha=1}^{p} \frac{1}{h_\alpha^2}\left(y^{(+1_\alpha)} - 2y + y^{(-1_\alpha)}\right) = -f,$$

which admits an alternative form of writing

$$\sum_{\alpha=1}^{p} \frac{2}{h_\alpha^2} y(x) = \sum_{\alpha=1}^{p} \frac{1}{h_\alpha^2}\left(y^{(+1_\alpha)} + y^{(-1_\alpha)}\right) + f(x).$$



in two-dimensional case

$$2\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right) y_0 = \frac{1}{h_1^2}(y_1 + y_3) + \frac{1}{h_2^2}(y_2 + y_4) + f_0$$

Let $x \in \omega_{h,1}^*$ be an irregular node. (in Fig 1 a)

$$\Lambda_1^* y_0 = \frac{1}{h_1}\left(\frac{y_1 - y_0}{h_1} - \frac{y_0 - y_3}{h_1^*}\right) = \frac{1}{h_1}\left(\frac{y_1}{h_1} + \frac{y_3}{h_1^*} - \frac{2h_1}{h_1 h_1^*} y_0\right),$$

where $h_1 = \frac{1}{2}(h_1 + h_1^*)$, and

$$\Lambda_2 y_0 = \frac{1}{h_2^2}(y_2 - 2y_0 + y_4).$$

From the eq. $\Lambda^* y = \Lambda_1^* y + \Lambda_2 y = -f$ we find that

$$\left(\frac{2h_1}{h_1^2 h_1^*} + \frac{2}{h_2^2}\right) y_0 = \frac{1}{h_1^2} y_1 + \frac{1}{h_1 h_1^*} y_3 + \frac{1}{h_2^2}(y_2 + y_4) + f_0.$$

In the case Fig 1, c), we deduce that

$$\left(\frac{2h_1}{h_1 h_{1-}^* h_{1+}^*} + \frac{2h_2}{h_2^2 h_2^*}\right) y_0 = \frac{1}{h_1 h_{1+}^*} y_1 + \frac{1}{h_1 h_{1-}^*} y_3 + \frac{1}{h_2 h_2^*} y_2 + \frac{1}{h_2^2} y_4 + f_0,$$

where $h_1 = \frac{1}{2}(h_{1-}^* + h_{1+}^*)$ and $h_2 = \frac{1}{2}(h_2 + h_2^*)$.

So all these eq.s can be represented in the canonical form

$$A(x) y(x) = \sum_{\xi \in \mathrm{Patt}'(x)} B(x,\xi)\, y(\xi) + F(x), \quad x \in \omega_h \qquad (8)$$

where $\mathrm{Patt}'(x)$ is the set consisting of $2p$ nodes of the $(2p+1)$-point "cross" pattern with center at the point x except for the node x itself, that is $\xi \neq x$.

We call the set $\mathrm{Patt}'(x)$ the neighborhood of the node x.

It is easily seen that

$$A(x) > 0, \quad B(x,\xi) > 0, \quad \sum_{\xi \in \mathrm{Patt}'(x)} B(x,\xi) = A(x) \quad \text{for all } x \in \omega_h. \qquad (9)$$

Eq. (8) is put together with the boundary conditions

$$y|_{\gamma_h} = \mu(x) . \qquad (10)$$

The Dirichlet difference problem is a special case of a more general problem in which it is required to find a grid function $y(x)$ defined on the grid $\overline{\omega}_h = \omega_h + \gamma_h$ and satisfying on ω_h the equation

$$A(x)y(x) = \sum_{\xi \in Patt'(x)} B(x,\xi)y(\xi) + F(x), \quad x \in \omega_h, \qquad (11)$$

$$y(x) = \mu(x), \quad x \in \gamma_h,$$

where

$$A(x) > 0, \quad B(x,\xi) > 0, \quad D(x) = A(x) - \sum_{\xi \in Patt'} B(x,\xi) \geq 0 \qquad (12)$$

for all $x \in \omega_h$

To prove the existence and uniqueness of a solution of problem (11)-(12), it suffices to check that the homogeneous eq.

$$\mathcal{L}[y] = A(x)y(x) - \sum_{\xi \in Patt'(x)} B(x,\xi)y(\xi) = 0, \quad x \in \omega_h, \qquad (13)$$

$$y(x) = 0, \quad x \in \gamma_h$$

has only the trivial solution $y(x) = 0$ for $x \in \overline{\omega}_h$.

The Maximum Principle

The maximum principle is suitable of difference elliptic and parabolic eq.s in the space C and is certainly true for grid eq.s of common structure, ~~which will be~~

Let ω be a finite set of nodes (a grid) in some bounded domain of the n-dimensional Euclidean space and let $P \in \omega$ a point of the grid ω.

Consider the eq.

$$A(P)y(P) = \sum_{Q \in Patt'(P)} B(P,Q)y(Q) + F(P), \quad P \in \omega. \tag{1}$$

related to a function $y(P)$ defined on the grid ω. Here the coefficients of the equation $A(P)$ and $B(P,Q)$ and the right-hand side of the eq. $F(P)$ are given grid functions; $Patt'(P) \subset \omega$ being the set of all the nodes of the grid ω except for the node P, is the neighborhood of the node P. The pattern of the grid eq. (1) at the node P consists, evidently, of the node P itself and its neighborhood $Patt'(P)$.

In what follows we will suppose that coefficients $A(P)$ and $B(P,Q)$ are subject to the conditions

$$A(P) > 0, \quad B(P,Q) > 0 \text{ for all } P \in \omega, \ Q \in Patt'(P), \tag{2}$$

$$D(P) = A(P) - \sum_{Q \in Patt'(P)} B(P,Q) \geq 0.$$

A point P is called a boundary node of the grid ω if at this point the value of the function $y(P)$ is known in advance:

$$y(P) = \mu(P) \text{ for } P \in \gamma, \tag{3}$$

where γ is the set of all boundary nodes.

Comparing (3) and (1) we see that on the boundary γ we must set formally $A(P) \equiv 1$, $B(P,Q) \equiv 0$ and $F(P) = \mu(P)$.

The grid $\bar{\omega}$ is taken to be connected, that is, for fixed points $\bar{P} \in \bar{\omega}$ and $\bar{\bar{P}} \in \bar{\omega}$ there always exists a sequence of neighborhoods $\{Patt'(P)\}$ such that the passage from $\bar{P}$ to $\bar{\bar{P}}$ can be done using only the nodes of those neighborhoods or, in other words, one can select nodes $P_1, P_2, \ldots, P_m$ of the grid ω such that

$P_1 \in Patt'(\bar{P})$, $P_2 \in Patt'(P_1)$, ..., $P_m \in Patt'(P_{m-1})$, $\bar{\bar{P}} \in Patt'(P_m)$

with

$B(P_i, P_{i+1}) \neq 0$, $i = 1, 2, \ldots, m-1$,
$B(\bar{P}, P_1) \neq 0$, $B(P_m, \bar{\bar{P}}) \neq 0$.

(4)

By using the notation

$$\mathcal{L}y(P) = A(P)y(P) - \sum_{Q \in Patt'(P)} B(P,Q)\,y(Q), \qquad (5)$$

the eq. (1) can be written in the form

$$\mathcal{L}y(P) = F(P), \qquad (6)$$

An alternative form of $\mathcal{L}y(P)$ may be useful in the further development:

$$\mathcal{L}y(P) = D(P)\,y(P) + \sum_{Q \in Patt'(P)} B(P,Q)\,(y(P) - y(Q)). \qquad (7)$$

Consider as one possible example the so-called scheme with weights for the heat conduction equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \quad 0 < x < 1, \ t > 0,$$

$$u(x,0) = u_0(x), \quad u(0,t) = \mu_1(t), \quad u(1,t) = \mu_2(t)$$

On the grid $\overline{\omega}_{h\tau} = \{(x_i = ih, t_j = j\tau), \ i = 0,1,\dots,N, \ hN = 1, \ j = 0,1,\dots\}$

this scheme takes the form

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \Lambda\left(\sigma y_i^{j+1} + (1-\sigma)y_i^j\right) + \varphi_i^j, \qquad (8)$$

$$\Lambda y = y_{\bar{x}x}, \quad y_i^0 = u_0(x_i), \quad y_0^j = \mu_1(t_j), \quad y_N^j = \mu_2(t_j).$$

$$\Rightarrow \left(\frac{1}{\tau} + \frac{2\sigma}{h^2}\right) y_i^{j+1} = \frac{\sigma}{h^2}\left(y_{i-1}^{j+1} + y_{i+1}^{j+1}\right) + \left(\frac{1}{\tau} + \frac{2(\sigma-1)}{h^2}\right) y_i^j + \frac{1-\sigma}{h^2}\left(y_{i-1}^j + y_{i+1}^j\right) + \varphi_i^j.$$

From here it is easily seen that $B(P,Q) \geq 0$ only if $\tau \leq \frac{h^2}{2(1-\sigma)}$ and $0 \leq \sigma \leq 1$. $D(P) = 0$.

The maximum principle

Theorem 1 Let $y(P) \not\equiv \text{const}$ be a grid function defined on a connected grid $\bar{\omega}$ and let both conditions (2) and (4) hold. Then the condition $\mathcal{L}y(P) \le 0$ $(\mathcal{L}y(P) \ge 0)$ on the grid ω_h implies that $y(P)$ cannot attain the maximal positive (minimal negative) value at the inner nodes $P \in \omega_h$.

Proof. Let $\mathcal{L}y(P) \le 0$ at all of the inner nodes $P \in \omega_h$. On the contrary, let $y(P)$ attain its maximal positive value at an inner node $\bar{P} \in \omega$, so that

$$y(\bar{P}) = \max_{\omega_h} y(P) = M_0 > 0.$$

Since $y(\bar{P}) \ge y(Q)$ for all $Q \in Patt'(\bar{P})$, we find by virtue of the relations $D(\bar{P}) \ge 0$ and $y(\bar{P}) > 0$ that

$$\mathcal{L}y(\bar{P}) = D(\bar{P})\, y(\bar{P}) + \sum_{Q \in Patt'(\bar{P})} B(\bar{P}, Q)\,(y(\bar{P}) - y(Q)) \ge D(\bar{P})\, y(\bar{P}) \ge 0.$$

Then only the case $\mathcal{L}y(\bar{P}) = 0$ needs investigation.

$\Leftrightarrow D(\bar{P}) = 0$ and $y(Q) = y(\bar{P})$ for all $Q \in Patt'(\bar{P})$.

We now face the node $P_1 \in Patt'(\bar{P})$ at which $y(P_1) = y(\bar{P}) = M_0$ and adopt these ideas. Since $y(P) \not\equiv \text{const}$ on the grid ω and the grid is connected, there exists a sequence of nodes $P_1, P_2, \dots, P_m, \bar{\bar{P}}$, satisfying conditions (4), such that

$$y(P_m) = y(\bar{P}) = M_0, \quad y(\bar{\bar{P}}) < M_0,$$

but

$$\bar{\bar{P}} \in Patt'(P_m).$$

Then

$$\mathcal{L}y(P_m) \ge D(P_m)\, y(P_m) + B(P_m, \bar{\bar{P}})\,(y(P_m) - y(\bar{\bar{P}})) \ge$$

$$\ge B(P_m, \bar{\bar{P}})\,(y(\bar{P}) - y(\bar{\bar{P}})) > 0$$

meaning $\bar{P} = P_m$ and justifying the first assertion of the theorem.

The second assertion will be reduced to the first one once we replace $y(P)$ by $-y(P)$.

Corollary 1. Let conditions (2) and (4) hold and let a grid function $y(P)$ defined on $\omega+\gamma$ be nonnegative on the boundary, that is, $y(P)\geq 0$ for $P\in\gamma$ and $\mathcal{L}y(P)\geq 0$ on ω. Then $y(P)$ is nonnegative on $\omega+\gamma$, that is $y(P)\geq 0$ for all $P\in\omega+\gamma$. But if $y(P)\leq 0$ on γ and $\mathcal{L}y(P)\leq 0$ on ω, then $y(P)\leq 0$ on $\omega+\gamma$.

Proof. . .

Corollary 2. The homogeneous eq. (1) subject to the boundary condition

$$\mathcal{L}y(P)=0 \text{ on } \omega, \quad y(P)=0 \text{ on } \gamma, \qquad (9)$$

has only the trivial solution $y(P)\equiv 0$.

Proof.

Corollary 3. Problem (1)–(4) possesses a unique solution.

2 Comparison theorem. The majorant.

Theorem. Let $y(P)$ be a solution of problem (1)–(4) and let $Y(P)$ be a solution of the problem

$$\mathcal{L}Y(P)=\bar{F}(P), \quad P\in\omega, \quad Y(P)=\bar{\mu}(P), \quad P\in\gamma. \qquad (10)$$

Then the conditions

$$|F(P)|\leq\bar{F}(P), \quad P\in\omega, \quad |\mu(P)|\leq\bar{\mu}(P), \quad P\in\gamma, \qquad (11)$$

provide the validity of the inequality

$$|y(P)|\leq Y(P), \text{ for } P\in\omega+\gamma. \qquad (12)$$

Proof. By Corollary 1 the inequality $\mathcal{Y}(P) \geq 0$ is valid on $\omega+\gamma$. The functions $u(P)=\mathcal{Y}(P)+y(P)$ and $v(P)=\mathcal{Y}(P)-y(P)$ solve the equations $\mathcal{L}u=F_u=\bar{F}+F\geq 0$ and $\mathcal{L}v=F_v=\bar{F}-F\geq 0$ subject to the boundary conditions $u|_\gamma=(\mathcal{Y}+y)|_\gamma=\bar{\mu}+\mu\geq 0$ and

$$v|_\gamma=(\mathcal{Y}-y)|_\gamma=\bar{\mu}-\mu\geq 0.$$

Since the conditions of Corollary 1 are satisfied, we have $u\geq 0$ or $y\geq -\mathcal{Y}$, $v\geq 0$ or $y\leq \mathcal{Y}$. It follows from the foregoing that $-\mathcal{Y}\leq y\leq \mathcal{Y}$ or $|y(P)|\leq \mathcal{Y}(P)$ on $\omega+\gamma$. The function $\mathcal{Y}(P)$ is called the <u>majorant</u> for a solution of problem (1)–(3).

<u>Corollary</u> For a solution of the problem

$$\mathcal{L}(y)=0 \text{ on } \omega, \quad y(P)=\mu(P) \text{ on } \gamma, \tag{13}$$

the estimate

$$\max_{P\in\omega+\gamma}|y(P)| = \|y\|_C \leq \|\mu\|_{C_\gamma}, \tag{14}$$

is valid with $\|\mu\|_{C_\gamma}=\max_{P\in\gamma}|\mu(P)|$.

Indeed, let us specify the majorant $\mathcal{Y}(P)$ by the conditions $\mathcal{L}\mathcal{Y}=0$ on ω and $\mathcal{Y}=\|\mu\|_{C_\gamma}$ on γ. The function $\mathcal{Y}(P)$ is nonnegative on $\omega+\gamma$ and attains its maximum at some node of the grid. This node is none the inner nodes if $\mathcal{Y}(P)\neq$ Const and, hence,

$$\|\mathcal{Y}\|_C=\max_{P\in\omega+\gamma}\mathcal{Y}(P)=\max_{P\in\gamma}\mathcal{Y}(P)=\|\mu\|_{C_\gamma}.$$

If $\mathcal{Y}(P)=$ Const, then $\mathcal{Y}(P)=\|\mu\|_{C_\gamma}$. In both cases $\|\mathcal{Y}\|_C=\|\mu\|_{C_\gamma}$. Combination of this relation and the inequality $\|y\|_C\leq\|\mathcal{Y}\|_C$ gives estimate (14).

The estimate of a solution to the nonhomogeneous equation.

In the further development a solution of problem (1)–(3) is viewed as a sum

$$y = \bar{y} + v,$$

where $\bar{y} = \bar{y}(P)$ is a solution to the homogeneous equation

$$\mathcal{L}\bar{y} = 0 \text{ on } \omega, \quad \bar{y} = \mu(P) \text{ on } \gamma, \tag{15}$$

and $v = v(P)$ is a solution to the nonhomogeneous equation

$$\mathcal{L}v(P) = F(P) \text{ on } \omega, \quad v(P) = 0 \text{ on } \gamma. \tag{16}$$

We have already obtained estimate (14) for $\bar{y}(P)$ and so it remains to evaluate (estimate) the function $v(P)$.

Theorem 3. If $D(P) > 0$ everywhere on the grid ω, then a solution of problem (16) admits the estimate

$$\|v\|_C \leq \left\| \frac{F}{D} \right\|_C. \tag{17}$$

Proof. Let a majorant $Y(P)$ be taken such that

$$\mathcal{L}Y = |F(P)|, \quad Y|_\gamma = 0, \quad Y(P) \geqslant 0 \text{ for } P \in \omega + \gamma$$

and $Y(P)$ attain its maximum at a node $P_0 \in \omega$. As far as $Y(P_0) = \|Y\|_C$ is concerned, the equation

$$D(P_0)Y(P_0) + \sum_{Q \in Patt'(P_0)} B(P_0, Q)(Y(P_0) - Y(Q)) = |F(P_0)|$$

$$\left\{ \begin{pmatrix} D(P_0) \\ A(P_0) - \sum_{Q \in Patt'(P_0)} B(P_0, Q) \end{pmatrix} Y(P_0) + \sum_{Q \in Patt'(P_0)} B(P_0, Q)(Y(P_0) - Y(Q)) \right\}$$

yields $D(P_0)Y(P_0) \leq |F(P_0)|$, $Y(P_0) \leqslant \dfrac{|F(P_0)|}{D(P_0)} \leqslant \left\| \dfrac{F}{D} \right\|_C$

thereby completing the proof.

Remark Estimate (17) is still valid for the solution of problem (16) provided that instead of (2) other conditions

$$|A(P)| \neq 0\,, \quad |B(P,Q)| \neq 0,$$

$$\mathcal{D}(P) = |A(P)| - \sum_{Q \in Patt'(P)} |B(P,Q)| > 0$$

hold.

Indeed, let $|V(P)| \geq 0$ take the maximal value at a node P_0. Because of this fact,

$$|A(P_0)| \cdot |V(P_0)| = \Big| \sum_{Q \in Patt'(P_0)} B(P_0, Q) V(Q) + F(P_0) \Big|$$

$$\leq \sum_{Q \in Patt'(P_0)} |B(P_0, Q)| \cdot |V(P_0)| + |F(P_0)|,$$

whence it follows that

$$\mathcal{D}(P_0)|V(P_0)| \leq |F(P_0)|, \quad \|V\|_C = |V(P_0)| \leq \frac{|F(P_0)|}{\mathcal{D}(P_0)} \leq \left\| \frac{F}{\mathcal{D}} \right\|_C .$$

It may happen that $\mathcal{D}(P) = 0$ on a subset $\mathring{\omega}$ of the grid ω and $\mathcal{D}(P) > 0$ on the complement of $\mathring{\omega}$ to ω: $\mathring{\omega} + \omega^* = \omega$. This type of situation is covered by the following assertion.

Theorem 4 Let the conditions

$\mathcal{D}(P) = 0$ for $P \in \mathring{\omega}$, $\mathcal{D}(P) > 0$ for $P \in \omega^*$

hold, where $\mathring{\omega}$ is a connected grid. Then for a solution of problem (16) with the right-hand side

$F(P) = 0$ for $P \in \mathring{\omega}$, $F(P) \not\equiv 0$ for $P \in \omega^*$,

the estimate

$$\|V\|_C \leq \left\| \frac{F}{\mathcal{D}} \right\|_{C^*} \qquad (18)$$

is valid in the norm $\|f\|_{C^*} = \max_{P \in \omega^*} |f(P)|$.

Proof. Let $\mathcal{Y}(P)$ be a majorant and $\mathcal{L}\mathcal{Y} = |F(P)|$ on the grid ω, $\mathcal{Y}|_\gamma = 0$, $\mathcal{Y} \geq 0$. The function $\mathcal{Y}(P)$ should attain its maximum on a finite set $\omega + \gamma$ at some node, not belonging to the boundary, because $\mathcal{Y}|_\gamma = 0$. Also, it does not enter the grid $\mathring{\omega}$ due to the connectedness of $\mathring{\omega}$ and the maximum principle. Hence,

$$\max_{P \in \omega} \mathcal{Y}(P) = \max_{P \in \omega^*} \mathcal{Y}(P) = \mathcal{Y}(P_0),$$

where P_0 is a node on the set ω^*.

By the initial hypothesis, $\mathcal{D}(P_0) > 0$. Arguing as in the proof of T. 3 we arrive at (18). An analog of the remark to T.3 is still valid for that case.

Stability and convergence of the Dirichlet difference problem

1. Estimation of a solution of the Dirichlet difference problem.

We construct a uniform estimate of a solution of the Dirichlet difference problem

$$\begin{aligned} \Lambda y &= -\varphi \quad \text{at the regular nodes} \\ \Lambda^* y &= -\varphi \quad \text{at the irregular nodes.} \\ y &= \mu(x) \quad \text{on the boundary,} \end{aligned} \tag{1}$$

where

$$\Lambda y = \sum_{\alpha=1}^{p} \Lambda_\alpha y, \quad \Lambda_\alpha y = y_{\overline{x}_\alpha x_\alpha},$$

$$\Lambda^* y = \sum_{\alpha=1}^{p} \Lambda_\alpha^* y, \quad \Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+\alpha)} - y}{h_{\alpha+}} - \frac{y - y^{(-\alpha)}}{h_{\alpha-}}\right).$$

The problem (1) can be written

$$\bar{\Lambda} y = -\varphi, \quad x \in \omega_h, \quad y|_{\gamma_h} = \mu(x), \tag{2}$$

where $\bar{\Lambda}$ coincides with $\mathcal{D}\Lambda^*$ at the near-boundary nodes, and with Λ at the remaining inner nodes.

The problem (1) also can be written as

$$A(x)y(x) = \sum_{\xi \in Patt(x)} B(x,\xi)y(\xi) + F(x), \quad x \in \omega, \quad y|_{\gamma_h} = \mu(x), \quad (3)$$

where

$$A(x) > 0, \quad B(x,\xi) > 0, \quad D(x) = A(x) - \sum_{\xi \in Patt(x)} B(x,\xi) \geq 0.$$

We now represent a solution of problem (1) as a sum

$$y = \bar{y} + \tilde{y},$$

where $\bar{y}$ and $\tilde{y}$ are, respectively, solutions of the appropriate problems

$$\Lambda \bar{y} = 0, \quad x \in \omega_h, \quad \bar{y} = \mu \text{ for } x \in \gamma_h, \quad (4)$$

$$\Lambda \tilde{y} = -\varphi, \quad x \in \omega_h, \quad \tilde{y} = 0 \text{ for } x \in \gamma_h. \quad (5)$$

An estimate for a solution of problem (4) such as

$$\|\bar{y}\|_C \leq \|\mu\|_{C_\gamma}, \quad (6)$$

has been derived before.

Having decomposition the right-hand side φ as

$$\varphi = \mathring{\varphi} + \varphi^*,$$

where $\mathring{\varphi} = \varphi$ and $\varphi^* = 0$ at the strictly inner nodes $x \in \omega_h^0$ and $\mathring{\varphi} = 0$ and $\varphi^* = \varphi$ at the near-boundary nodes $x \in \omega_h^*$, we consider

$$\tilde{y} = v + w$$

with v and w being solutions of the problems

$$\Lambda v = -\mathring{\varphi} \text{ for } x \in \omega_h, \quad v|_\gamma = 0, \quad (7)$$

$$\Lambda w = -\varphi^* \text{ for } x \in \omega_h, \quad w|_{\gamma_h} = 0. \quad (8).$$

We are going to evaluate separately each of the functions $v(x)$ and $w(x)$.

In order to estimate $v(x)$, it is necessary to construct a majorant $\mathcal{V}(x)$. Assume that the origin is inside the domain G, we try to determine a majorant of the type

$$Y(x) = K(R^2 - r^2), \quad r^2 = \sum_{\alpha=1}^{P} x_\alpha^2,$$

where $K > 0$ is a constant and R is the radius of a P-dimensional ball (a circle for $P=2$) with center at the origin containing entirely the domain G. K will be chosen.

Since, $\Lambda_\alpha x_\beta^2 = 0$ for $\alpha \neq \beta$ and

$$\Lambda_\alpha x_\alpha^2 = \frac{(x_\alpha + h_\alpha)^2 - 2x_\alpha^2 + (x_\alpha - h_\alpha)^2}{h_\alpha^2} = 2,$$

$$\Lambda_\alpha^* x_\alpha^2 = \frac{1}{h_\alpha}\left[\frac{(x_\alpha + h_{\alpha+})^2 - x_\alpha^2}{h_{\alpha+}} - \frac{x_\alpha^2 - (x_\alpha - h_{\alpha-})^2}{h_{\alpha-}}\right] =$$

$$= \frac{1}{h_\alpha}\left[\frac{x_\alpha^2 + 2x_\alpha h_{\alpha+} + h_{\alpha+}^2 - x_\alpha^2}{h_{\alpha+}} - \frac{x_\alpha^2 - x_\alpha^2 + 2x_\alpha h_{\alpha-} - h_{\alpha-}^2}{h_{\alpha-}}\right]$$

$$= \frac{1}{h_\alpha}\left[2x_\alpha + h_{\alpha+} - 2x_\alpha + h_{\alpha-}\right] = \frac{h_{\alpha+} + h_{\alpha-}}{h_\alpha} = 2\cdot\frac{h_{\alpha+} + h_{\alpha-}}{2h_\alpha} = 2\theta_\alpha,$$

$$\theta_\alpha = \frac{h_{\alpha+} + h_{\alpha-}}{2h_\alpha}$$

We find that

$$\Lambda Y = \sum_{\alpha=1}^{P} \Lambda_\alpha Y = -2PK \quad \text{for } x \in \mathring{\omega}_h,$$

$$\Lambda^* Y = -2P\theta K \quad \text{for } x \in \omega_h^*,$$

where $\theta = P^{-1}\sum_{\alpha=1}^{P}\theta_\alpha$. Here $\theta_\alpha = 1$ if a node $x \in \omega_\alpha^*$ is regular along the direction x_α. Thus, Y is a solution of the problem

$$\overline{\Lambda} Y = -\overline{F}(x), \quad Y|_{\gamma_h} = K(R^2 - r^2)|_{\gamma_h} \geq 0,$$

where $\overline{F}(x) = 2PK$ for $x \in \mathring{\omega}_h$ and $\overline{F}(x) = 2P\theta K$ for $x \in \omega_h^*$.

Comparison with problem (7), where $F=\mathring{\varphi}$, that is, $F=0$ for $x\in\omega_h^*$ and $v|_{\gamma_h}=0$, shows that $\bar{F}(x)\geq|F(x)|=|\mathring{\varphi}(x)|$ if we accept $K=\frac{1}{2P}\|\mathring{\varphi}\|_C$. Here the conditions of the comparison theorem are valid as long as $\bar{F}(x)\geq|F(x)|=0$ for $x\in\omega^*$, assuring the relation $\|v\|_C\leq\|y\|_C$.

Since $\|y\|_C\leq KR^2$, so

$$\|v\|_C\leq\frac{R^2}{2P}\|\mathring{\varphi}\|_C=\frac{R^2}{2P}\|\varphi\|_C \qquad (9)$$

is valid in the norm $\|\varphi\|_C=\max_{x\in\mathring{\omega}_h}|\varphi(x)|$.

Our next step is the estimation of the function $w(x)$, which is a solution of problem (8):

$$(8)\quad \Lambda w=-\varphi^*,\ x\in\omega_h,\ w|_{\gamma_h}=0,\ \varphi^*=\begin{cases}0, & x\in\mathring{\omega}_h\\ \varphi, & x\in\omega_h^*\end{cases}$$

First, we are going to show that for problem (8)

$$D(x)\geq\frac{1}{h^2},\ \text{where } h=\max_\alpha h_\alpha,\ x\in\omega_h^*, \qquad (10)$$

$$D(x)=0 \text{ for } x\in\mathring{\omega}_h. \qquad (11)$$

(11) is (clear) simple to show.

After that, we look at eq. (8) at a near-boundary node $x\in\omega_h^*$

$$A(x)w(x)=\sum_{\xi\in Patt(x)}B(x,\xi)w(\xi)+F(x), \qquad (12)$$

$$F(x)=\varphi^*(x),\ w|_{\gamma_h}=0.$$

If one of the nodes $\xi=\xi_0$, say $\xi_0=x^{(+1_\alpha)}$, happens to be a boundary node, then $w(\xi_0)=0$ and the neighborhood $Patt(x)$ contains no point ξ_0

In this case the function $D(x)$ takes on the value

$$D(x) = A(x) - \sum_{\substack{\xi \in Patt'(x) \\ \xi \neq \xi_0}} B(x,\xi)$$

$$= A(x) - \left[\sum_{\xi \in Patt'(x)} B(x,\xi) - B(x,\xi_0) \right] = B(x,\xi_0),$$

since $A(x) = \sum_{\xi \in Patt'(x)} B(x,\xi)$ for the Laplace equation.

$$\Rightarrow \quad D(x) = B(x, x^{(+\alpha)}) > 0.$$

If a node x is near-boundary not only with respect to x_α, but also in other directions, then sum (12) contains no other terms for $\xi = \xi_1, \xi_2, \dots \xi_k$, so that

$$D(x) = B(x,\xi_0) + B(x,\xi_1) + \dots + B(x,\xi_k) > 0.$$

Let $x \in \omega_h^*$ be a near-boundary and irregular node only in some direction x_α and $\xi_0 = x^{(+\alpha)} \in \gamma_h$, $x^{(-\alpha)} \in \omega_h$.

From the equation

$$\Lambda_\alpha^* w + \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{p} \Lambda_\beta w = -\varphi^*(x),$$

where

$$\Lambda_\beta y = y_{x_\beta x_\beta},$$

$$\Lambda_\alpha^* w = \frac{1}{h_\alpha} \left(\frac{w^{(+\alpha)} - w}{h_{\alpha+}} - \frac{w - w^{(-\alpha)}}{h_\alpha} \right) = -\frac{1}{h_\alpha} \left(\frac{w}{h_{\alpha+}} + \frac{w - w^{(-\alpha)}}{h_\alpha} \right),$$

we establish the relations

$$A(x) = \frac{1}{h_\alpha h_{\alpha+}} + \frac{1}{h_\alpha^2} + \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{p} \frac{2}{h_\beta^2}, \qquad \sum_{\xi \in Patt(x)} B(x,\xi) = \frac{1}{h_\alpha^2} + \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{p} \frac{2}{h_\beta^2}$$

$$\Rightarrow \quad D(x) = A(x) - \sum_{\xi \in Patt'(x)} B(x,\xi) = \frac{1}{h_\alpha h_{\alpha+}} \geqslant \frac{1}{h^2}$$

If x is a regular near-boundary node only with respect to x_α, then

$$D(x) = \frac{1}{h_\alpha^2} \geq \frac{1}{h^2} .$$

When, in addition, it turns out to be near-boundary in other directions, the function $D(x)$ can only increase.

Then

$$\|w\|_C \leq \left\| \frac{\varphi^*}{D} \right\|_{C^*} \leq h^2 \|\varphi^*\|_C . \qquad (13)$$

Collecting estimates (6), (9), (13) and then involving the well-known inequality

$$\|y\|_{\bar{C}} \leq \|\bar{y}\|_{\mathring{C}} + \|v\|_C + \|w\|_C ,$$

we establish the following theorem.

Theorem 1. For a solution of the Dirichlet difference problem the estimate

$$\|y\|_{\bar{C}} \leq \|\mu\|_{C_\gamma} + \frac{R^2}{2P} \|\varphi\|_{\mathring{C}} + h^2 \|\varphi\|_{C^*} \qquad (14)$$

holds with

$$\|f\|_{\bar{C}} = \max_{x \in \omega_h + \gamma_h} |f(x)| , \quad \|f\|_{\mathring{C}} = \max_{x \in \mathring{\omega}_h} |f(x)| ,$$

$$\|f\|_{C^*} = \max_{x \in \omega_h^*} |f(x)| , \quad \|f\|_{C_\gamma} = \max_{x \in \gamma_h} |f(x)| .$$

This theorem expresses the stability of the Dirichlet difference problem (1) with respect to the boundary data and the right-hand side.

2. The uniform convergence and the order of accuracy of a difference scheme.

2. The uniform convergence and the order of accuracy of a difference scheme.

Let

$$z = y - u,$$

where y is a solution of problem (1), and $u = u(x)$ is a solution of problem (1). Substitution $y = z + u$ into (1) or (2) yields

$$\bar{\Lambda} y = -\varphi \Rightarrow \bar{\Lambda}(z+u) = \bar{\Lambda} z + \Lambda u = -\varphi$$

$$\Rightarrow \quad \bar{\Lambda} z = -\underbrace{(\Lambda u + \varphi)}_{\psi(x)}, \quad x \in \omega, \quad z|_{\gamma} = 0.$$

$$\Rightarrow \quad \bar{\Lambda} z = -\psi(x), \quad x \in \omega, \quad z|_{\gamma} = 0, \qquad (15)$$

where $\psi(x) = \bar{\Lambda} u + \varphi(x)$ is the residual.

We stated before that

$$\psi(x) = O(|h|^2) = O(h^2), \quad \text{at the regular nodes,}$$

$$\psi(x) = O(1) \quad \text{at the irregular nodes.}$$

or, more specifically,

$$|\psi| \le \frac{M_4 |h|^2}{12} \le p \frac{M_4}{12} h^2 \quad \text{at the regular nodes}$$

$$|\psi| \le p M_2 \quad \text{at the irregular nodes.}$$

where

$$M_k = \max_{\substack{x \in \bar{G} \\ 1 \le \alpha \le p}} \left| \frac{\partial^k u}{\partial x_\alpha^k} \right|, \quad k = 2, 3, 4 \ldots, \quad |h|^2 = \sum_{\alpha=1}^{p} h_\alpha^2, \quad h = \max_{1 \le \alpha \le p} h_\alpha$$

By Theorem 1 (estimate (14)), we have

$$\|z\|_C \le \frac{R^2}{2p} \|\psi\|_C + h^2 \|\psi\|_C.$$

$$\Rightarrow \quad \|z\|_C = \|y - u\|_C \le$$

$$\le \left(\frac{R^2}{24} M_4 + p M_2 \right) h^2$$

Therefore, we have proved

Theorem 2 If $u(x) \in C^4(\overline{G})$, that is, a solution possesses continuous derivatives in $\overline{G} = G + \Gamma$ of the first four orders, then the difference scheme converges uniformly with the rate $O(h^2)$, that is, it is of second-order accuracy, so that estimate (16) is valid.

Higher-accuracy schemes for Poisson's equation.

On the basis of the "cross" scheme it is possible to construct a scheme with the error of approximation $O(|h|^4)$ or $O(|h|^6)$ on a solution in the case of a square (cube) grid. In order to raise the order of approximation, we use the fact that $u = u(x)$ is a solution of Poisson's eq.

$$\Delta u = -f(x) \qquad (1)$$

Without loss of generality we may restrict ourselves to the careful analysis of the two-dimensional case ($p=2$) where

$$\Delta u = L_1 u + L_2 u, \quad L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2},$$

Let

$$\Lambda u = (\Lambda_1 + \Lambda_2) u, \quad \Lambda_\alpha u = u_{\bar{x}_\alpha x_\alpha}.$$

Then

$$\Lambda u - Lu = \frac{h_1^2}{12} L_1^2 u + \frac{h_2^2}{12} L_2^2 u + O(|h|^4) \qquad (2)$$

From the equation

$$L_1 u + L_2 u = -f(x),$$

we find that

$$L_1^2 u = -L_1 f - L_1 L_2 u, \quad L_2^2 u = -L_2 f - L_1 L_2 u.$$

So that

$$\Lambda u = Lu + \frac{h_1^2}{12}\left(-L_1 f - L_1 L_2 u\right) + \frac{h_2^2}{12}\left(-L_2 f - L_1 L_2 u\right) + O(|h|^4)$$

$$= Lu - \frac{h_1^2}{12} L_1 f - \frac{h_2^2}{12} L_2 f - \frac{h_1^2 + h_2^2}{12} L_1 L_2 u + O(|h|^4) \qquad (3)$$

We substitute here $Lu = -f$ and replace $L_1 L_2 u$ by the difference operator. We use the approximation

$$L_1 L_2 u = \frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} \sim \Lambda_1 \Lambda_2 u = u_{\bar{x}_1 x_1 \bar{x}_2 x_2}$$

The expression for $\Lambda_1\Lambda_2 u$ such that

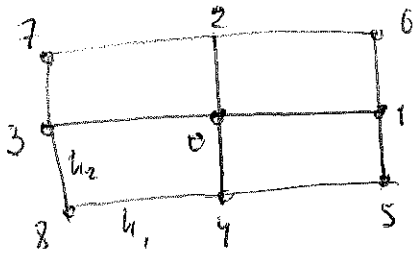
$$\Lambda_1\Lambda_2 u = \Lambda_1\left[\frac{u(x_1, x_2-h_2) - 2u(x_1,x_2) + u(x_1, x_2+h_2)}{h_2^2}\right]$$

$$= \frac{1}{h_1^2 h_2^2}\Big\{ u(x_1-h_1, x_2-h_2) - 2u(x_1-h_1, x_2) + u(x_1-h_1, x_2+h_2)$$

$$- 2u(x_1, x_2-h_2) + 4u(x_1,x_2) - 2u(x_1, x_2+h_2)$$

$$+ u(x_1+h_1, x_2-h_2) - 2u(x_1+h_1, x_2) + u(x_1+h_1, x_2+h_2)\Big\}$$

is needed in the estimation of the error of approximation to $\Lambda_1\Lambda_2 u - L_1 L_2 u$ by virtue of the well-established expansion



$$\Lambda v = v_{\bar{x}x} = \frac{v(x+h) - 2v(x) + v(x-h)}{h^2} = v''(\xi), \quad \xi = x+\theta h, \ |\theta|\le 1. \tag{4}$$

assuming that $v(x)$ has a continuous second derivative on the segment $[x-h, x+h]$;

$$\Lambda v = v_{\bar{x}x} = v''(x) + \frac{h^2}{12} v^{(4)}(\xi^*), \quad \xi^* = x+\theta^* h, \ |\theta^*|\le 1, \tag{5}$$

$v(x)$ has a continuous fourth derivative on the segment $[x-h, x+h]$. By relating x_1 to be fixed we might have

$$\Lambda_2 u = L_2 u(x_1, x_2) + \frac{h_2^2}{12}\frac{\partial^4 u}{\partial x_2^4}(x_1, \xi_2), \quad \xi_2 = x+\theta_2 h_2, \ |\theta_2|\le 1.$$

Then

$$\Lambda_1\Lambda_2 u(x_1,x_2) = \Lambda_1 L_2 u(x_1,x_2) + \frac{h_2^2}{12}\Lambda_1\frac{\partial^4 u}{\partial x_2^4}(x_1, \xi_2)$$

Applying formula (5) with $v = L_2 u$ and $x = x_1$ to the first summand yields

$$\frac{\partial^4}{\partial x_1^4}\left(\frac{\partial^4 u}{\partial x_2^4}\right)(\xi_1^*, x_2) = \frac{\partial^8 u(\xi_1^*, x_2)}{\partial x_1^4 \partial x_2^4}$$

$$\Lambda_1 L_2 u(x_1,x_2) = L_1 L_2 u(x_1,x_2) + \frac{h_1^2}{12}\frac{\partial^6 u}{\partial x_1^4}(\xi_1^*, x_2), \quad \xi_1^* = x_1+\theta_1^* h_1, \quad |\theta_1^*|\le 1.$$

By the formula (4):

$$\frac{h_2^2}{12}\Lambda_1\frac{\partial^4 u}{\partial x_2^4}(x_1,\xi_2) = \frac{h_2^2}{12}\frac{\partial^6 u}{\partial x_1^2 \partial x_2^4}(\xi_1, \xi_2), \quad \xi_1 = x_1+\theta_1 h_1, \ |\theta_1|\le 1.$$

Therefore,

$$(\Lambda_1\Lambda_2 - L_1L_2)u = O(h_1^2) + O(h_2^2) = O(|h|^2).$$

Substituting into (3) the expression for $\Lambda_1\Lambda_2 u$ in place of L_1L_2u

$$L_1L_2u = \Lambda_1\Lambda_2 u + O(|h|^2),$$

and involving the eq. $Lu = -f(x)$, we finally get

$$\Lambda u = Lu - \frac{h_1^2+h_2^2}{12}\Lambda_1\Lambda_2 u - \frac{h_1^2}{12}L_1 f - \frac{h_2^2}{12}L_2 f + O(|h|^4)$$
$$= -\left(f + \frac{h_1^2}{12}L_1 f + \frac{h_2^2}{12}L_2 f\right) - \frac{h_1^2+h_2^2}{12}\Lambda_1\Lambda_2 u + O(|h|^4). \qquad (6)$$

Because of this, the equation

$$\begin{cases} \Lambda' y = -\varphi, \quad \Lambda' y = \Lambda y + \dfrac{h_1^2+h_2^2}{12}\Lambda_1\Lambda_2 y, \\ \varphi = f + \dfrac{h_1^2}{12}L_1 f + \dfrac{h_2^2}{12}L_2 f, \end{cases} \qquad (7)$$

provides an approximation of order 4 on a solution $u = u(x)$ of Poisson's equation (1). Indeed, formula (6) gives

$$\Lambda' u + \varphi = (\Lambda' u + \varphi) - (Lu + f) = O(|h|^4), \quad L = L_1 + L_2.$$

The operator Λ' is defined on the nine-point "box" pattern

(7) is representable by

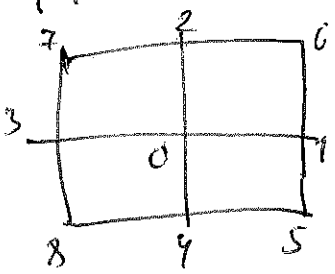
$$\frac{5}{3}\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)y = \frac{1}{6}\left(\frac{5}{h_1^2} - \frac{1}{h_2^2}\right)\left(y^{(+1_1)} + y^{(-1_1)}\right)$$
$$+ \frac{1}{6}\left(\frac{5}{h_2^2} - \frac{1}{h_1^2}\right)\left(y^{(+1_2)} + y^{(-1_2)}\right)$$
$$+ \frac{1}{12}\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)\left(y^{(+1_1,+1_2)} + y^{(+1_1,-1_2)}\right)$$
$$+ \left(y^{(-1_1,-1_2)} + y^{(-1_1,+1_2)}\right) + \varphi.$$

where $y^{(\pm 1_1)} = y(x_1 \pm h_1, x_2)$, $y^{(+1_1,-1_2)} = y(x_1+h_1, x_2-h_2)$, etc.

On the square grid $(h_1 = h_2 = h)$ the final result is:

$$y_0 = \frac{4(y_1 + y_2 + y_3 + y_4) + y_5 + y_6 + y_7 + y_8}{20} + \frac{3}{10} h^2 \varphi_0$$

$$\varphi = f + \frac{h_1^2}{12} \Lambda_1 f + \frac{h_2^2}{12} \Lambda_2 f .$$



2. Estimation of a solution of the difference boundary-value problem.

Consider now the difference Dirichlet problem for the scheme of accuracy $O(|h|^4)$ in the rectangle $\overline{G} = \{0 \le x_\alpha \le l_\alpha, \alpha = 1,2\}$:

$$\begin{cases} \Lambda' y = -\varphi, \quad x \in \omega_h, \quad y|_{\gamma_h} = \mu(x) \\ \varphi = f + \frac{h_1^2}{12} \Lambda_1 f + \frac{h_2^2}{12} \Lambda_2 f, \end{cases} \tag{9}$$

where $\Lambda' y$ is given by (7). Each of the grid nodes is regular, the boundary γ_h of the grid contains all the nodes on the boundary Γ including the vertices of the rectangle.

Let $z = y - u$:

$$\Lambda' z = -\psi, \quad x \in \omega_h, \quad z = 0 \text{ on } \gamma_h, \tag{10}$$

where $\psi = \Lambda' u + \varphi = O(|h|^4)$ for $x \in \omega_h$, if $u \in C^{(6)}$.

Also, we have

$$B(x,\xi) \ge 0 \quad \text{for} \quad \frac{1}{\sqrt{5}} \le \frac{h_1}{h_2} \le \sqrt{5} \tag{11}$$

$$\left.\begin{array}{l} \left(\frac{5}{h_1^2} - \frac{1}{h_2^2} \ge 0 \Rightarrow 5 \ge \left(\frac{h_1}{h_2}\right)^2 \Rightarrow \frac{h_1}{h_2} \le \sqrt{5}\right. \\ \left.\left(\frac{5}{h_2} - \frac{1}{h_1^2} \ge 0 \quad \left(\frac{h_1}{h_2}\right)^2 \ge \frac{1}{5} \Rightarrow \frac{h_1}{h_2} \ge \frac{1}{\sqrt{5}}\right.\right. \end{array}\right\} \Rightarrow \frac{1}{\sqrt{5}} \le \frac{h_1}{h_2} \le \sqrt{5}$$

To estimate the solution of problem (10), we take the majorant of the type

$$\gamma(x) = K\left(\ell_1^2 - x_1^2 + \ell_2^2 - x_2^2\right).$$

Taking into account that $\Lambda\gamma = -4K$, $\Lambda_1\Lambda_2\gamma = 0$,

$\|\gamma\| \le K(\ell_1^2 + \ell_2^2)$ and accepting $4K = \|\psi\|_C$,

we obtain for the solution of problem (10), we have

$$\|z\|_C \le \frac{\ell_1^2 + \ell_2^2}{4}\|\psi\|_C \quad \text{provided} \quad \frac{1}{\sqrt{5}} \le \frac{h_1}{h_2} \le \sqrt{5}$$

This implies that

Scheme (9) is of fourth-order accuracy when $u \in C^{(6)}$, $f \in C^{(4)}$ and condition (11) is satisfied.

Remark

The condition (11) is automatically fulfilled, when $h_1 = h_2 = h$ (on the square grid). When $u \in C^{(8)}$ and with a proper choice of ψ guarantees the sixth order of accuracy of scheme (9) on the square grid.

ADVANCED NUMERICAL ANALYSIS

Part 2

Parabolic equations

Transformation to non-dimensional form

1D case

$$\frac{\partial U}{\partial T} = K \frac{\partial^2 U}{\partial X^2}, \quad K \text{ is constant}, \qquad (1)$$

The solution of which gives the temperature U at a distance X from one end of a thin uniform rod after a time T.

Let L represent the length of a rod and U_0 some particular temperature such as the maximum or minimum temperature at zero time.

Put
$$x = \frac{X}{L} \quad \text{and} \quad u = \frac{U}{U_0}.$$

Then
$$\frac{\partial U}{\partial X} = \frac{\partial U}{\partial x} \cdot \frac{dx}{dX} = \frac{\partial U}{\partial x} \cdot \frac{1}{L}$$

$$\frac{\partial^2 U}{\partial X^2} = \frac{\partial}{\partial X}\left(\frac{\partial U}{\partial X}\right) = \frac{\partial}{\partial X}\left(\frac{\partial U}{\partial x} \cdot \frac{1}{L}\right) = \frac{\partial}{\partial x}\left(\frac{\partial^2 U}{\partial x^2} \cdot \frac{1}{L}\right) \cdot \frac{dx}{dX}$$

$$= \frac{1}{L^2} \frac{\partial^2 U}{\partial x^2}$$

$$\Rightarrow \frac{\partial (u \cdot U_0)}{\partial T} = \frac{K}{L^2} \frac{\partial^2 (u \cdot U_0)}{\partial x^2} \Rightarrow \frac{L^2}{K} \frac{\partial u}{\partial T} = \frac{\partial^2 u}{\partial x^2}$$

Writing $t = \frac{KT}{L^2}$, we have $\frac{\partial u}{\partial T} = \frac{\partial u}{\partial t} \cdot \frac{dt}{dT} = \frac{\partial u}{\partial t} \cdot \frac{K}{L^2}$

$$\Rightarrow \frac{L^2}{K} \cdot \frac{K}{L^2} \frac{\partial u}{\partial t} = \frac{\partial u}{\partial t}$$

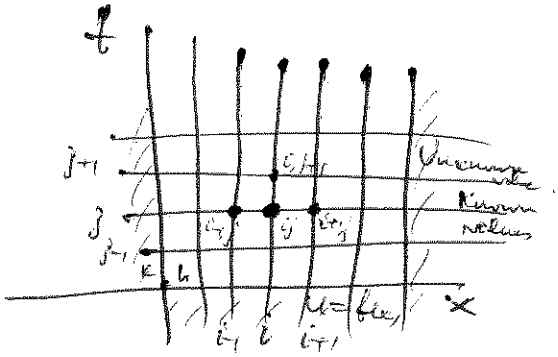
$$\Rightarrow \boxed{\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}} \qquad (2)$$

is the non-dimensional form of (1).

An explicit method of solution

One of the explicit methods of finite-difference approximation to

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad (3)$$



is

$$\frac{u_{i,j+1} - u_{ij}}{k} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2},$$

where

$$x = ih, \quad (i = 0, 1, 2, \ldots),$$

and

$$t = jk, \quad (j = 0, 1, 2, \ldots).$$

This can be written as

$$u_{i,j+1} = u_{ij} + r(u_{i-1,j} - 2u_{ij} + u_{i+1,j}), \qquad (4)$$

where $r = \frac{\delta t}{(\delta x)^2} = \frac{k}{h^2}$, and gives a formula for the unknown "temperature" $u_{i,j+1}$ at the $(i, j+1)$-th mesh point in terms of known "temperature" along the j-th time row.

Example 1 (see G. D. Smith P.13)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

(i) $u = 0$ when $x = 0$ and 1 for all t. (The boundary conditions)

(ii) $u = 2x$ for $0 \le x \le \frac{1}{2}$, and $u = 2(1-x)$ for $\frac{1}{2} \le x \le 1$ for $t = 0$. (The initial condition)

It will be proved that this explicit method is valid only when $0 < r \le \frac{1}{2}$ (for stability $\Rightarrow$ and convergence)

$$0 < r = \frac{k}{h^2} \le \frac{1}{2}$$

Case 1. $h = \frac{1}{10}$, $k = \frac{1}{100}$

So $r = \frac{\frac{1}{100}}{\frac{1}{100}} = \frac{1}{10} < \frac{1}{2}$

Case 2 $h = \frac{1}{10}$, $k = \frac{5}{1000} \Rightarrow r = \frac{\frac{5}{1000}}{\frac{1}{100}} = \frac{5}{10} = 0.5$

Case 3. $h = \frac{1}{10}$, $k = \frac{1}{100} \Rightarrow r = \frac{\frac{1}{100}}{\frac{1}{100}} = 1$

Solution of F.D eq.s meaningless.

Home work.

Take a) $h = \frac{1}{4}$, $k = \frac{1}{20}$

b) $h = \frac{1}{4}$, $k = \frac{1}{10}$.

The analytical solution:

$$u = \frac{8}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \left(\sin \frac{1}{2} n\pi\right)\left(\sin n\pi x\right) e^{-n^2 \pi^2 t}$$

Crank-Nicolson implicit method

Although the explicit method is computationally simple it has one serious drawback. The time step $\delta t = k$ is necessarily very small because the process is valid only for $0 < \frac{k}{h^2} \le \frac{1}{2}$, i.e., $k \le \frac{1}{2}h^2$, and $h = \delta x$ must be kept small in order to attain reasonable accuracy.

Crank and Nicolson (1947) proposed, and used, a method that reduces the total volume of calculation and is valid (i.e., convergent and stable) for all finite values of $r = \frac{k}{h^2}$.

They replaced $\frac{\partial^2 u}{\partial x^2}$ by the mean of its FD representations on the $(j+1)$th and jth time rows and approximated the eq.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad (1)$$

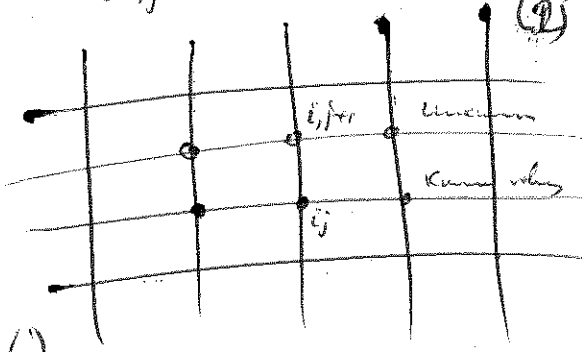
by $$\frac{u_{i,j+1} - u_{ij}}{k} = \frac{1}{2}\left\{ \frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{h^2} + \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} \right\}$$

giving

$$2u_{i,j+1} - 2u_{ij} = r\left(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} + u_{i+1,j} - 2u_{ij} + u_{i-1,j}\right)$$

$$\Rightarrow \quad -r u_{i-1,j+1} + (2+2r)u_{i,j+1} - r u_{i+1,j+1} = r u_{i-1,j} + (2-2r)u_{ij} + r u_{i+1,j} \qquad (2)$$

where $r = \frac{k}{h^2}$.



$\Rightarrow$ Implicit method.

See Example 2. (In the text book)

Derivative boundary conditions

Example 1.

Solve the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (0 < x < 1) \qquad (1)$$

satisfying the initial condition

$u = 1$ for $0 \le x \le 1$ when $t = 0$,

and the b.c.s

$\frac{\partial u}{\partial x} = u$ at $x = 0$, for all t,

$\frac{\partial u}{\partial x} = -u$ at $x = 1$, for all t.

Using an explicit method and employing central-differences for the boundary conditions.

One explicit finite-difference representation of eq. (1) is

$$\frac{u_{i,j+1} - u_{ij}}{\delta t} = \frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{(\delta x)^2}, \quad i = 1, 2, \ldots, N-1,$$

i.e.,

$$u_{i,j+1} = u_{ij} + r(u_{i-1,j} - 2u_{ij} + u_{i+1,j}), \qquad (2)$$

where $r = \frac{\delta t}{(\delta x)^2}$.

At $x = 0$, (i.e., $i = 0$)

$$u_{0,j+1} = u_{0j} + r(u_{-1,j} - 2u_{0j} + u_{1,j}). \qquad (3)$$

The b.c. at $x = 0$, in terms of central differences, can be written as

$$\frac{u_{1,j} - u_{-1,j}}{2\delta x} = u_{0,j}. \qquad (4)$$

Eliminating $u_{-1,j}$ between (3) and (4) gives

$$u_{-1,j} = u_{1,j} - 2\delta x\, u_{0,j}$$

$$\Rightarrow u_{0,j+1} = u_{0,j} + r\left(u_{0,j} - 2\delta x\, u_{0,j} - 2u_{0,j} + u_{1,j}\right)$$

$$\Rightarrow \boxed{u_{0,j+1} = u_{0,j} + 2r\left\{u_{1,j} - (1+\delta x)u_{0,j}\right\}.} \qquad (5)$$

At $x=1$ (i.e., $i=N$), from (2), we have

$$u_{N,j+1} = u_{N,j} + r\left(u_{N-1,j} - 2u_{N,j} + \underset{\sim}{u_{N+1,j}}\right). \qquad (6)$$

(? over $u_{N+1,j}$)

and the central difference approximation of the b.c. at $x=1$, we have

$$\frac{u_{N+1,j} - u_{N-1,j}}{2\delta x} = -u_{N,j}. \qquad (7)$$

(? over $u_{N+1,j}$)

From (6) and (7) we obtain

$$\begin{cases} u_{N+1,j} = u_{N-1,j} - 2\delta x\, u_{N,j} \\ u_{N,j+1} = u_{N,j} + r\left(u_{N-1,j} - 2u_{N,j} + u_{N-1,j} - 2\delta x\, u_{N,j}\right) \end{cases}$$

$$\Rightarrow \boxed{u_{N,j+1} = u_{N,j} + 2r\left\{u_{N-1,j} - (1+\delta x)u_{N,j}\right\}} \qquad (8)$$

The system of eq.s (2), (5), and (8) is an approximation of the boundary value problem.

Example 2. Re-solve the example 1 using an explicit method and employing a forward-difference for the boundary conditions at $x=0$ $(x=1)$

From (2):

$$u_{i,j+1} = u_{ij} + r(u_{i-1,j} - 2u_{ij} + u_{i+1,j})$$

for $i=1$

$$u_{1,j+1} = u_{1,j} + r(u_{0,j} - 2u_{1,j} + u_{2,j}) \qquad (9)$$

The boundary condition at $x=0$, namely $\frac{\partial u}{\partial x} = u$, in terms of a forward difference is

$$\frac{u_{1,j} - u_{0,j}}{\delta x} = u_{0,j}$$

so $$u_{0,j} = u_{1,j} - \delta x\, u_{0,j} \Rightarrow$$

$$\Rightarrow u_{0,j} = \frac{u_{1,j}}{1+\delta x} \qquad (10)$$

From (9) and (10), we obtain

$$u_{1,j+1} = u_{1,j} + r\left(\frac{u_{1,j}}{1+\delta x} - 2u_{1,j} + u_{2,j}\right)$$

$$\Rightarrow u_{1,j+1} = \left(1 - 2r + \frac{r}{1+\delta x}\right)u_{1,j} + r\,u_{2,j} \qquad (11)$$

Similarly the finite-difference eq. for $\frac{\partial u}{\partial x} = -u$ at $x=1$ can be obtained.

Example 3 Solve example 1 by the Crank-Nicolson method:

$$\frac{u_{i,j+1}-u_{ij}}{\delta t} = \frac{1}{2}\left\{\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{(\delta x)^2} + \frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{(\delta x)^2}\right\}$$

$$\Rightarrow -ru_{i-1,j+1}+(2+2r)u_{i,j+1}-ru_{i+1,j+1} = ru_{i-1,j}+(2-2r)u_{ij}+ru_{i+1,j} \quad (12)$$

The central difference representation of the b.c. at $x=0$ is

$$\frac{u_{1j}-u_{-1,j}}{2\delta x} = u_{0,j},$$

from which it follows that

$$u_{-1,j} = u_{1,j} - 2\delta x\, u_{0,j}$$

and $$u_{-1,j+1} = u_{1,j+1} - 2\delta x\, u_{0,j+1}.$$

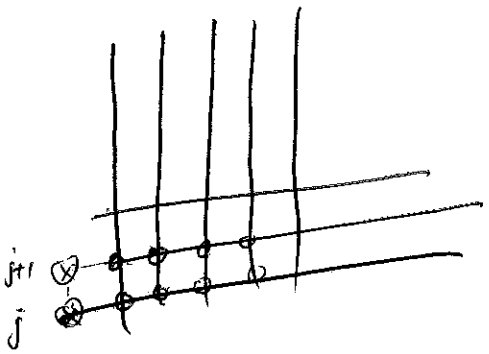
From (12) for $i=0$:

$$-ru_{-1,j+1} + (2+2r)u_{0,j+1} - ru_{1,j+1} = ru_{-1,j} + (2-2r)u_{0j} + ru_{1,j}$$

$$-r(u_{1,j+1}-2\delta x\, u_{0,j+1}) + (2+2r)u_{0,j+1} - ru_{1,j+1} = r(u_{1,j}-2\delta x\, u_{0,j}) + (2-2r)u_{0j} + ru_{1,j}$$

$$-2ru_{1,j+1} + (2r\delta x + 2 + 2r)u_{0,j+1} = 2ru_{1,j} + (2r\delta x + 2 - 2r)u_{0j}$$

The obtained system can be solved by the direct elimination method.



The local truncation error

Consider the DE:

$$h(U_{ij}) = \left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{i,j} = 0$$

and let

$$F(u_{ij}) = 0 = \frac{u_{i,j+1} - u_{ij}}{k} - \frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h^2} \qquad (1)$$

be the classical explicit eq.

The local truncation error is defined as:

$$T_{ij} = F(U_{ij}) = \frac{U_{i,j+1} - U_{ij}}{k} - \frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} \qquad (2)$$

This can be written

$$U_{i,j+1} = kT_{ij} + rU_{i-1,j} + (1-2r)U_{ij} + rU_{i+1,j}. \qquad (3)$$

Assume now that the solution U of the PDE is known at all mesh points up to and including those at the j-th time-level. We could then calculate a local approximation $u_{i,j+1}$ to $U_{i,j+1}$ by means of (1) and get

$$u_{i,j+1} = rU_{i-1,j} + (1-2r)U_{ij} + rU_{i+1,j} \qquad (4)$$

By (3) and (4) it follows that

$$U_{i,j+1} - u_{i,j+1} = \text{(local) discretization error at } (i,j+1) = kT_{ij}. \qquad (5)$$

This shows that the local truncation error at the point (i,j) is a measure of the (local) discretization error at the point $(i,j+1)$ when the finite-difference scheme is applied once only to the exact solution values of the PDE, all arithmetic being exact, i.e., without rounding errors.

By means of Taylor's expansion it is easy to express T_{ij} in terms of powers of h and k and part-derivatives of U at (ih, jk).

Example 3. Calculate the order of the local truncation error of the classical explicit difference approximation to

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0.$$

By Taylor's expansion

$$U_{i\pm1,j} = U\{(i\pm1)h, k\} = U(x_i \pm h, t_j)$$

$$= U_{ij} \pm h\left(\frac{\partial U}{\partial x}\right)_{i,j} + \frac{1}{2}h^2\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j} \pm \frac{1}{6}h^3\left(\frac{\partial^3 U}{\partial x^3}\right)_{i,j} + \frac{h^4}{24}\left(\frac{\partial^4 U}{\partial x^4}\right)_{i,j} \pm \cdots$$

$$U_{i,j\pm1} = U(x_i, t_j \pm k) = U_{ij} \pm k\left(\frac{\partial U}{\partial t}\right)_{i,j} + \frac{1}{2}k^2\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j} \pm \frac{k^3}{6}\left(\frac{\partial^3 U}{\partial t^3}\right)_{i,j} + \cdots$$

Hence

$$T_{ij} = \frac{U_{i,j+1} - U_{ij}}{k} - \frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} = \left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{i,j} + \frac{k}{2}\left(\frac{\partial^2 U}{\partial t^2}\right)_{ij}$$

$$- \frac{h^2}{12}\left(\frac{\partial^4 U}{\partial x^4}\right)_{i,j} + \frac{k^2}{6}\left(\frac{\partial^3 U}{\partial t^3}\right)_{ij} - \frac{1}{360}h^4\frac{\partial^6 U}{\partial x^6} + \cdots \qquad (6)$$

But U is the solution of the DE, so

$$\left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{i,j} = 0$$

Therefore the principal part of the local truncation error is

$$\left(\frac{1}{2}k\frac{\partial^2 U}{\partial t^2} - \frac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}\right)_{i,j}.$$

Hence

$$T_{ij} = O(k) + O(h^2).$$

When $k = rh^2$, $0 < r \le \frac{1}{2}$, T_{ij} is $O(k)$ or $O(h^2)$.

This error may be further reduced by choosing a special value for $\frac{k}{h^2}$ because eq. (6) can be written as

$$T_{ij} = \frac{1}{12}h^2\left(6\frac{K}{h^2}\frac{\partial^2 U}{\partial t^2} - \frac{\partial^4 U}{\partial x^2}\right)_{ij} + O(K^2) + O(h^4).$$

By DE

$$\frac{\partial}{\partial t} = \frac{\partial^2}{\partial x^2},$$

so

$$\frac{\partial}{\partial t}\left(\frac{\partial U}{\partial t}\right) = \frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 U}{\partial x^2}\right)$$

assuming that these derivatives exist. If we put $\frac{6K}{h^2} = 1$, $\left(K = \frac{1}{6}h^2\right)$ then

$$T_{ij} = O(K^2) + O(h^4).$$

(ii) Consistency or Compatibility

It is sometimes possible to approximate a parabolic or hyperbolic equation by a finite-difference scheme that is stable but which has a solution that converges to the solution of a different DE as the mesh lengths tend to zero. Such a difference scheme is said to be inconsistent or incompatible with the PDE.

Example. The equation

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$$

is approximated at the point (ih, jk) by the difference eq.

$$\frac{U_{i,j+1} - U_{i,j-1}}{2k} - \frac{U_{i+1,j} - 2\{\theta U_{i,j+1} + (1-\theta) U_{i,j-1}\} + U_{i-1,j}}{h^2} = 0.$$

Show that the local truncation error at this point is

$$\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{h^2}{12}\frac{\partial^4 U}{\partial x^4} + (2\theta - 1)\frac{2k}{h^2}\frac{\partial U}{\partial t} + \frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2} + O\left(\frac{k^2}{h^2}, h^4, k^4\right).$$

Discuss the consistency of this scheme with the PDE when:

(i) $k = rh$ and (ii) $k = rh^2$,

where r is a positive constant and θ a variable parameter.

Expansion of the terms $U_{i,j+1}$, $U_{i,j-1}$, $U_{i+1,j}$, and $U_{i-1,j}$ about (ph,jk) by Taylor's series

$$T_{ij} = \frac{U_{i,j+1} - U_{i,j-1}}{2k} - \frac{U_{i+1,j} - 2\{\theta U_{i,j+1} + (1-\theta)U_{i,j-1}\} + U_{i-1,j}}{h^2}$$

leads to

$$T_{ij} = \left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{i,j} + \left\{\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{h^2}{12}\frac{\partial^4 U}{\partial x^4} + (2\theta-1)\frac{2k}{h^2}\frac{\partial U}{\partial t} + \frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2}\right\} + O\left(\frac{k^3}{h^2}, h^4, k^4\right).$$

Case (i) $k = rh$

As $h \to 0$,

$$T_{ij} = F(U_{ij}) \to \left\{\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + (2\theta-1)\frac{2r}{h}\frac{\partial U}{\partial t} + r^2\frac{\partial^2 U}{\partial t^2}\right\}_{i,j}$$

When $\theta \neq \frac{1}{2}$ the third term tends to infinity. When $\theta = \frac{1}{2}$ the limiting value of T_{ij} is

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + r^2\frac{\partial^2 U}{\partial t^2}.$$

In this case the finite-difference equation is consistent with the hyperbolic equation

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + r^2\frac{\partial^2 U}{\partial t^2} = 0.$$

Hence the difference equation is always <u>inconsistent</u> with $\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$ when $k = rh$.

Case (ii) $K = rh^2$

As $h \to 0$,

$$T_{ij} \to \frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + 2(2\theta - 1) r \frac{\partial U}{\partial t}.$$

When $\theta \neq \frac{1}{2}$ the difference scheme is consistent with the parabolic eq.

$$\left\{1 + 2(2\theta - 1) r\right\} \frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0.$$

It is only when $\theta = \frac{1}{2}$ that the difference-scheme is consistent with the given DE. This is then the well-known Du Fort and Frankel three-level explicit scheme which is also stable for all $r > 0$ (Show! At Home)

Home work

Show that the explicit scheme

$$\frac{U_{i,j+1} - U_{i,j-1}}{2k} - \frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} = 0$$

gives a local truncation error of $O(k^2) + O(h^2)$, but unconditional instable.

Lax's equivalence theorem

Given a properly posed linear initial-value problem and a linear finite-difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence.

A problem is properly posed if:

(i) The solution is unique when it exists.

(ii) The solution depends continuously on the initial data

Convergence, stability, and consistency

Let U be the exact solution of a PDE with independent variables x and t, and u the exact solution of the difference equations used to approximate the PDE. Then the FDE is said to be convergent when u tends to U at a fixed point or along a fixed t-level as δx and δt both tend to zero.

The difference $(U-u)$ is called the discretization error.

Let $F_{ij}(u)=v$ represent the difference equation at the (i,j)th mesh point. The value $F_{ij}(U)$ is called the local truncation error at the (i,j)th mesh point. If this tends to zero as the mesh lengths tend to zero the difference equation is said to be consistent with the PDE.

In practice the solution actually computed is not u but, N, (say). N will be called the numerical solution and $(u-N)$ the global rounding error R. The total error at the (i,j)th mesh point is

$$U_{ij}-N_{ij} = (U_{ij}-u_{ij}) + (u_{ij}-N_{ij}) = \text{discretization error} + \text{global rounding error}.$$

If the growth of R_{ij} is bounded for all i as j tends to infinity, then we say that the difference eq.s are stable.

The local truncation error and convergency

The local truncation error

Let $F_{ij}(u)=0$ represent the difference equation approximating the PDE at the (i,j)-th mesh point, with exact solution u. If u is replaced by U at the mesh points of the difference equation, where U is the exact solution of PDE, the value $F_{ij}(U)$ is called the local truncation error T_{ij} at the (i,j) mesh point.

Example. Calculate the order of the local truncation error of the classical explicit difference approximation to

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}=0$$

at the point (ih, jk).

$$F_{ij}(u)=\frac{u_{i,j+1}-u_{ij}}{k}-\frac{u_{i-1,j}-2u_{ij}+u_{i+1,j}}{h^2}=0$$

Therefore,

$$T_{ij}=F_{ij}(U)=\frac{U_{i,j+1}-U_{ij}}{k}-\frac{U_{i-1,j}-2U_{ij}+U_{i+1,j}}{h^2}$$

By Taylor's expansion

$$U_{i+1,j}=U\{(i+1)h,jk\}=U(x_i+h,t_j)$$
$$=U_{ij}+h\left(\frac{\partial U}{\partial x}\right)_{ij}+\frac{h^2}{2}\left(\frac{\partial^2 U}{\partial x^2}\right)_{ij}+\frac{h^3}{6}\left(\frac{\partial^3 U}{\partial x^3}\right)_{ij}+\frac{h^4}{24}\left(\frac{\partial^4 U}{\partial x^4}\right)_{ij}+\cdots$$

$$U_{i-1,j}=U\{(i-1)h,jk\}=U(x_i-h,t_j)$$
$$=U_{ij}-h\left(\frac{\partial U}{\partial x}\right)_{ij}+\frac{h^2}{2}\left(\frac{\partial^2 U}{\partial x^2}\right)_{ij}-\frac{h^3}{6}\left(\frac{\partial^3 U}{\partial x^3}\right)_{ij}+\cdots$$

$$U_{i,j+1}=U(x_i,t_j+k)=U_{ij}+k\left(\frac{\partial U}{\partial t}\right)_{ij}+\frac{k^2}{2}\left(\frac{\partial^2 U}{\partial t^2}\right)_{ij}+\frac{k^3}{6}\left(\frac{\partial^3 U}{\partial t^3}\right)_{ij}+\cdots$$

Substitution into the expression for T_{ij} the gives

$$T_{ij}=\left(\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right)_{ij}+\frac{k}{2}\left(\frac{\partial^2 U}{\partial t^2}\right)_{ij}-\frac{h^2}{12}\left(\frac{\partial^4 U}{\partial x^4}\right)_{ij}+\frac{k^2}{6}\left(\frac{\partial^3 U}{\partial t^3}\right)_{ij}-\cdots$$
$$-\frac{h^4}{360}\frac{\partial^6 U}{\partial x^4}+\cdots$$

But U is the solution of the differential eq. so

$$\left(\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right)_{ij}=0.$$

Therefore the principal part of the local truncation error is

$$\left(\frac{1}{2} k \frac{\partial^2 U}{\partial t^2} - \frac{1}{12} h^2 \frac{\partial^4 U}{\partial x^4} \right)_{i,j}.$$

Hence

$$T_{i,j} = O(k) + O(h^2).$$

when $k = rh^2$, $0 < r \leq \frac{1}{2}$, $T_{i,j}$ is $O(k)$ or $O(h^2)$, as one would expect by equations

$$\frac{\partial^2 U}{\partial x^2} \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}, \quad \frac{\partial U}{\partial t} \approx \frac{U_{i,j+1} - U_{i,j}}{k}.$$

This error may be further reduced by choosing a special value for $\frac{k}{h^2}$ because the eq. for $T_{i,j}$ can be written as

$$T_{i,j} = \frac{1}{12} h^2 \left(6 \frac{k}{h^2} \frac{\partial^2 U}{\partial t^2} - \frac{\partial^4 U}{\partial x^4} \right)_{i,j} + O(k^2) + O(h^4)$$

By the differential eq.

$$\frac{\partial}{\partial t} = \frac{\partial^2}{\partial x^2}$$

So

$$\frac{\partial}{\partial t} \left(\frac{\partial U}{\partial t} \right) = \frac{\partial^2}{\partial x^2} \left(\frac{\partial^2 U}{\partial x^2} \right),$$

assuming that these derivatives exist. If we put $\frac{6k}{h^2} = 1$, the expression in the braces is then zero and $T_{i,j}$ is $O(k^2) + O(h^4)$. This is of little use in practice because $k = \frac{1}{6} h^2$ is very small for small h so the volume of arithmetic needed to advance the solution to a large time-level is substantial

Consistency or compatibility

It is sometimes possible to approx. a parabolic or hyperbolic eq. by a FD scheme that is stable but which has solution that converges to the solution of a different differential eq. as the mesh lengths tend to zero. Such a difference scheme is said to be inconsistent or ~~inconsistence~~ incompatible with the PDE.

Let $L(U)=0$ represent the PDE in the independent variables x and t, with exact solution U.

Let $F(u)=0$ represent the approximating FDE with exact solution u.

Let v be a continuous function of x and t with a sufficient number of continuous derivatives to enable $L(v)$ to be evaluated at the point (ih, jk).

Then the truncation error $T_{ij}(v)$ at the point (ih, jk) is defined by

$$T_{ij}(v) = F(v_{ij}) - L(v_{ij}). \qquad (1)$$

If $T_{ij}(v) \to 0$ as $h \to 0, k \to 0$, the difference eq. is said to be consistent or compatible with the PDE.

Most authors put $v=U$ because $L(U)=0$. It then follows by eq. (1) that

$$T_{ij}(U) = F(U_{ij}) \rightarrow \text{local truncation error}$$

$\Rightarrow$ Def: The difference eq. is then consistent if the limiting value of the local truncation error is zero as $h \to 0, k \to 0$.

Example 1. The eq.

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$$

is approximated at the point (ih, jk) by the difference eq.

$$\frac{U_{i,j+1} - U_{i,j-1}}{2k} - \frac{U_{i+1,j} - 2\{\theta U_{i,j+1} + (1-\theta)U_{i,j-1}\} + U_{i-1,j}}{h^2} = 0,$$

Show that the local truncation error at this point is

$$\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{h^2}{12}\frac{\partial^4 U}{\partial x^4} + (2\theta - 1)\frac{2k}{h^2}\frac{\partial U}{\partial t} + \frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2} + O\left(\frac{k^3}{h^2}, h^4, k^4\right)$$

Discuss the consistency of this scheme with the PDE when:

(i) $k = rh$ and (ii) $k = rh^2$,

where r is a positive constant and θ a variable parameter.

Expansion of the terms $U_{i,j+1}$, $U_{i,j-1}$, $U_{i+1,j}$, and $U_{i-1,j}$ about (ih, jk) by Taylor's series, and substitution into

$$T_{ij} = \frac{U_{i,j+1} - U_{i,j-1}}{2k} - \frac{U_{i+1,j} - 2\{\theta U_{i,j+1} + (1-\theta)U_{i,j-1}\} + U_{i-1,j}}{h^2}$$

$$U_{i,j\pm1} = U(x_i, t_j \pm k) = U_{ij} \pm k\frac{\partial U}{\partial t} + \frac{k^2}{2}\frac{\partial^2 U}{\partial t^2} \pm \frac{k^3}{6}\frac{\partial^3 U}{\partial t^3} + \frac{k^4}{24}\frac{\partial^4 U}{\partial t^4} \pm \cdots$$

$$U_{i\pm1,j} = U_{ij} \pm h\frac{\partial U}{\partial x} + \frac{h^2}{2}\frac{\partial^2 U}{\partial x^2} \pm \frac{h^3}{6}\frac{\partial^3 U}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 U}{\partial x^4} \pm \frac{h^5}{5!}\frac{\partial^5 U}{\partial x^5} + \cdots$$

leads to

$$T_{ij} = \frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{2U_{ij} + h^2\frac{\partial^2 U}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4 U}{\partial x^4} + \cdots - 2\theta\left(2k\frac{\partial U}{\partial t} + 2U_{ij}\right)}{h^2}$$

$$\left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{(i,j)} - \frac{+2k\frac{\partial U}{\partial t} - k^2\frac{\partial^2 U}{\partial t^2}}{h^2} + O\left(\frac{k^3}{h^2}, h^4, k^4\right)$$

$$T_{ij} = \left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{(i,j)} + \left\{\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{h^2}{12}\frac{\partial^4 U}{\partial x^4}\right.$$

$$\left. + (2\theta - 1)\frac{2k}{h^2}\frac{\partial U}{\partial t} + \frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2}\right\} + O\left(\frac{k^3}{h^2}, h^4, k^4\right).$$

Case (i) $k = rh$

As $h \to 0$,

$$T_{ij} = F(U_{ij}) \to \left\{\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + (2\theta - 1)\frac{2r}{h}\frac{\partial U}{\partial t} + r^2\frac{\partial^2 U}{\partial t^2}\right\}_{(i,j)}$$

+ ~~O(...)~~

when $\theta \neq \frac{1}{2}$ the third term tends to infinity.

when $\theta = \frac{1}{2}$ the limiting value of T_{ij} is

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + r^2\frac{\partial^2 U}{\partial t^2}.$$

In this case the FD eq. is consistent with the hyperbolic equation

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + r^2\frac{\partial^2 U}{\partial t^2} = 0$$

Hence the difference equation is always inconsistent with $\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$ when $k = rh$.

Case (ii) $k = rh^2$.

As $h \to 0$

$$T_{ij} \to \frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} + 2(2\theta - 1)r\frac{\partial U}{\partial t}$$

when $\theta \neq \frac{1}{2}$ the difference scheme is consistent with the parabolic eq.

$$\{1 + 2(2\theta - 1)r\}\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0.$$

It is only when $\theta = \frac{1}{2}$ that the difference scheme is consistent with the given differential equation. This is then well-known Du Fort and Frankel three-level explicit scheme which is also stable for all $r > 0$.

The Richardson explicit scheme

$$\frac{U_{i,j+1} - U_{i,j-1}}{2k} - \frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} = 0.$$

has a local truncation error of $O(k^2) + O(h^2)$ but it is unconditionally ~~instability~~

Analytical treatment of Convergence (Direct proof).

Consider the eq.

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0<x<1, \qquad (1)$$

where U is known for $0<x<1$ when $t=0$, and at $x=0$ and 1 when $t>0$.

The simplest explicit finite-difference approximation to (1) is

$$\frac{u_{i,j+1}-u_{ij}}{k} = \frac{u_{i-1,j}-2u_{ij}+u_{i+1,j}}{h^2} \qquad (2)$$

Let $e=U-u$, and at the mesh points.

$$u_{ij} = U_{ij} - e_{ij}, \quad u_{i,j+1} = U_{i,j+1} - e_{i,j+1}, \text{ etc.}$$

Substitution into (2) leads to

$$u_{i,j+1} = ru_{i-1,j} + (1-2r)u_{ij} + ru_{i+1,j}$$

$$U_{i,j+1} - e_{i,j+1} = rU_{i-1,j} - re_{i-1,j} + (1-2r)(U_{ij}-e_{ij}) + r(U_{i+1,j}-e_{i+1,j})$$

$$\rightarrow \quad e_{i,j+1} = re_{i-1,j} + (1-2r)e_{ij} + re_{i+1,j} + U_{i,j+1} - U_{ij} + r(2U_{ij} - U_{i-1,j} - U_{i+1,j}) \qquad (3)$$

By Taylor's theorem

$$U_{i+1,j} = U(x_i+h, t_j) = U_{ij} + h\left(\frac{\partial U}{\partial x}\right)_{ij} + \frac{h^2}{2}\frac{\partial^2 U(x_i+\theta_1 h, t_j)}{\partial x^2},$$

$$U_{i-1,j} = U(x_i-h, t_j) = U_{ij} - h\left(\frac{\partial U}{\partial x}\right)_{ij} + \frac{h^2}{2}\frac{\partial^2 U(x_i-\theta_2 h, t_j)}{\partial x^2},$$

$$U_{i,j+1} = U(x_i, t_j+k) = U_{ij} + k\frac{\partial U(x_i, t_j+\theta_3 k)}{\partial t}.$$

where $0<\theta_1<1$, $0<\theta_2<1$ and $0<\theta_3<1$.

$$\Rightarrow \quad e_{i,j+1} = r e_{i-1,j} + (1-2r) e_{ij} + r e_{i+1,j} + \frac{}{}$$

$$+ k \left\{ \frac{\partial U(x_i, t_j + \theta_3 k}{\partial t} - \frac{\partial^2 U(x_i + \theta_4 h, t_j)}{\partial x^2} \right\}, \qquad (4)$$

where $-1 < \theta_4 < 1$.

Let $E_j = \max_i |e_{ij}|$, $M = \max_{i,j} |\{\quad\}|$. When $r \le \frac{1}{2}$, $\Rightarrow 1-2r > 0$, so

$$|e_{i,j+1}| \le r|e_{i-1,j}| + (1-2r)|e_{ij}| + r|e_{i+1,j}| + kM$$

$$\le E_j + kM \quad \text{for all values of } i \Rightarrow \text{this is true for } \max_i |e_{i,j+1}|.$$

Hence

$$E_{j+1} \le E_j + kM \le (E_{j-1} + kM) + kM \le E_{j-2} + 3kM$$

$$\le E_{j-3} + 4kM \le \cdots \le$$

$$E_j \le E_{j-1} + kM \le E_{j-2} + 2kM \le E_0 + jkM = tM,$$

because the initial values for u and U are the same, i.e., $E_0 = 0$, when h tends to zero, $k = rh^2$ also tends to zero and M tends to

$$\left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} \right)_{i,j} \Rightarrow 0$$

$\Rightarrow E_j \Rightarrow 0$. As $|U_{ij} - u_{ij}| \le E_j$, this proves that u converges to U as h tends to zero when $r \le \frac{1}{2}$ and t is finite.

~~When $r > \frac{1}{2}$ it can be shown that the~~

Analytic treatment of stability

There are two standard ways of investigating the boundedness of the solution of the finite-difference equation.

(1) Matrix method
(2) The Fourier series method (Von Neumann's method)

Matrix method

Explicit finite-difference scheme

Consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1,$$

where $U = 0$ at $x = 0$ and 1, $t > 0$, and U is known when $t = 0$.

The explicit finite-difference approximation

$$\frac{u_{i,j+1} - u_{ij}}{k} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2}$$

$$\Rightarrow \quad u_{i,j+1} = r u_{i-1,j} + (1-2r) u_{ij} + r u_{i+1,j} \qquad (1)$$

leads to the eq.s

$$i=1, \quad u_{1,j+1} = 0 + (1-2r)u_{1j} + r u_{2j}$$

$$i=2, \quad u_{2,j+1} = r u_{1j} + (1-2r)u_{2j} + r u_{3j}$$

$$\vdots$$

$$i=N-1, \quad u_{N-1,j+1} = r u_{N-2,j} + (1-2r)u_{N-1,j} + 0,$$

where $Nh = 1$, and $u_{0,j} = u_{N,j} = 0$. These can be written as

$$\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ \vdots \\ u_{N-1,j+1} \end{bmatrix} = \begin{bmatrix} (1-2r) & r & & & \\ r & (1-2r) & r & & \\ & r & (1-2r) & r & \\ & & \ddots & \ddots & \ddots \\ & & & r & (1-2r) \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ u_{3,j} \\ \vdots \\ u_{N-1,j} \end{bmatrix}$$

$$\vec{u}_{j+1} \qquad = \qquad A \qquad \vec{u}_j$$

or as

$$\vec{u}_{j+1} = A\vec{u}_j .$$

Hence

$$\vec{u}_j = A\vec{u}_{j-1} = A(A\vec{u}_{j-2}) = A^2\vec{u}_{j-2} = A^3\vec{u}_{j-3} = \cdots = A^j\vec{u}_0 ,$$

where $\vec{u}_0$ is the vector of initial values.

Now suppose we introduce errors at every pivotal point along $t=0$ and start the computation with the vector of values $\vec{u}_0^*$ instead of $\vec{u}_0$. We shall then calculate

$$\vec{u}_1^* = A\vec{u}_0^* , \quad \vec{u}_2^* = A\vec{u}_1^* = A^2\vec{u}_0^* ,$$

and finally

$$\vec{u}_j^* = A^j\vec{u}_0^* ,$$

assuming we introduce no further errors.

Define the error vector $\vec{e}$ by

$$\vec{e} = \vec{u} - \vec{u}^* .$$

Then

$$\vec{e}_j = \vec{u}_j - \vec{u}_j^* = A^j(\vec{u}_0 - \vec{u}_0^*) = A^j\vec{e}_0 ,$$

showing that the formula for the propagation of the errors is the same as that for the calculation of $\vec{u}$.

The finite difference scheme will be stable when $\vec{e}_j$ remains bounded as j increases indefinitely. This can always be investigated by expressing the error vector in terms of the eigenvectors of A.

Since A is real and symmetric, the eigenvalues λ are all distinct. Then A has $(N-1)$ LI eigenvectors $\vec{v}_s$. So these eigenvectors can be used as a basis for our $(N-1)$-dimensional vector space and the error $\vec{e}_0$, with its $(N-1)$ components, can be expressed uniquely as a linear combination of them, namely,

$$\vec{e}_0 = \sum_{s=1}^{N-1} c_s \vec{v}_s ,$$

where the c_s, $s = 1(1)(N-1)$, are known scalars.

The errors along the time-level $t = k$, resulting from the initial errors $\vec{e}_0$ will be given by

$$\vec{e}_1 = A\vec{e}_0 = A \sum c_s \vec{v}_s = \sum c_s A \vec{v}_s = \sum c_s \lambda_s \vec{v}_s$$

($\vec{A}\vec{v}_s = \lambda_s \vec{v}_s$)

Similarly, $\left(\vec{e}_2 = A \sum c_s \lambda_s \vec{v}_s = \sum c_s \lambda_s A \vec{v}_s = \sum c_s \lambda_s^2 \vec{v}_s\right)$ ($A\vec{v}_s = \lambda \vec{v}_s$)

$$\vec{e}_j = \sum c_s \lambda_s^j \vec{v}_s \qquad (2)$$

This shows that the errors will not increase exponentially with j provided the eigenvalue with the largest modulus has a modulus less than or equal to unity.

The matrix A can be written as

$$\begin{bmatrix} (1-2r) & r & & & \\ r & (1-2r) & r & & \\ & \ddots & \ddots & \ddots & \\ & & r & (1-2r) & r \\ & & & r & (1-2r) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \end{bmatrix}$$

$$+ r \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}$$

i.e., as

$$A = I + r T_{N-1}$$

where T_{N-1} is an $(N-1)\times(N-1)$ matrix whose eigenvalues λ_s and eigenvectors $\vec{v}_s$ are given by

$$\lambda_s = -4\sin^2\left(\frac{s\pi}{2N}\right), \quad (s=1,2,\dots,N-1),$$

$$\vec{v}_s = \left(\sin\frac{s\pi}{N}, \sin\frac{2s\pi}{N}, \dots, \sin\frac{(N-1)s\pi}{N}\right)$$

The eigenvalues and vectors of a common tridiagonal matrix.

Let

$$A = \begin{bmatrix} a & b & & & \\ c & a & b & & \\ & c & a & b & \\ & & \ddots & \ddots & \ddots \\ & & & c & a \end{bmatrix}$$

be a square matrix of order N, where a, b, and c may be real or complex numbers.

$$A\vec{v} = \lambda\vec{v}, \quad \vec{v} = (v_1, v_2, \dots, v_N)$$

$$\begin{bmatrix} a & b & & & \\ c & a & b & & \\ & c & a & b & \\ & & \ddots & \ddots & \ddots \\ & & & c & a \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_N \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_N \end{bmatrix} \Rightarrow$$

$$\begin{aligned} av_1 + bv_2 &= \lambda v_1 \\ cv_1 + av_2 + bv_3 &= \lambda v_2 \\ cv_2 + av_3 + bv_4 &= \lambda v_3 \\ &\vdots \\ cv_{N-1} + av_N &= \lambda v_N \end{aligned}$$

$$\begin{aligned} (a-\lambda)v_1 + bv_2 &= 0 \\ cv_1 + (a-\lambda)v_2 + bv_3 &= 0 \\ &\cdots \\ cv_{j-1} + (a-\lambda)v_j + bv_{j+1} &= 0 \\ &\cdots \\ cv_{N-1} + (a-\lambda)v_N &= 0 \end{aligned}$$

If we define $v_0 = v_{N+1} = 0$ then these N eq.s can be combined into the single difference equation.

$$c v_{j-1} + (a-\lambda) v_j + b v_{j+1} = 0, \quad j = 1(1)N. \tag{3}$$

$\underline{v_j = m^j \neq 0}$

$$c m^{j-1} + (a-\lambda) m^j + b m^{j+1} = 0$$

$$\Rightarrow \quad c + (a-\lambda) m + b m^2 = 0 \tag{3'}$$

If $m_1 \neq m_2$ (are real), then

$$v_j = B m_1^j + C m_2^j$$

is the solution of (3'). Since $v_0 = v_{N+1} = 0$, then

$$\begin{cases} 0 = B + C & \longrightarrow \quad B = -C \\ 0 = B m_1^{N+1} + C m_2^{N+1} & \Rightarrow \quad 0 = B m_1^{N+1} - B m_2^{N+1} \end{cases}$$

Hence $$\left(\frac{m_1}{m_2}\right)^{N+1} = 1 = e^{i 2 s \pi}, \quad s = 1(1)N, \quad i = \sqrt{-1}.$$

$$\Rightarrow \quad \frac{m_1}{m_2} = e^{\frac{i 2 s \pi}{N+1}} \tag{4}$$

By eq. (3')

$$m_1 m_2 = \frac{c}{b}, \tag{5}$$

and elimination of m_2 between (4) and (5) leads to

$$m_2 = \frac{c}{b m_1}, \quad \frac{m_1}{\frac{c}{b m_1}} = e^{\frac{i 2 s \pi}{N+1}} \Rightarrow m_1^2 = \frac{c}{b} e^{\frac{i 2 s \pi}{N+1}}$$

$$\Rightarrow \quad m_1 = \left(\frac{c}{b}\right)^{\frac{1}{2}} e^{\frac{i s \pi}{N+1}}.$$

$$m_2 = \frac{c}{b\left(\frac{c}{b}\right)^{\frac{1}{2}} e^{\frac{i s \pi}{N+1}}} = \left(\frac{c}{b}\right)^{\frac{1}{2}} e^{-\frac{i s \pi}{N+1}}.$$

Again, by (3'),

$$u_1+u_2=\frac{\lambda-a}{b} \quad \overset{\lambda=}{\Rightarrow} a+b(u_1+u_2)$$

$$\Rightarrow \lambda = a+b\left(\frac{c}{b}\right)^{\frac{1}{2}}\left(e^{\frac{is\pi}{N+1}}+e^{-\frac{is\pi}{N+1}}\right)$$

Hence the N eigenvalues are given by

$$\lambda_s = a+2b\left(\frac{c}{b}\right)^{\frac{1}{2}}\cos\frac{s\pi}{N+1}, \quad s=1\,(1)\,N.$$

$$\boxed{\lambda_s = a+2\sqrt{bc}\,\cos\frac{s\pi}{N+1}, \quad s=1\,(1)\,N}$$

The j-th component of the eigenvector is

$$v_j = Bu_1^j+Cu_2^j = B\left(\frac{c}{b}\right)^{\frac{1}{2}j}\left(e^{\frac{ijs\pi}{N+1}}-e^{-\frac{ijs\pi}{N+1}}\right)$$

$$= 2iB\left(\frac{c}{b}\right)^{\frac{1}{2}j}\sin\frac{js\pi}{N+1}.$$

So the eigenvector $\vec{v}_s$ corresponding to λ_s can be taken as

$$\vec{v}_s^{\,T} = \left\{\left(\frac{c}{b}\right)^{\frac{1}{2}}\sin\frac{s\pi}{N+1}, \frac{c}{b}\sin\frac{2s\pi}{N+1}, \left(\frac{c}{b}\right)^{\frac{3}{2}}\sin\frac{3s\pi}{N+1}, \ldots, \left(\frac{c}{b}\right)^{\frac{N}{2}}\sin\frac{Ns\pi}{N+1}\right\}.$$

It is easily shown that the roots (3') cannot be equal because if we assume $u_1=u_2$ the solution (3') is then

$$v_j = (B+Cj)u_1^j$$

and $v_0=v_{N+1}=0 \Rightarrow B=C=0$, giving $\vec{v}=0$, which is not possible.

$a=-2, \; b=1, \; c=1$

$$\Rightarrow \lambda_s = -2+2\left(\frac{1}{1}\right)^{\frac{1}{2}}\cos\frac{s\pi}{N} = -2\left(1-\cos\frac{s\pi}{N}\right) = -4\sin^2\frac{s\pi}{2N}$$

$$\Rightarrow \boxed{\lambda_s = -4\sin^2\frac{s\pi}{2N}} \quad (s=1,2,\ldots,N-1)$$

$$\vec{v}_s = \left\{\sin\frac{s\pi}{N}, \sin\frac{2s\pi}{N}, \ldots, \sin\frac{(N-1)s\pi}{N}\right\}$$

A note on eigenvalues and eigenvectors

Let $\vec{x}$ be an eigenvector of A corresponding to the eigenvalue λ. Then $A\vec{x} = \lambda\vec{x}$. Hence

$A(A\vec{x}) = A^2\vec{x} = \lambda A\vec{x} = \lambda^2\vec{x}$,

$\Rightarrow$ A^2 has an eigenvalue λ^2 corresponding to the eigenvector $\vec{x}$. Similarly, $A^p\vec{x} = \lambda^p\vec{x}$, $p = 3, 4, \ldots$

(i) If $f(A) = a_p A^p + a_{p-1}A^{p-1} + \ldots + a_0 I$ is a polynomial in A with scalar coefficients $a_p, \ldots, a_0$, then

$f(A)\vec{x} = (a_p\lambda^p + \ldots + a_0)\vec{x} = f(\lambda)\vec{x}$

$\Rightarrow$ $f(A)$ has an eigenvalue $f(\lambda)$ corresponding to the eigenvector $\vec{x}$.

(ii) The eigenvalue of $[f_1(A)]^{-1} f_2(A)$ corresponding to the eigenvector $\vec{x}$ is $\frac{f_2(\lambda)}{f_1(\lambda)}$, where $f_1(A)$ and $f_2(A)$ are polynomial in A.

Proof

$$f_1(A)\vec{x} = f_1(\lambda)\vec{x}$$
$$f_2(A)\vec{x} = f_2(\lambda)\vec{x}$$

Premultiply both eq.s by $[f_1(A)]^{-1}$ and write as

$[f_1(A)]^{-1}\vec{x} = \frac{\vec{x}}{f_1(\lambda)}$ and $[f_1(A)]^{-1} f_2(A)\vec{x} = f_2(\lambda)[f_1(A)]^{-1}x$

$\Rightarrow [f_1(A)]^{-1} f_2(A)\vec{x} = f_2(\lambda)\cdot\frac{1}{f_1(\lambda)}\vec{x} = \frac{f_2(\lambda)}{f_1(\lambda)}\vec{x}$

Hence the eigenvalues of A, ~~as shown later~~ are

$$1 + r\left\{-4\sin^2\frac{s\pi}{2N}\right\}.$$

Therefore the condition for stability of the explicit scheme is

$$\left| 1 - 4r\sin^2\frac{s\pi}{2N} \right| \leq 1.$$

$$\Rightarrow \quad -1 \leq 1 - 4r\sin^2\frac{s\pi}{2N} \leq 1,$$

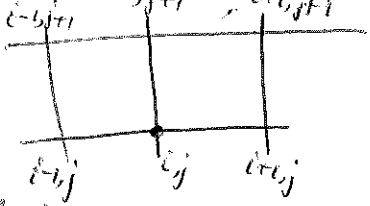
$$\Rightarrow \quad -1 \leq 1 - 4r\sin^2\frac{s\pi}{2N} \Rightarrow 4r\sin^2\frac{s\pi}{2N} \leq 2$$

$$\Rightarrow r \leq \frac{1}{2\sin^2\frac{s\pi}{2N}} > \frac{1}{2}$$

proving that the scheme is stable for $r \leq \frac{1}{2}$.

Crank-Nicolson method for $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$

$$\frac{u_{i,j+1} - u_{ij}}{\delta t} = \frac{1}{2}\left\{ \frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{(\delta x)^2} + \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{(\delta x)^2} \right\}$$



$$\Rightarrow 2u_{i,j+1} - 2u_{ij} = r\left(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}\right) + r\left(u_{i+1,j} - 2u_{ij} + u_{i-1,j}\right)$$

$$-r u_{i-1,j+1} + (2+2r)u_{i,j+1} - r u_{i+1,j+1} = r u_{i-1,j} + (2-2r)u_{ij} + r u_{i+1,j}$$

$$\Rightarrow \begin{bmatrix} (2+2r) & -r & & & & \\ -r & (2+2r) & -r & & & \\ & \ddots & \ddots & \ddots & & \\ & & -r & (2+2r) & -r & \\ & & & -r & (2+2r) & \end{bmatrix} \begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \vdots \\ u_{N-1,j+1} \end{bmatrix}$$

$$= \begin{bmatrix} (2-2r) & r & & & & \\ r & (2-2r) & r & & & \\ & \ddots & \ddots & \ddots & & \\ & & r & (2-2r) & r & \\ & & & r & (2-2r & \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{N-1,j} \end{bmatrix}$$

$$\Rightarrow (2I - rT_{N-1})\vec{u}_{j+1} = (2I + rT_{N-1})\vec{u}_j$$

$$\Rightarrow \vec{u}_{j+1} = (2I - rT_{N-1})^{-1}(2I + rT_{N-1})u_j$$

By the previous argument these finite-difference eq.s will be stable when the moduli of the eigenvalues of

$$A = (2I - rT_{N-1})^{-1}(2I + rT_{N-1})$$

are each less than one. As the eigenvalues of T_{N-1} are $-4\sin^2\left(\frac{s\pi}{2N}\right)$, the eigenvalues of A are

$$\frac{2 - 4r\sin^2\left(\frac{s\pi}{2N}\right)}{2 + 4r\sin^2\left(\frac{s\pi}{2N}\right)}, \quad s = 1, 2, \ldots, N-1$$

and these are clearly less than one for all positive values of r proving that the Crank-Nicholson eq.s have unrestricted stability.

Useful theorems on bounds for eigenvalues

Gerschgorin's first theorem

The largest of the moduli of the eigenvalues of the square matrix A cannot exceed the largest sum of the moduli of the elements along any row or any column.

Proof.

Let λ_i be an eigenvalue of the $N \times N$ matrix A, and $\vec{x}_i$ the corresponding eigenvector with components $v_1, v_2, \dots, v_n$. Then the eq.

$$A\vec{x}_i = \lambda_i \vec{x}_i$$

in detail, is

$$a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n = \lambda_i v_1,$$

$$a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n = \lambda_i v_2,$$

$$\vdots$$

$$a_{s,1}v_1 + a_{s2}v_2 + \cdots + a_{s,n}v_n = \lambda_i v_s,$$

$$\vdots$$

Let v_s be the largest in modulus of $v_1, v_2, \dots, v_n$. Select the s-th equation and divide by v_s, giving

$$\lambda_i = a_{s,1}\left(\frac{v_1}{v_s}\right) + a_{s2}\left(\frac{v_2}{v_s}\right) + \cdots + a_{s,n}\left(\frac{v_n}{v_s}\right).$$

Therefore

$$|\lambda_i| \le |a_{s,1}| + |a_{s,2}| + \cdots + |a_{s,n}|,$$

because $\left|\frac{v_i}{v_s}\right| \le 1$, $i = 1, 2, \dots, n$.

In particular this holds for $|\lambda_i| = \max_s |\lambda_s|$, $s = 1(1)N$.

Since the eigenvalues of the transpose of A are the same as those of A the theorem is also true for columns.

Gerschgorin's circle theorem or Brauer's theorem

Let P_s be the sum of the moduli of the elements along the s-th row excluding the diagonal element $a_{s,s}$.

Then each eigenvalue of A lies inside or on the boundary of at least one of the circles $|\lambda - a_{s,s}| = P_s$.

Proof

Since

$$\lambda_i = a_{s1}\left(\frac{v_1}{v_s}\right) + a_{s2}\left(\frac{v_2}{v_s}\right) + \cdots + a_{ss} + \cdots + a_{s,n}\left(\frac{v_n}{v_s}\right),$$

Hence

$$|\lambda_i - a_{ss}| = \left| a_{s1}\left(\frac{v_1}{v_s}\right) + \cdots + 0 + \cdots + a_{s,n}\left(\frac{v_N}{v_s}\right)\right|$$

$$\leq |a_{s,1}| + |a_{s,2}| + \cdots + 0 + \cdots + |a_{s,n}| = P_s$$

which completes the proof.

As an illustrative example consider the Crank-Nicolson eq.s

$$(2I - rT_{N-1})\vec{U}_{j+1} = (2I + rT_{N-1})\vec{U}_j = \{4I - (2I - rT_{N-1})\}\vec{U}_j$$

which can be written as

$$B\vec{u}_{j+1} = (4I - B)\vec{u}_j \,.$$

giving

$$\vec{u}_{j+1} = (4B^{-1} - I)\vec{u}_j \,,$$

where

$$B = \begin{bmatrix} (2+2r) & -r & & & \\ -r & (2+2r) & -r & & \\ & \ddots & \ddots & \ddots & \\ & & -r & (2+2r) & -r \\ & & & -r & (2+2r) \end{bmatrix}$$

The finite-difference eq.s will be stable when the modulus of every eigenvalue of $(4B^{-1} - I)$ does not exceed one, that is, when

$$\left| \frac{4}{\lambda} - 1 \right| \leq 1 \,,$$

where λ is an eigenvalue of B.

$$\Rightarrow \quad -1 \leq \frac{4}{\lambda} - 1 \leq 1 \quad \Rightarrow \quad \underline{\lambda \geq 2} \,.$$

For the matrix B, $a_{ss} = 2+2r$, $\max P_s = 2r$, so Brauer's theorem leads to

$$|\lambda - 2 - 2r| \leq 2r \Rightarrow -2r \leq \lambda - 2 - 2r \leq 2r$$

$$\Rightarrow 2 \leq \lambda \leq 2 + 4r \Rightarrow \underline{\underline{\lambda \geq 2}} \text{ for all value of } r.$$

Stability criteria for derivative boundary conditions

Consider the eq.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1,$$

and the cond.s

$$\frac{\partial u}{\partial x} = h_1(u - v_1) \quad \text{at} \quad x = 0, \ t \geq 0,$$

$$\frac{\partial u}{\partial x} = -h_2(u - v_2) \quad \text{at} \quad x = 1, \ t \geq 0.$$

where h_1, h_2, v_1, v_2 are constants, $h_1 \geq 0$, $h_2 \geq 0$.

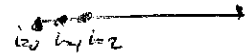
When the b.c.s are approximated by the central-difference eq.s

$$\frac{u_{1,j} - u_{-1,j}}{2\delta x} = h_1(u_{0,j} - v_1), \quad \Rightarrow u_{-1,j} = u_{1,j} - 2h_1 \delta x\, u_{0,j} + 2h_1 \delta x\, v_1$$

$$\frac{u_{N+1,j} - u_{N-1,j}}{2\delta x} = -h_2(u_{N,j} - v_2), \quad (N\delta x = 1). \Rightarrow u_{N+1,j} = u_{N-1,j} - 2h_2 \delta x\, u_{N,j} + 2h_2 \delta x\, v_2$$

and the differential eq. by the explicit scheme

$$u_{i,j+1} = r u_{i-1,j} + (1-2r) u_{i,j} + r u_{i+1,j},$$



elimination of $u_{-1,j}, u_{N+1,j}$, leads to the eq.s

$i = 0$:

$$u_{0,j+1} = r u_{-1,j} + (1-2r) u_{0,j} + r u_{1,j}, \Leftrightarrow u_{-1,j} = u_{1,j} - 2\delta x\, h_1 u_{0,j} + 2\delta x h_1 v_1$$

$$\Rightarrow u_{0,j+1} = r u_{1,j} - 2r h_1 \delta x\, u_{0,j} + 2r h_1 \delta x\, v_1 + (1-2r) u_{0,j} + r u_{1,j}$$

$$\boxed{u_{0,j+1} = 2r u_{1,j} + \{1 - 2r(1 + h_1 \delta x)\} u_{0,j} + 2r h_1 \delta x\, v_1}$$

$i = N$:

$$u_{N,j+1} = r u_{N-1,j} + (1-2r) u_{N,j} + r u_{N+1,j}$$

$$u_{N,j+1} = r u_{N-1,j} - 2r h_2 \delta x\, u_{N,j} + 2r h_2 \delta x\, v_2 + r u_{N-1,j} + (1-2r) u_{N,j}$$

$$\boxed{u_{N,j+1} = 2r u_{N-1,j} + \{1 - 2r(1 + h_2 \delta x)\} u_{N,j} + 2r h_2 \delta x\, v_2}$$

$$\Rightarrow \begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ u_{2,j+1} \\ \vdots \\ u_{N-1,j+1} \\ u_{N,j+1} \end{bmatrix} = \begin{bmatrix} \{1-2r(1+h_1^{a,h}\delta x)\} & 2r & & & \\ r & (1-2r) & r & & \\ & r & \ddots & \ddots & \\ & & r & (1-2r) & r \\ & & & 2r & \{1-2r(1+h_2^{a,h}\delta x)\} \end{bmatrix}$$

$$\times \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{N-1,j} \\ u_{N,j} \end{bmatrix} + \begin{bmatrix} 2r h_1^{a_1,b_1,h} v_1 \delta x \\ 0 \\ 0 \\ \vdots \\ 0 \\ 2r h_1^{a_2,b_2,h} v_2 \delta x \end{bmatrix}$$

As each component of the last column vector is a constant the matrix determining the propagation of the error is

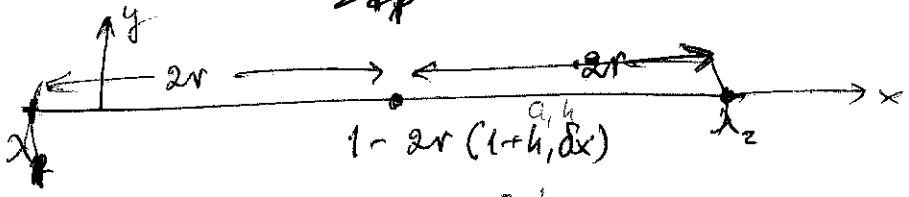
$$\begin{bmatrix} \{1-2r(1+h_1^{a,h}\delta x)\} & 2r & & & \\ r & (1-2r) & r & & \\ & r & \ddots & \ddots & \\ & & r & (1-2r) & r \\ & & & 2r & \{1-2r(1+h_2^{a,h}\delta x)\} \end{bmatrix}$$

Application of Brauer's theorem to this matrix, with

$$a_{ss} = 1-2r(1+h_1^{a,h}\delta x) \quad \text{and} \quad P_s = 2r,$$

shows that its eigenvalues λ lie on or within the circle

$$|\lambda - \{1-2r(1+h_1^{a,h}\delta x)\}| \leq 2r$$



$$-2r \le \lambda - \{1-2r(1+h_{1,\delta x}^{a,h})\} \le 2r$$

$$1-2r-2rh_{1,\delta x}^{a,h}-2r \le \lambda \le 2r+1-2r-2rh_{1,\delta x}^{a,h}$$

$$1-4r-2rh_{1,\delta x}^{a,h}$$

$$\Rightarrow \lambda_1 = 1-2r(2+h_{1,\delta x}^{a,h}), \quad \lambda_2 = 1-2rh_{1,\delta x}^{a,h}$$

and for stability,

$$|\lambda_1| \le 1, \quad |\lambda_2| \le 1.$$

Hence ↓

$$-1 \le 1-2r(2+h_{1,\delta x}^{a,h}) \le 1, \text{ giving}$$

$$2r(2+h_{1,\delta x}^{a,h}) \le 2 \Rightarrow \boxed{r \le \frac{1}{2+h_{1,\delta x}^{a,h}}},$$

and

$$|1-2rh_{1,\delta x}^{a,h}| \le 1 \Rightarrow -1 \le 1-2rh_{1,\delta x}^{a,h}$$

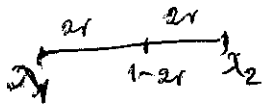
$$2rh_{1,\delta x}^{a,h} \le 2 \Rightarrow \boxed{r_1 \le \frac{1}{h_{1,\delta x}^{a,h}}}$$

The least of these is

$$\boxed{r \le \frac{1}{2+h_{1,\delta x}^{a,h}}}.$$

Similarly we require $\boxed{r \le \frac{1}{2+h_{2,\delta x}^{a_2,h}}}$.

For row 2 (1)(N-1) $a_{ss} = 1-2r$, $P_s = 2r$, giving



$$|\lambda-(1-2r)| \le 2r$$

$$\Rightarrow -2r \le \lambda-(1-2r) \le 2r \Rightarrow -2r \le \lambda-1+2r \le 2r$$

$$1-2r-2r \le \lambda \le 2r+1-2r$$

$$\underbrace{1-4r}_{\lambda_1} \le \lambda \le \underbrace{1}_{\lambda_2}$$

$$|\lambda_1| \le 1 \rightarrow |1-4r| \le 1$$

$$-1 \le 1-4r \le 1$$

$$4r \le 2 \rightarrow \boxed{r \le \frac{1}{2}}$$

For overall stability,

$$r \le \min\left\{ \frac{1}{2+\underset{a_1 h}{h_1 \delta x}} \; ; \; \frac{1}{2+\underset{a_2 h}{h_2 \delta x}} \right\}$$

Crank–Nicolson equations

$$-r u_{i-1,j+1} + (2+2r) u_{i,j+1} - r u_{i+1,j+1} = r u_{i-1,j} + (2-2r) u_{ij} + r u_{i+1,j}$$

$$\begin{cases} u_{-1,j} = u_{1,j} - 2\underset{a_1 h}{h_1 \delta x} u_{0,j} + 2\underset{a_1 h}{h_1 \delta x} \underset{b_1}{v_1}, \\ u_{N+1,j} = u_{N-1,j} - 2\underset{a_2 h}{h_2 \delta x} u_{N,j} + 2\underset{a_2 h}{h_2 \delta x} \underset{b_2}{v_2}, \end{cases} \quad \begin{cases} u_{-1,j+1} = u_{1,j+1} - 2\underset{a_1 h}{h_1 \delta x} u_{0,j+1} + 2\underset{a_1 h}{h_1 \delta x} \underset{b_1}{v_1}, \\ u_{N+1,j+1} = u_{N-1,j+1} - 2\underset{a_2 h}{h_2 \delta x} u_{N,j+1} + 2\underset{a_2 h}{h_2 \delta x} \underset{b_2}{v_2}, \end{cases}$$

$i=0$:

$$-r u_{-1,j+1} + (2+2r) u_{0,j+1} - r u_{1,j+1} = r u_{-1,j} + (2-2r) u_{0,j} + r u_{1,j}$$

$$-r u_{1,j+1} + 2r h_1 \delta x\, u_{0,j+1} - 2r h_1 \delta x\, v_1 + (2+2r) u_{0,j+1} - r u_{1,j+1}$$

$$= r u_{1,j} - 2r h_1 \delta x\, u_{0,j} + 2r h_1 \delta x\, v_1 + (2-2r) u_{0,j} + r u_{1,j}$$

$$\boxed{\{2+2r(1+h_1\delta x)\} u_{0,j+1} - 2r u_{1,j+1} = \{2-2r(1+h_1\delta x)\} u_{0,j} + 2r u_{1,j} + 4r h_1 \delta x\, v_1}$$

$i=N$:

$$-r u_{N-1,j+1} + (2+2r) u_{N,j+1} - r u_{N+1,j+1} = r u_{N-1,j} + (2-2r) u_{N,j} + r u_{N+1,j}$$

$$-r u_{N-1,j+1} + (2+2r) u_{N,j+1} - r u_{N-1,j+1} + 2r h_2 \delta x\, u_{N,j+1} - 2r h_2 \delta x\, v_2$$

$$= r u_{N-1,j} + (2-2r) u_{N,j} + r u_{N-1,j} - 2r h_2 \delta x\, u_{N,j} + 2r h_2 \delta x\, v_2$$

$$\Rightarrow \boxed{\begin{aligned} &-2r u_{N-1,j+1} + \{2+2r(1+h_2\delta x)\} u_{N,j+1} = \\ &= 2r u_{N-1,j} + \{2-2r(1+h_2\delta x)\} u_{N,j} + 4r h_2 \delta x\, v_2 \end{aligned}}$$

$\Rightarrow$

$$\begin{bmatrix} \{2+2r(1+h_1^{a_1}\delta x)\} & -2r & & & & \\ -r & (2+2r) & -r & & & \\ & -r & (2+2r) & -r & & \\ & & \ddots & \ddots & \ddots & \\ & & & -r & (2+2r) & -r \\ & & & & 2r & \{2+2r(1+h_2^{a_2}\delta x)\} \end{bmatrix}$$

$$\times \begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ \vdots \\ u_{N-1,j+1} \\ u_{N,j+1} \end{bmatrix} = \begin{bmatrix} \{2-2r(1+h_1^{a_1}\delta x)\} & +2r & & & & \\ r & (2-2r) & r & & & \\ & r & (2-2r) & r & & \\ & & \ddots & \ddots & \ddots & \\ & & & r & (2-2r) & r \\ & & & & 2r & \{2-2r(1+h_2^{a_2}\delta x)\} \end{bmatrix}$$

$$\times \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ \vdots \\ u_{N-1,j} \\ u_{N,j} \end{bmatrix} + \begin{bmatrix} 4rh_1\delta x v_1 \\ 0 \\ \vdots \\ 0 \\ 4rh_2\delta x v_2 \end{bmatrix}$$

$\Rightarrow$ $B\vec{e}_{j+1} = (4I - B)\vec{e}_j$

$\vec{e}_{j+1} = (4B^{-1} - I)\vec{e}_j$

$\Rightarrow B\vec{u}_{j+1} = \{4I - B\}\vec{u}_j + \vec{g} \Rightarrow \vec{u}_{j+1} = (4B^{-1} - I)\vec{u}_j + B^{-1}\vec{g}$

$\Rightarrow \vec{e}_{j+1} = \{4B^{-1} - I\}\vec{e}_j$

By Gerschgorin's circle theorem

$$|\lambda - \{2+2r(1+h_1^{a_1}\delta x)\}| \le 2r$$

$$-2r + 2 + 2r(1+h_1^{a_1}\delta x) \le \lambda \le 2r + 2 + 2r(1+h_1\delta x)$$

$$2 + 2rh_1^{a_1}\delta x \le \lambda \le 2 + 4r + 2rh_1^{a_1}\delta x$$

$$\Rightarrow \lambda \ge 2 + 2rh_1^{a_1}\delta x \Rightarrow \boxed{\lambda \ge 2}$$

$$\left|\frac{4}{\lambda} - 1\right| \le 1$$

$$-1 \le \frac{4}{\lambda} - 1 \le 1$$

$$0 \le \frac{4}{\lambda} \le 2$$

$\boxed{\lambda \ge 2}$!

Similarly $\lambda \ge 2$ and $\lambda \ge 2 + 2rh_2^{a_2}\delta x$ for the remaining rows.

Hence the equations are unconditionally stable.

The stability of three or more time-level difference eq.s

The following theorem is useful

Theorem. If the matrix A can be written as

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,m} \\ \vdots & & & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,m} \end{bmatrix},$$

where each $A_{i,j}$ is an $n \times n$ matrix, and all the $A_{i,j}$ have a common set of n L.I. eigenvectors, then the eigenvalues of A are given by the eigenvalues of the matrices

$$\begin{bmatrix} \lambda_{1,1}^{(k)} & \lambda_{1,2}^{(k)} & \cdots & \lambda_{1,m}^{(k)} \\ \lambda_{2,1}^{(k)} & \lambda_{2,2}^{(k)} & \cdots & \lambda_{2,m}^{(k)} \\ \vdots & & & \vdots \\ \lambda_{m,1}^{(k)} & \lambda_{m,2}^{(k)} & \cdots & \lambda_{m,m}^{(k)} \end{bmatrix}, \quad k = 1(1)n,$$

where $\lambda_{i,j}^{(k)}$ is the k-th eigenvalues of $A_{i,j}$ corresponding to the k-th eigenvector $\vec{v}_k$ common to all the $A_{i,j}$'s.

Example. Investigate the stability of the and Francel approximation to the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 \le x \le 1,$$

given that U is known initially for $0 \le x \le$ and is known on the boundaries $x = 0$ and

The approximation at the (ih, jk) is

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

which may be written as

$$u_{i,j+1} - u_{i,j+1} = 2r(u_{i-1,j} + u_{i+1,j}) - 2r\, u_{i,j-1}$$

$$(1+2r)u_{i,j+1} = 2r(u_{i-1,j} + u_{i+1,j}) + (1-2r)u$$

where $r = \frac{k}{h^2}$. For known boundary values and equations in matrix form are

$$(1+2r)\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ \vdots \\ u_{N-1,j+1} \end{bmatrix} = 2r\begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & 0 & 1 & \\ & & & \ddots & \\ & & & 1 & 0 \end{bmatrix}\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}$$

$$+ (1-2r)\begin{bmatrix} u_{1,j-1} \\ u_{2,j-1} \\ u_{3,j-1} \\ \vdots \\ u_{N-1,j-1} \end{bmatrix} + 2r\begin{bmatrix} u_{0,j} \\ 0 \\ 0 \\ \vdots \\ u_{N,j} \end{bmatrix}$$

giving

$$\vec{U}_{j+1} = \frac{2r}{1+2r} A\vec{U}_j + \frac{1-2r}{1+2r}\vec{U}_{j-1} + \vec{C}_j$$

where A is as displayed and $\vec{C}_j$ is a vector of known values.

Put

$$\vec{v}_j = \begin{bmatrix} \vec{u}_j \\ \vec{u}_{j-1} \end{bmatrix}.$$

Then eq. (1) and the identity $\vec{u}_j = \vec{u}_j$ can be written as

$$\begin{bmatrix} \vec{u}_{j+1} \\ \hline \vec{u}_j \end{bmatrix} = \left[\begin{array}{cc} \frac{2r}{1+2r} A & \frac{(1-2r)}{(1+2r)} I_{N-1} \\ \hline I_{N-1} & 0 \end{array}\right] \begin{bmatrix} \vec{u}_j \\ \hline \vec{u}_{j-1} \end{bmatrix} + \begin{bmatrix} \vec{C}_j \\ \hline 0 \end{bmatrix}$$

where I_{N-1} is the unit matrix of order $(N-1)$,

i.e., as

$$\vec{V}_{j+1} = P\vec{V}_j + \vec{d}_j,$$

where P is the matrix shown and $\vec{d}_j$ a column vector of known constants. This technique has reduced a three-level difference equation to a two-level one. The equation will be stable when each eigenvalue of P has a modulus ≤ 1.

The matrix A has $(N-1)$ different eigenvalues so it has $(N-1)$ LI eigenvectors $\vec{v}_s$, $s = 1(1)(N-1)$. Although the matrix I_{N-1} has $(N-1)$ eigenvalues each equal to 1 it has $(N-1)$ LI eigenvectors which may be taken as $\vec{v}_s$, $s = 1(1)(N-1)$. ($I_{N-1}\vec{v}_s = 1\cdot\vec{v}_s$). Hence the eigenvalues λ of P are the eigenvalues of the matrix

$$\begin{bmatrix} \frac{2r\lambda_k}{1+2r} & \frac{1-2r}{1+2r} \\ 1 & 0 \end{bmatrix},$$

$\left[\lambda_s = a + 2b\left(\frac{c}{b}\right)^{\frac{1}{2}} \cos\frac{s\pi}{N+1}, \quad s = 1(1)N \;\; (a=0,\ c=b=1)\right]$

where λ_k is the k-th eigenvalue of A. $\left(\lambda_k = 2\cos\left(\frac{k\pi}{N}\right),\ k = 1(1)(N-1)\right)$

$$\Rightarrow \quad \det\begin{bmatrix} \left\{\frac{2r\lambda_k}{1+2r} - \lambda\right\} & \frac{1-2r}{1+2r} \\ 1 & -\lambda \end{bmatrix} = 0$$

$$\Rightarrow \quad \lambda^2 - \frac{2r\lambda_{ic}}{1+2r}\lambda - \frac{1-2r}{1+2r} = 0 .$$

Hence

$$\lambda^2 - \frac{2r \cdot 2\cos\frac{k\pi}{N}}{1+2r}\lambda - \frac{1-2r}{1+2r} = 0 ,$$

$$\lambda = 2r\cos\frac{k\pi}{N} \Big/ 1+2r \pm \sqrt{\frac{4r^2\cos^2\frac{k\pi}{N}}{(1+2r)^2} + \frac{1-2r}{1+2r}}$$

$$= \frac{2r\cos\frac{k\pi}{N}}{1+2r} \pm \sqrt{\frac{4r^2\cos^2\frac{k\pi}{N} + 1 - 4r^2}{(1+2r)^2}}$$

$$= \frac{2r\cos\frac{k\pi}{N}}{1+2r} \pm \frac{\sqrt{4r^2\left(\cos^2\frac{k\pi}{N} - 1\right) + 1}}{1+2r}$$

$$= \frac{2r\cos\frac{k\pi}{N}}{1+2r} \pm \frac{\sqrt{1 - 4r^2\sin^2\frac{k\pi}{N}}}{1+2r}$$

$$\lambda = \frac{2r\cos\frac{k\pi}{N} \pm \sqrt{1 - 4r^2\sin^2\frac{k\pi}{N}}}{1+2r}$$

Case (i) $\quad 1 > 1 - 4r\sin^2\frac{k\pi}{N} \geq 0$

Then

$$|\lambda| < \frac{2r+1}{1+2r} = 1$$

Case (ii) $\quad 1 - 4r^2\sin^2\frac{k\pi}{N} < 0 .$

Then

$$|\lambda|^2 = \frac{1}{(2r+1)^2}\left\{\left(2r\cos\frac{k\pi}{N}\right)^2 + 4r^2\sin^2\frac{k\pi}{N} - 1\right\}$$

$$= \frac{4r^2-1}{4r^2+4r+1} < 1 \quad \text{since} \quad r > 0 .$$

Therefore the equations are unconditionally stable for all positive r.

Stability by the Fourier series method (von Neumann's method)

This method expresses

The idea of this method consists of that, an initial line of errors in terms of a finite Fourier series, and considers the growth of a function that reduces to this series for $t=0$ by a "variables seperable" method identical with that commonly used for deriving analytical solutions of PDEs. The Fourier series can be formulated in terms of sines or cosines but the algebra is easier if the complex exponential form is used, i.e., with $\sum a_n \cos\frac{n\pi x}{l}$ or $\sum b_n \sin\frac{n\pi x}{l}$ replaced by the equivalent $\sum A_n e^{i\frac{n\pi x}{l}}$, $i=\sqrt{-1}$. l is the interval throughout which the function is defined.

We use the notation $U(ph, qk) = U_{p,q}$ (instead of U_{ij}).

Then $A_n e^{i\frac{n\pi x}{l}} = A_n e^{i\left(\frac{n\pi ph}{Nh}\right)} = A_n e^{i\beta_n ph}$, where

$\beta_n = \frac{n\pi}{Nh}$ and $Nh = l$

Denote the errors at the pivotal points along $t=0$, between $x=0$ and Nh, by $E(ph) = E_p$, $p=0,1,\ldots,N$.

Then the $(N+1)$ equations

$$E_p = \sum_{n=0}^{N} A_n e^{i\beta_n ph}, \quad (p=0,1,\ldots,N),$$

are sufficient to determine the $(N+1)$ unknowns $A_0, A_1, \ldots, A_N$ uniquely, showing that an arbitrary distribution of initial errors can be expressed in this exponential form.

To investigate the propagation of this error as t increases we need to find a solution of the finite-difference eq. which reduces to $e^{i\beta ph}$ when $t = qk = 0$.

Assume

$$E_{p,q} = e^{i\beta x} e^{\alpha t} = e^{i\beta ph} e^{\alpha qk} = e^{i\beta ph} \xi^q, \qquad (1)$$

where $\xi = e^{\alpha k}$, and α, in general, is a complex constant. This obviously reduces to $e^{i\beta ph}$ when $q=0$. The error will not increase as t increases provided

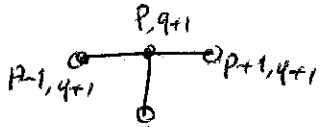
$$|\xi| \leq 1 .$$

for the stability

The criterion $|\xi| \leq 1$ is necessary and sufficient for two time-level difference eq.s but is not always sufficient for three or more level eq.s: although it is always necessary.

Example 1 Investigate the stability of the fully-implicit finite-difference equation

$$\frac{u_{p,q+1} - u_{p,q}}{k} = \frac{u_{p-1,q+1} - 2u_{p,q+1} + u_{p+1,q+1}}{h^2} \qquad (2)$$



approximating the parabolic eq.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} .$$

As the error function $E_{p,q}$ satisfies the same difference equation as $u_{p,q}$, substitution of $E_{p,q}$ from eq. (1) into eq. (2) gives

$$e^{i\beta ph}\xi^{q+1} - e^{i\beta ph}\xi^q = r\{e^{i\beta(p-1)h}\xi^{q+1} - 2e^{i\beta ph}\xi^{q+1} + e^{i\beta(p+1)h}\xi^{q+1}\},$$

where $r = k/h^2$. Division by $e^{i\beta ph}\xi^q$ leads to

$$\xi - 1 = r\xi(e^{-i\beta h} - 2 + e^{i\beta h})$$
$$= r\xi(2\cos\beta h - 2) = -4r\xi \sin^2\left(\frac{\beta h}{2}\right).$$

Hence

$$\xi = \frac{1}{1 + 4r\sin^2\left(\frac{\beta h}{2}\right)} .$$

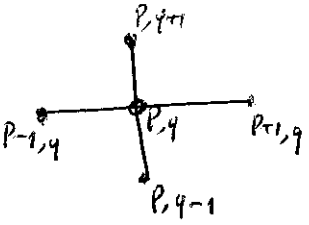
For stability $|\xi| \le 1$, which is so for all positive values of r.

Example 2. The hyperbolic equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

is approximated by the explicit scheme

$$\frac{u_{p,q+1} - 2u_{p,q} + u_{p,q-1}}{k^2} = \frac{u_{p+1,q} - 2u_{p,q} + u_{p-1,q}}{h^2} \qquad (3)$$



Investigate its stability.

Putting $E_{p,q} = e^{i\beta ph}\xi^q$, $\xi = e^{\alpha k}$ in (3), we obtain

$$\xi - 2 + \xi^{-1} = r^2\left(e^{i\beta h} - 2 + e^{-i\beta h}\right)$$

$$\Rightarrow \xi^2 - 2\xi + 1 = \xi r^2 (2\cos\beta h - 2) = -4\xi r^2 \sin^2\left(\frac{\beta h}{2}\right)$$

$$\Rightarrow \xi^2 - 2\underbrace{\left(1 - 2r^2\sin^2\left(\frac{\beta h}{2}\right)\right)}_{A}\xi + 1 = 0, \quad r = \frac{k}{h}, \qquad (4)$$

$$\Rightarrow \xi^2 - 2A\xi + 1 = 0$$

$$\Rightarrow \xi_1 = A + \sqrt{A^2 - 1} \quad \text{and} \quad \xi_2 = A - \sqrt{A^2 - 1}.$$

For stability

$$|\xi| \le 1.$$

As r, k, β are real $\underline{A \le 1}$

When $A < -1$, $|\xi_2| > 1$, giving instability.

When $-1 \le A \le 1$, $A^2 \le 1$, $\xi_1 = A + i(1-A^2)^{\frac{1}{2}}$, $\xi_2 = A - i(1-A^2)^{\frac{1}{2}}$,

Hence

$$|\xi_1| = |\xi_2| = \sqrt{A^2 + (1-A^2)} = 1,$$

providing that eq. (3) is stable for $-1 \le A \le 1$. By (4) we then have

$$-1 \le 1 - 2r^2\sin^2\left(\frac{\beta h}{2}\right) \le 1 \Rightarrow -1 \le 1 - 2r^2\sin^2\left(\frac{\beta h}{2}\right)$$

$$\Rightarrow r^2 \le \frac{1}{\sin^2\left(\frac{\beta h}{2}\right)} \Rightarrow r \le \frac{1}{\sin\left(\frac{\beta h}{2}\right)} \Rightarrow \boxed{r \le 1}$$

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(n)} \right\}, \quad i = 1(1)m.$$

Jacobi method:

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(n)} \right\}, \quad i = 1(1)m$$

If the G-S are written as | Gauss-Seidel method

$$\Rightarrow \quad x_i^{n+1} = x_i^{(n)} + \frac{1}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i}^{m} a_{ij} x_j^{(n)} \right\}$$

SOR (Successive over-relaxation method):

$$x_i^{(n+1)} = x_i^{(n)} + \frac{\omega}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i}^{m} a_{ij} x_j^{(n)} \right\}, \quad i = 1(1)m.$$

it may be written as

$$x_i^{(n+1)} = x_i^{(n)} + \frac{\omega}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - a_{ii} x_i^{(n)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(n)} \right\}$$

$$= x_i^{(n)} - \omega x_i^{(n)} + \frac{\omega}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(n)} \right\}$$

$$\Rightarrow \quad x_i^{(n+1)} = \frac{\omega}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(n)} \right\} + (1-\omega) x_i^{(n)}$$

$$= \omega \left(\text{R.H.S. of the Gauss-Seidel iteration equation} \right) - (\omega - 1) x_i^{(n)}.$$

The factor ω, called the accelation parameter or relaxation factor, generally lies in the range $1 < \omega < 2$.

$\omega = 1 \Rightarrow$ Gauss-Seidel iteration.

Iterative methods in matrix form

$$Ax = b \quad (1) \qquad \underbrace{\begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}}_{A\ 4\times4} \begin{bmatrix} x_1 \\ \vdots \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_4 \end{bmatrix}$$

We write A as

$$A = D - L - U, \text{ where}$$

For 4×4

$$D = \begin{bmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{44} \end{bmatrix}, \quad -L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ a_{31} & a_{32} & 0 & 0 \\ a_{41} & a_{42} & a_{43} & 0 \end{bmatrix},$$

$$-U = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} \\ 0 & 0 & a_{23} & a_{24} \\ 0 & 0 & 0 & a_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The eq. (1) can then be written as

$$(D - L - U)x = b \Rightarrow Dx = (L+U)x + b$$

Jacobi iter

$$\Rightarrow Dx^{(n+1)} = (L+U)x^{(n)} + b \qquad (1)$$

$$\Rightarrow x^{(n+1)} = D^{-1}(L+U)x^{(n)} + D^{-1}b$$

The matrix $D^{-1}(L+U)$ is called the ~~point~~ Jacobi iteration matrix.

The Gauss-Seidel iteration is defined by

$$Dx^{(n+1)} = Lx^{n+1} + Ux^{(n)} + b \qquad (2)$$

Hence $$(D-L)x^{(n+1)} = Ux^{n} + b$$

giving that

$$x^{(n+1)} = (D-L)^{-1}Ux^{(n)} + (D-L)^{-1}b.$$

which shows that the Gauss-Seidel iteration matrix is $(D-L)^{-1}U$.

SOR:

$$x_i^{(n+1)} = x_i^{(n)} + \frac{\omega}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i}^{m} a_{ij} x_j^{(n)} \right\}, \quad i = 1(1)m.$$

or

$$x_i^{(n+1)} = \frac{\omega}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{m} a_{ij} x_i^{(n)} \right\} - (\omega - 1) x_i^{n}$$

$$= \omega \, (\text{R.H.S. of the Gauss-Seidel iteration eq.s}) - (\omega - 1) x_i^{(n)}.$$

The correction or displacement vector $d^{(n)} = x^{(n+1)} - x^{(n)}$ of the SOR method is taken to be ω times the displacement vector $d_1^{(n)}$ defined by the Gauss-Seidel iteration.

$$D d_1^{(n+1)} = D\left(x^{(n+1)} - x_n^{(n)}\right) = L x^{(n+1)} + U x^{(n)} + b - D x^{(n)}$$

Hence the SOR iteration, defined by

$$d^{(n)} = \omega d_1^{(n)}$$

can be written as

$$x^{(n+1)} - x^{(n)} = \omega D^{-1} \left(L x^{(n+1)} + U x^{(n)} + b - D x^{(n)} \right).$$

Therefore

$$(I - \omega D^{-1} L) x^{(n+1)} = x^{(n)} + \omega D^{-1} U x^{(n)} + \omega D^{-1} b - \omega x^{(n)}$$

$$= \{ (1 - \omega) I + \omega D^{-1} U \} x^{(n)} + \omega D^{-1} b.$$

Hence

$$x^{(n+1)} = (I - \omega D^{-1} L)^{-1} \{ (1 - \omega) I + \omega D^{-1} U \} x^{(n)} + (I - \omega D^{-1} L)^{-1} \omega D^{-1} b \quad (3)$$

showing the SOR iteration matrix is

$$(I - \omega D^{-1} L)^{-1} \{ (1 - \omega) I + \omega D^{-1} U \}.$$

A necessary and sufficient condition for the convergence iterative methods

Each of the three iterative methods described above can be written as

$$X^{(n+1)} = GX^{(n)} + C, \qquad (4)$$

where G is the iteration matrix and C a column vector of known values. This equation was derived from the original equations by rearranging them into the form

$$X = GX + C, \qquad (5)$$

i.e., the unique solution of the n linear eq.s $Ax=b$ is the solution of equation (5). The error $e^{(n)}$ in the n-th approximation to the exact solution is defined by $e^{(n)} = X - X^{(n)}$:

$$e^{(n+1)} = Ge^{(n)}.$$

Therefore

$$e^{(n)} = Ge^{(n-1)} = G^2 e^{(n-2)} = \ldots = G^n e^{(0)}. \qquad (6)$$

The sequence of iterative values $X^{(0)}, X^{(2)}, \ldots, X^{(n)}, \ldots$ will converge to X as n tends to infinity if

$$\lim_{n\to\infty} e^{(n)} = 0.$$

Since $X^{(0)}$ and therefore $e^{(0)}$ is arbitrary it follows that the iteration will converge if and only if

$$\lim_{n\to\infty} G^n = 0.$$

Assume now that the matrix G of order n has n LI eigenvectors v_s, $s = 1(1)n$. Then these eigenvectors can be used as a basis for our n-dimensional vector space and the arbitrary error vector $e^{(0)}$, with its n

components, can be expressed uniquely as a linear combination of them, namely

$$e^{(0)} = \sum_{s=1}^{m} c_s v_s ,$$

where the c_s are scalers. Hence

$$e^{(1)} = G e^{(0)} = \sum_{s=1}^{m} c_s G v_s ,$$

But $G v_s = \lambda_s v_s$, where λ_s is the eigenvalue corresponding to v_s. Hence

$$e^{(1)} = \sum_{s=1}^{n} c_s \lambda_s v_s , \quad e^{(2)} = G e^{(1)} = G \sum c_s \lambda_s v_s$$
$$= \sum c_s \lambda_s G v_s = \sum_{s=1}^{m} c_s \lambda_s^2 v_s$$

Similarly, $$e^{(n)} = G e^{(n+1)} = \sum_{s=1}^{m} c_s \lambda_s^n v_s . \qquad (7)$$

Therefore $e^{(n)}$ will tend to the null vector as n tends to infinity, for arbitrary $e^{(0)}$, if and only if $|\lambda_s| < 1$ for all s. In other words, the iteration will converge for arbitrary $x^{(0)}$ if and only if the spectral radius $\rho(G)$ of G is less that one.

As a corollary to this result a sufficient condition for convergence is that $\|G\| < 1$. To prove this we have that $G v_s = \lambda_s v_s$. Hence

$$\| G v_s \| = \| \lambda_s v_s \| = |\lambda_s| \| v_s \| .$$

On the other hand $$\| G v_s \| \le \|G\| \| v_s \|$$

$$\Rightarrow \quad |\lambda_s| \|v_s\| \le \|G\| \cdot \|v_s\| \Rightarrow |\lambda_s| \le \|G\|, \quad s = 1(1)m.$$

It follows from this that a sufficient condition for convergence is that $\|G\| < 1$. It is not a necessary condition because the norm of G can exceed one even when $\rho(G) < 1$.

As an example consider the Jacobi iteration matrix $D^{-1}(L+U)$. If the i-th equation of $Ax=b$ is

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{ii}x_i + \dots + a_{im}x_m = b_i,$$

then the ith row of $D^{-1}(L+U)$ is

$$\frac{a_{i1}}{a_{ii}} \quad \frac{a_{i2}}{a_{ii}} \quad \dots \quad \frac{a_{i,i-1}}{a_{ii}} \quad 0 \quad \frac{a_{i,i+1}}{a_{ii}} \quad \dots \quad \frac{a_{im}}{a_{ii}}.$$

Then if we take our matrix norm as the infinity norm, the Jacobi iteration will converge if

$$|a_{i1}| + |a_{i2}| + \dots + |a_{i,i-1}| + 0 + |a_{i,i+1}| + \dots + |a_{im}| < |a_{ii}|.$$

$\Rightarrow$ The Jacobi method applied to the equation $Ax=b$ will converge if A is a strictly diagonally dominant matrix.