

# Numerical Methods for PDE

D.Y. Kwak

March 14, 2016



# Contents

<b>1</b>	<b>Finite Difference Method</b>	<b>1</b>
1.1	2nd order linear p.d.e. in two variables . . . . .	1
1.2	Finite Difference Method . . . . .	3
1.2.1	Error of difference operator . . . . .	7
1.3	Elliptic equation in 2D . . . . .	10
1.3.1	Basic finite difference method for elliptic equation . . .	10
1.3.2	Treatment of irregular boundaries(Dirichlet boundary conditions . . . . .	16
1.3.3	Convection -diffusion equation . . . . .	24
1.4	Parabolic p.d.e's . . . . .	29
1.4.1	Discretization of parabolic p.d.e, General Case . . . . .	38
1.5	Finite element method for parabolic problems . . . . .	41
1.5.1	One dimensional model problem . . . . .	41
1.6	Semi discretization in space . . . . .	42
1.7	Fully discrete Scheme . . . . .	44
1.8	Hyperbolic Equation . . . . .	46
1.8.1	Method of Characteristics . . . . .	49
1.8.2	FDM for Hyperbolic equations . . . . .	50
1.8.3	Implicit method for second order hyperbolic equations .	84



# Chapter 1

## Finite Difference Method

### 1.1 2nd order linear p.d.e. in two variables

General 2nd order linear p.d.e. in two variables is given in the following form:

$$L[u] = Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G \text{ in } \Omega,$$

where  $\Omega$  is an open set in  $\mathbb{R}^2$ . According to the relations between coefficients, the p.d.e.s are classified into 3 categories, namely,

elliptic	if $AC - B^2 > 0$
hyperbolic	if $AC - B^2 < 0$
parabolic	if $AC - B^2 = 0$ .

Furthermore, if the coefficients  $A, B$  and  $C$  are constant, it can be written as

$$\left[ \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right] \begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{bmatrix} + Du_x + Eu_y + Fu = G.$$

Auxiliary condition

$$\left\{ \begin{array}{l} \text{B.C. Dirichlet, Neumann, Robin} \\ \text{I.C. for parabolic problem} \\ \text{Interface Cond} \end{array} \right.$$

The condition  $u = g_0$  on  $\Gamma_0 \subset \partial\Omega$  is called the *Dirichlet B.C.*, the condition  $\frac{\partial u}{\partial n} = g_1$  on  $\Gamma_1 \subset \partial\Omega$  is called the *Neumann B.C.*, the condition  $\alpha \frac{\partial u}{\partial n} + u =$

$g_2$  on  $\Gamma_2 \subset \partial\Omega$  is called the *Robin B.C.* Here  $\Gamma_0 \cup \Gamma_1 \cup \Gamma_2 = \partial\Omega$  is a disjoint union except measure zero set. If some of these conditions are mixed, we say it is a mixed B.C.

### Elliptic Problem

In general a 2nd order linear p.d.e. in  $\mathbb{R}^d$  can be given in the following convenient form:

$$L[u] = -\sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left( a_{ij} \frac{\partial u}{\partial x_j} \right) + cu = -\nabla \cdot \mathcal{A} \nabla u + cu = f \text{ in } \Omega \quad (1.1)$$

+ B.C.s or I.C.s

Here  $\mathcal{A} = (a_{ij})_{i,j=1}^d$  is the coefficient matrix.  $L$  is called a **differential operator**. The equation is called *elliptic* if  $\mathcal{A}$  is positive definite, in which case a typical B.C. is  $u = g$  on  $\partial\Omega$ . Here  $u$  may represent some physical quantities such as electromagnetic potential, displacement of elastic membrane, temperature, concentration of chemical component, or pressure of a fluid (in porous media), etc.

- behavior near boundary
- Equation (1.1) holds in an open set  $\Omega$ .

### Notations

Let

$$C(\Omega), C^1(\Omega), C(\bar{\Omega}), C^k(\bar{\Omega}), C(\partial\Omega)$$

denote some functions spaces. Define

$$\partial_i u = \frac{\partial u}{\partial x_i}, \quad \partial_{ij} u = \frac{\partial^2 u}{\partial x_j \partial x_i}, \quad \Delta = (\partial_{11} + \cdots + \partial_{dd})$$

so that  $\nabla : C^1(\Omega) \rightarrow (C(\Omega))^d$  and  $\nabla \cdot : (C^1(\Omega))^d \rightarrow C(\Omega)$  represent differential operators via

$$\nabla u = (\partial_1 u, \cdots, \partial_d u)^T, \quad \nabla \cdot \mathbf{v} = \partial_1 v_1 + \cdots + \partial_d v_d.$$

We call  $\Delta$  the *Laplace operator* and also write

$$\Delta = \nabla \cdot \nabla = \nabla^2$$

**Definition 1.1.1** (Classical solution). Assume  $f \in C(\Omega)$ ,  $g \in C(\partial\Omega)$ . A function  $u$  is called a *classical solution* of (1.1) if  $u \in C^2(\Omega) \cap C(\bar{\Omega})$ . If  $f \notin C(\Omega)$  or  $g \notin C(\partial\Omega)$  then the solution  $u$  may not belong to  $C^2(\Omega) \cap C(\bar{\Omega})$ . Still the solution may exist in some sense, called a *weak solution*.

We say a pde is “well posed” if a unique solution exists and the solution depends continuously on the data. There are basically three classes of method to discretize it,

- (1) Finite Difference method
- (2) Finite Element method
- (3) Finite Volume method

## 1.2 Finite Difference Method

Let  $u(x)$  be a function defined on  $\Omega \subset \mathbb{R}^n$ . Let  $U_{i,j}$  be the function defined over discrete domain  $\{(x_i, y_j)\}$  (such points are grid points) that may approximate  $u_{i,j} = u(x_i, y_j)$ . Such functions are called grid functions.

Difference operator

$$\begin{aligned}\partial^+ U_i &= \frac{U_{i+1} - U_i}{h_{i+1}}, & \text{forward difference} \\ \partial^- U_i &= \frac{U_i - U_{i-1}}{h_i}, & \text{backward difference} \\ \partial^0 U_i &= \frac{U_{i+1} - U_{i-1}}{h_i + h_{i+1}}, & \text{central difference} \\ \partial^2 U_i &= \frac{2(\partial^+ - \partial^-)}{h_i + h_{i+1}}, & \text{central 2nd difference}\end{aligned}$$

**Example 1.2.1.** Note that

$$\begin{aligned}\partial^+ U_i &= \frac{U_{i+1} - U_i}{h_{i+1}} = \partial^0 U_{i+1/2}, & \text{central difference at } x_{i+1/2} \\ \partial^- U_i &= \frac{U_i - U_{i-1}}{h_i} = \partial^0 U_{i-1/2}, & \text{central difference at } x_{i-1/2}\end{aligned}$$

**Exercise 1.2.2.** (1) We can interpret  $\partial^2 U_i$  as a central difference  $2 \frac{\partial^0 U_{i+1/2} - \partial^0 U_{i-1/2}}{h_i + h_{i+1}}$ . Derive the truncation error.

**Example 1.2.3.** Consider the following second order two point boundary value problem :

$$-u''(x) = f(x), u(a) = c, u(b) = d.$$

Assume a mesh  $a = x_0 < x_1 < \dots < x_N = b$ ,  $\Delta x_i = x_{i+1} - x_i = h$ . If  $u \in C^4$ , we replace the derivative by a difference quotient, and obtain

$$-\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + O(h^2) = f(x_i), \quad i = 1, \dots, N-1, u_0 = c, u_N = d$$

Dropping the error term, we obtain a system of linear equations in the approximate values  $U_i$ :

$$-\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} = f(x_i) := f_i, \quad i = 1, \dots, N-1, U_0 = c, U_N = d.$$

This is an  $(N-1) \times (N-1)$  matrix equations.

$$h^{-2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \begin{pmatrix} U_1 \\ \cdot \\ \vdots \\ \cdot \\ U_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ \cdot \\ \vdots \\ \cdot \\ f_{N-1} \end{pmatrix} + h^{-2} \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \\ d \end{pmatrix}$$

Above equation can be written as  $L_h U^h = F^h$ , where  $U^h = (U_1, \dots, U_{N-1})$  and  $F^h = (f_i) +$  boundary terms. It is called a difference equation for the given differential equation.

**Exercise 1.2.4.** Write down a matrix equation for the same problem with second boundary condition changed to the Neumann condition at  $b$ , i.e,  $u'(b) = d$ . If one uses first order difference for derivative, we lose accuracy.

We need an extra equation in this case. There are several choices:

- (1) Use first order backward difference scheme

$$d = u'(b) \approx \frac{U_N - U_{N-1}}{h}$$

append it to the last equation. (first order accuracy)

- (2) Assuming the D.E. holds at the boundary, form a central difference equa-



tion at  $U_N$  by introducing a fictitious point  $U_{N+1}$  :

$$-\frac{1}{h^2}(U_{N-1} - 2U_N + U_{N+1}) = f(1) \quad (1.2)$$

$$\frac{1}{2h}(U_{N+1} - U_{N-1}) = d. \quad (1.3)$$

Eliminating  $U_{N+1}$  from these eqs., we have

$$\frac{U_N - U_{N-1}}{h^2} = \frac{d}{h} + \frac{f(1)}{2}. \quad (1.4)$$

Append it as  $N$ -th equation. The matrix is still symmetric; Eq. (1.4) can be viewed as centered difference approximation to  $u'(x_n - \frac{h}{2})$  and rhs as the first two terms of Taylor expansion

$$u'(x_n - \frac{h}{2}) = u'(x_n) - \frac{h}{2}u''(x_n) + \dots$$

(3) Approximate  $u'(1)$  by higher order scheme such as

$$\frac{u_{N-2} - 4u_{N-1} + 3u_N}{2h} = d.$$

In this case one has second order truncation error (Show it) but the matrix loses symmetry.

**Exercise 1.2.5.** (1) Solve the D.E. in Example 1.2.3 with  $f = 2 - 6x$  so that  $u = x - x^2 + x^3$  and the following BCs (with  $h = 1/n$ ,  $n = 5, 10, 20, 40, \dots$ ). Report the error  $\|u - u_h\|_{L^2} \doteq \sqrt{\sum_i |(u - u_h)(x_i)|^2}$  using a Table.

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

(b)  $u(0) = 0, u'(1) = 2$  (Neumann) For this problem implement all three method in the previous exercise (1), (2) and (3).

(2) Write down the stiffness matrix of 2D problem with Neumann condition at  $x = 1$  on the unit square with  $4 \times 4$  grid. Label the node  $x_1, x_2, x_3, \dots$  from the bottom row.(excluding the boundary) There are two possibilities to treat the Neumann condition: One is to use backward difference. Another is to assume fictitious values and use central difference, then incorporate them into the five point stencil. In other words, use  $u_x \doteq \frac{u_9 - u_3}{2h} = g_2(1, \frac{1}{4})$  and substitute into the stencil, the third equation

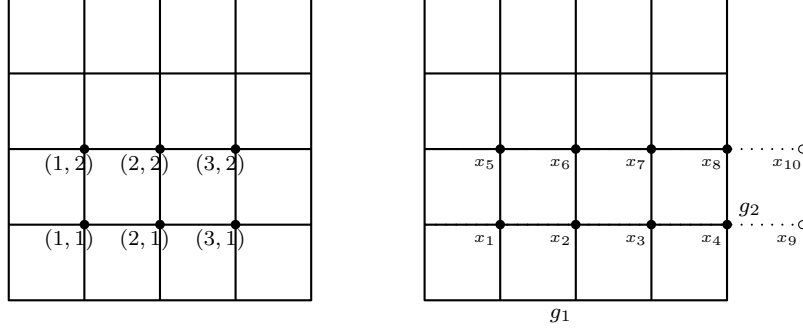


Figure 1.1: Numbering and Grid for the Neumann problem

becomes

$$\frac{1}{h^2}(-2u_3 + 4u_4 - u_8) = (f + \frac{2}{h}g_2)(1, \frac{1}{4}) + \frac{g_1}{h^2}.$$

**Example 1.2.6** (Heat equation). We consider

$$\begin{aligned} u_t &= \sigma u_{xx}, & \text{for } 0 < x < 1, \quad 0 < t < T \\ u(t, 0) &= u(t, 1) = 0 \\ u(0, x) &= g(x), \quad g(0) = g(1) = 0 \end{aligned}$$

Let  $x_i = ih, i = 0, \dots, N, \Delta x = 1/N$  and  $t_n = n\Delta t, \Delta t = \frac{T}{J}$ . Then we have the following difference scheme

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} = \sigma \left[ \frac{U_{i-1}^n - 2U_i^n + U_{i+1}^n}{\Delta x^2} \right],$$

for  $i = 1, 2, \dots, N - 1$  and  $n = 1, 2, \dots, M - 1$  where  $U_i^n \approx u(t_i, x_n)$ . From the boundary condition and initial condition we have

$$U_i^0 = g(x_i), U_0^n = 0, U_N^n = 0.$$

$$U_i^{n+1} = U_i^n + \frac{\sigma \Delta t}{\Delta x^2} [U_{i-1}^n - 2U_i^n + U_{i+1}^n].$$

In vector notation

$$U_h^{n+1} = U_h^n - \frac{\sigma \Delta t}{\Delta x^2} A U_h^n$$

where  $A$  is the same matrix as in example 1. If  $n = 0$ , right hand side is known. Thus

$$U_h^n = (I - \sigma \frac{\Delta t}{\Delta x^2} A)^n G, \quad G = (g(x_1), \dots, g(x_{N-1}))^T.$$

This is called **forward Euler** or **explicit scheme**. If we change the right hand side to

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{\Delta t} &= \sigma \left[ \frac{U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}}{\Delta x^2} \right] \\ U_i^{n+1} &= U_i^n + \frac{\sigma \Delta t}{\Delta x^2} [U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}]. \end{aligned}$$

$$(I + \sigma \frac{\Delta t}{\Delta x^2} A)^n U_h^n = G, \quad G = (g(x_1), \dots, g(x_{N-1}))^T.$$

This is called **backward Euler** or **implicit scheme**.

### 1.2.1 Error of difference operator

For  $u \in C^2$ , use the Taylor expansion about  $x_i$

$$u_{i+1} = u(x_i + h_i) = u(x_i) + h_i u'(x_i) + \frac{h_i^2}{2} u''(\xi), \quad \xi \in (x_i, x_{i+1})$$

$$\therefore \frac{u_{i+1} - u_i}{h_i} - u'(x_i) = \frac{h_i}{2} u''(\xi).$$

Expanding  $u(x_i)$  about  $x_{i+1}$ ,

$$u(x_i) = u(x_{i+1}) - h_i u'(x_{i+1}) + \frac{h_i^2}{2} u''(x_{i+1}) - \frac{h_i^3}{6} u'''(\theta).$$

These are first order accurate. To derive a second order scheme, expand about  $x_{i+1/2}$ ,

$$\begin{aligned} u_{i+1} &= u_{i+1/2} + \frac{h_i}{2} u'(x_{i+1/2}) + \frac{1}{2} \left(\frac{h_i}{2}\right)^2 u''(x_{i+1/2}) + \frac{1}{6} \left(\frac{h_i}{2}\right)^3 u^{(3)}(\xi) \\ u_i &= u_{i+1/2} - \frac{h_i}{2} u'(x_{i+1/2}) + \frac{1}{2} \left(\frac{h_i}{2}\right)^2 u''(x_{i+1/2}) - \frac{1}{6} \left(\frac{h_i}{2}\right)^3 u^{(3)}(\xi). \end{aligned}$$

Subtracting, we obtain

$$\frac{u_{i+1} - u_i}{h_i} = u'(x_{i+1/2}) + \frac{h_i^2}{24} u^{(3)}(x_{i+1/2}) + O(h_i^3).$$

Thus we obtain a second order approximation to  $u'(x_{i+1/2})$ . By translation, we have

$$\frac{u_{i+1} - u_{i-1}}{2h_i} - u'(x_i) = O(h_i^2/6) \quad \text{if } h_i = h_{i+1}. \quad (1.5)$$

Assume  $h_i = h_{i+1}$  and we substitute the solution of differential equation into the difference equation. Using  $-u'' = f$  we obtain

$$\begin{aligned} & \frac{(-u_{i-1} + 2u_i - u_{i+1})}{h^2} - f(x_i) := L_h u - F^h \\ = & \frac{1}{h^2}(-u_i + hu'_i - \frac{h^2}{2}u''_i + \frac{h^3}{6}u^{(3)} - \frac{h^4}{24}u^{(4)}(\theta_1) + 2u_i) \\ & + \frac{1}{h^2}(-u_i - hu'_i - \frac{h^2}{2}u''_i - \frac{h^3}{6}u^{(3)} - \frac{h^4}{24}u^{(4)}(\theta_2)) - f(x_i) \\ = & -u''_i - f(x_i) - \frac{h^2}{24}(u^{(4)}(\theta_1) + u^{(4)}(\theta_2)) \\ \leq & \frac{h^2}{24} \max |u^{(4)}|. \end{aligned}$$

Thus we obtain a discrete equation

$$L_h U_h = F^h. \quad (1.6)$$

We let  $\tau_h = L_h u - F^h$  and call it the **truncation error**.

**Definition 1.2.7.** We say a difference scheme is **consistent** if the truncation error approaches zero as  $h$  approaches zero, in other words, if  $L_h u - f \rightarrow 0$  in some norm.

Truncation error measures how well the difference equation approximates the differential equation. But it does not measure the actual error in the solution. However, for a stable scheme, these two errors are equivalent.

Use of different quadrature for  $f$ . Instead of  $f(x_i)$  we can use

$$\frac{1}{12}[f(x_{i-1}) + 10f(x_i) + f(x_{i+1})] = \frac{5}{6}f(x_i) + \frac{\mu_0}{6}f(x_i)$$

where  $\mu_0 f(x_i)$  is the average of  $f$  which is  $f(x_i) + O(h^2)$ .

### Nonuniform grid(irregular mesh)

We use central difference scheme at  $x_{i\pm 1/2}$  to get

$$u'(x_{i+1/2}) \approx \frac{u_{i+1} - u_i}{h_{i+1}} \quad \text{and} \quad u'(x_{i-1/2}) \approx \frac{u_i - u_{i-1}}{h_i}.$$

Thus, it is natural to approximate as

$$u''(x_i) \approx \left( \frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i} \right) / \left( \frac{h_i + h_{i+1}}{2} \right). \quad (1.7)$$

**Definition 1.2.8.**  $L_h$  is said to be **stable** if there is a constant  $C$  independent of  $h$  such that

$$\|U_h\| \leq C \|F^h\| \quad \text{for all } h > 0$$

where  $U^h$  is the solution of the difference equation,  $L_h U^h = F^h$ . In other word,  $L_h$  is stable if and only if  $L_h^{-1}$  is bounded.

**Definition 1.2.9.** A finite difference scheme is said to **converge** if

$$\|U_h - u\| \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

$e_h = U_h - u$  is called the **discretization error**.

**Theorem 1.2.10** (Lax equivalence theorem). *Given a consistent scheme, stability is equivalent to convergence.*

*Proof.* Assume stability. From  $L_h u - f = \tau^h$ ,  $L_h U_h - F^h = 0$ , we have  $L_h(u - U_h) = \tau^h$ . Thus,

$$\|u - U_h\| \leq C \|L_h(u - U_h)\| = C \|\tau^h\| \rightarrow 0.$$

Hence the scheme converges and the error is bounded by truncation error. Obviously a convergent scheme must be stable. From the theory of p.d.e, we know  $\|u\| \leq C \|f\|$ . Hence

$$\|U_h\| \leq \|U_h - u\| + \|u\| \leq O(\tau^h) + C \|f\| \leq C \|f\| \leq C \|F^h\|.$$

□

**Exercise 1.2.11.** (1) Derive a truncation error for (1.5) in case of irregular mesh.(use weighted difference)

(2) Find truncation error for of difference scheme for  $-u''(x_i)$  in (1.7) in case

of nonuniform grid.

$$\begin{aligned}
 L_h u &= 2[-h_i u_{i+1} + (h_i + h_{i+1})u_i - h_{i+1}u_{i-1}]/h_i h_{i+1}(h_i + h_{i+1}) \\
 &= 2 \left[ -h_i(u_i + h_{i+1}u'_i + \frac{h_{i+1}^2}{2}u''_i + \frac{h_{i+1}^3}{6}u_i^{(3)} + O(h^4)) + (h_i + h_{i+1})u_i \right. \\
 &\quad \left. - h_{i+1}(u_i - h_i u'_i + \frac{h_i^2}{2}u''_i - \frac{h_i^3}{6}u_i^{(3)} + O(h^4)) \right] / h_i h_{i+1}(h_i + h_{i+1}) \\
 &= -u_i^{(2)} + \frac{1}{3}(h_{i+1} - h_i)u_i^{(3)} + O(h_i^2 + h_{i+1}^2). \tag{1.8}
 \end{aligned}$$

- (3) Use  $\frac{1}{3}[f(x_i) + f(x_{i+1}) + f(x_{i-1})]$  for the right hand side. What is the truncation error?
- (4) Show for uniform grid, we have

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = \frac{1}{12}[f(x_{i-1}) + 10f(x_i) + f(x_{i+1})] + Ch^4 \max |u^{(6)}(x)|.$$

## 1.3 Elliptic equation in 2D

### 1.3.1 Basic finite difference method for elliptic equation

First consider the following elliptic problem:(Dirichlet problem)

$$\begin{aligned}
 -\Delta u &= f \text{ in } \Omega \\
 u &= g \text{ on } \partial\Omega
 \end{aligned}$$

- (1) Approx. D.E.  $-(u_{xx} + u_{yy}) = f$  by a finite difference at interior mesh pts.
- (2) The unknown function  $u$  is approximated by a grid function  $U_h$ . Assume  $u \in C^4$ .

In one dim case,

$$\begin{aligned}
 u(x+h) &= u(x) + hu_x(x) + \frac{h^2}{2}u_{xx}(x) + \frac{h^3}{6}u_{xxx}(x) + O(h^4) \\
 u(x-h) &= \dots
 \end{aligned}$$

Thus

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u_{xx}(x) + O(h^2)$$

while in 2D case

$$\begin{aligned} u_{xx}(x, y) &\doteq [u(x+h, y) - 2u(x, y) + u(x-h, y)]/h^2 \\ u_{yy}(x, y) &\doteq [u(x, y+h) - 2u(x, y) + u(x, y-h)]/h^2 \end{aligned}$$

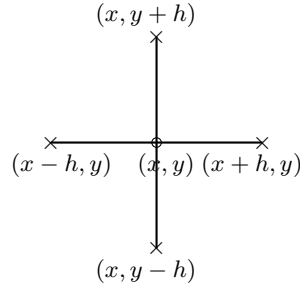


Figure 1.2: 5-point Stencil

This is called a *5 point Stencil* or *5 point Star*, etc. Approximating  $\nabla^2 u = \Delta u$  by 5-point stencil for each point (interior mesh pt), we obtain a linear system of equations  $A\mathbf{x} = \mathbf{f}$  in unknowns  $\mathbf{x} = u_{ij}$ . By Gershgorin disc theorem, the matrix  $A$  is nonsingular.  $L_h[u]$  is called **finite difference operator**, e.g.,

$$L_h[u](x, y) \doteq [-4u(x, y) + u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h)]/h^2$$

### Problems with variable coefficients

More generally,

$$L[u] = - \left[ \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right] \text{Diag}\{a_{11}, a_{22}\} \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{bmatrix} + cu = -(a_{11}u_x)_x - (a_{22}u_y)_y + cu$$

With uniform meshes, the central differences gives

$$u_x(x) \doteq \frac{u(x+h) - u(x-h)}{2h} \quad (1.9)$$

$$(u_x)_x(x) \doteq \frac{u_x(x + \frac{h}{2}) - u_x(x - \frac{h}{2})}{h} \quad (1.10)$$

$$u_x(x + \frac{h}{2}) \doteq \frac{u(x+h) - u(x)}{h} \quad (1.11)$$

$$u_x(x - \frac{h}{2}) \doteq \frac{u(x) - u(x-h)}{h} \quad (1.12)$$

For a problem with variable coefficients  $a(x, y)$ , we use central difference (1.10, 1.11)

$$(a_{11}u_x)_x \doteq [(a_{11}u_x)(x + \frac{h}{2}) - (a_{11}u_x)(x - \frac{h}{2})]/h \quad (1.13)$$

$$\doteq \frac{a_{11}(x + \frac{h}{2})(u(x+h) - u(x)) - a_{11}(x - \frac{h}{2})(u(x) - u(x - \frac{h}{2}))}{h^2} \quad (1.14)$$

$$\begin{aligned} u(x+h_2) &= u(x) + h_2u_x(x) + \frac{h_2^2}{2}u_{xx} + \frac{h_2^3}{3!}u^{(3)} + \dots \quad \times h_1 \\ u(x-h_1) &= u(x) - h_1u_x(x) + \frac{h_1^2}{2}u_{xx} - \frac{h_1^3}{3!}u^{(3)} \dots \quad \times h_2 \end{aligned}$$

$$\begin{aligned} &h_1u(x+h_2) - h_2u(x-h_1) \\ &= (h_1 - h_2)u(x) + 2h_1h_2u_x(x) + \frac{h_1h_2}{2}(h_2 - h_1)u_{xx} + \dots \end{aligned}$$

$$\therefore u_x(x) = \frac{h_1u(x+h_2) - h_2u(x-h_1) - (h_1 - h_2)u(x)}{2h_1h_2} + O(h)$$

This is only first order accurate. To get a second order method multiply two equations respectively by  $h_1^2, h_2^2$  and subtract to get(i.e, eliminate  $u_{xx}$ )

$$\begin{aligned} &h_1^2u(x+h_2) - (h_1^2 - h_2^2)u(x) - h_2^2u(x-h_1) \\ &= (h_2h_1^2 + h_1h_2^2)u_x(x) + \left(\frac{h_1^2h_2^3}{6} + \frac{h_2^2h_1^3}{6}\right) \max|u'''|. \end{aligned}$$

Hence

$$u_x \approx \frac{h_1^2u(x+h_2) - (h_1^2 - h_2^2)u(x) - h_2^2u(x-h_1)}{h_1h_2(h_1 + h_2)}$$

is second order accurate. Compare this with (1.5).

As a simple example, we consider a differential eq. of the form(with  $\gamma > 0$ )

$$L[u] \equiv -[u_{xx} + u_{yy}] + \gamma u = f$$

whose discretized form

$$L_h[U] = \frac{1}{h^2} \begin{pmatrix} 4 + \gamma h^2 & -1 & -1 & 0 \\ -1 & 4 + \gamma h & 0 & -1 \\ -1 & 0 & 4 + \gamma h^2 & -1 \\ 0 & -1 & -1 & 4 + r\gamma h^2 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} = F$$

satisfies



- (1)  $L_h[u] = L[u] + O(h^2)$  as  $h \rightarrow 0$ .  $u$  is true solution.
- (2)  $AU = F + Bdy$ ,  $Au = [\Delta u - \gamma u + O(h^2)] + Bdy$

With abuse of notation, we write it as

$$L_h(U - u) = O(h^2) = \tau_h.$$

Let  $A$  be the matrix representation of  $L_h$  then the discretization error  $U - u$  has the form  $A^{-1}\tau_h$  (depends on  $h$ ) and satisfies

$$\|U - u\| \leq \|A^{-1}\| \cdot \|\tau_h\| \leq \|A^{-1}\| O(h^2).$$

If we put  $D = \text{diag}A = \{a_{11}, \dots, a_{nn}\}$ , then  $D^{-1}A(U - u) = D^{-1}\tau_h$ . Write  $D^{-1}A = I + B$ , where  $B$  is off diagonal. Then we know  $\|B\|_\infty = \frac{4}{4+\gamma h^2} < 1$  if  $\gamma > 0$ . Thus  $(D^{-1}A)^{-1} = (I + B)^{-1}$  exists and

$$\|(D^{-1}A)^{-1}\|_\infty = \|(I + B)^{-1}\|_\infty \leq \frac{1}{1 - \|B\|_\infty} \leq \frac{4 + \gamma h^2}{\gamma h^2}.$$

Hence

$$\|U - u\|_\infty \leq \|(D^{-1}A)^{-1}\|_\infty \cdot \|D^{-1}\tau_h\|_\infty \leq \frac{4 + \gamma h^2}{\gamma h^2} \cdot \frac{h^2}{4 + \gamma h^2} O(h^2) = O(h^2) \rightarrow 0.$$

Thus, we have proved the following result.

**Theorem 1.3.1** (Convergence of FDM -special case). *Assume*

- (1)  $u \in C^4(\Omega)$
- (2)  $\gamma > 0$
- (3) *uniform mesh*

Then  $\|U - u\|_\infty = O(h^2)$  as  $h \rightarrow 0$ .

### General Elliptic problems

Generally,  $A, B, C$  are not constant. In this case, we can still put the problem into a conservative form as follows:

$$\begin{aligned} L[u] &= Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0 \\ &= \nabla^T \cdot \begin{pmatrix} A & B \\ B & C \end{pmatrix} \nabla u - (A_x + B_y - D)u_x - (B_x + C_y - E)u_y + Fu + G, \end{aligned}$$

where  $\nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)$ , so that  $\nabla u = \begin{pmatrix} u_x \\ u_y \end{pmatrix}$ . If  $A_x + B_y - D = 0$  and  $B_x + C_y - E$ , it is self-adjoint.

### Treating the cross term like $u_{xy}$

Assume  $u_{xy} = u_{yx}$ , we approximate  $\frac{\partial}{\partial y} \frac{\partial u}{\partial x}$  by  $\delta_y^0 \delta_x^0 U^h$  where  $\delta_x^0 U^h(P) = \frac{U^h(E) - U^h(W)}{2\Delta x}$  is the central difference. Then from

$$\delta_y^0 U^h(P) = \frac{U^h(N) - U^h(S)}{2\Delta y}$$

and forward -backward difference formula we get

$$\delta_y^0 \delta_x^0 U^h = \frac{1}{2\Delta_y} \left[ \frac{U^h(NE) - U^h(NW)}{2\Delta x} - \frac{U^h(SE) - U^h(SW)}{2\Delta x} \right]$$

### Change of variable method to eliminate the cross term

One can transform the variable so that the resulting equation in new variable does not have cross term.

**Lemma 1.3.2.** *Let  $s = s(x, y)$ ,  $t = t(x, y)$  be a coordinate transform which is locally one-to-one onto. Denote its derivative by  $\frac{\partial(s,t)}{\partial(x,y)} = P$ , Jacobian matrix. Then we have*

$$\nabla_{(x,y)} u = \begin{bmatrix} u_x \\ u_y \end{bmatrix} = \begin{bmatrix} u_s s_x + u_t t_x \\ u_s s_y + u_t t_y \end{bmatrix} = P^T \cdot \begin{bmatrix} u_s \\ u_t \end{bmatrix} = P^T \cdot \nabla_{(s,t)} \cdot u$$

In other words,

$$\nabla_{(x,y)} = \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \end{pmatrix} = \begin{pmatrix} \frac{\partial s}{\partial x} \cdot \frac{\partial}{\partial s} + \frac{\partial t}{\partial x} \cdot \frac{\partial}{\partial t} \\ \frac{\partial s}{\partial y} \cdot \frac{\partial}{\partial s} + \frac{\partial t}{\partial y} \cdot \frac{\partial}{\partial t} \end{pmatrix} = \begin{pmatrix} \frac{\partial s}{\partial x} & \frac{\partial t}{\partial x} \\ \frac{\partial s}{\partial y} & \frac{\partial t}{\partial y} \end{pmatrix} \begin{pmatrix} \partial/\partial s \\ \partial/\partial t \end{pmatrix} = P^T \cdot \nabla_{(s,t)}$$

**Remark 1.3.3.** If we let  $(s, t) = F(x, y)$  then  $\text{grad}_{(x,y)} = DF^T \text{grad}_{(s,t)}$ .

Hence  $\nabla_{(x,y)}^T = \nabla_{(s,t)}^T \cdot P$  and we see that

$$\nabla_{(x,y)}^T A \nabla_{(x,y)} u = \nabla_{(s,t)}^T P A P^T \nabla_{(s,t)} u.$$

If  $A$  is symmetric, there exists a  $P$  such that  $P A P^T = \text{diagonal} = \{d_1, d_2\}$ .

If we choose  $s(x, y)$ ,  $t(x, y)$  so that  $\frac{\partial(s,t)}{\partial(x,y)} = P$ , then

$$\nabla_{(x,y)}^T A \nabla_{(x,y)} u = \frac{\partial}{\partial s} \left( d_1 \frac{\partial u}{\partial s} \right) + \frac{\partial}{\partial t} \left( d_2 \frac{\partial u}{\partial t} \right).$$

**Example 1.3.4.** Transform the problem  $u_{xx} + 4u_{xy} + u_{yy} = 0$  so that it does not have cross term.

Since  $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ , its eigenvalues are 3, -1 with corresponding eigenvectors (1, 1) and (1, -1), we see that with  $P = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ , we have

$$PAP^T = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}.$$

Hence the transformed equation is

$$\frac{\partial}{\partial s} \left( 3 \frac{\partial u}{\partial s} \right) - \frac{\partial}{\partial t} \left( \frac{\partial u}{\partial t} \right) = 0.$$

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} \partial s / \partial x & \partial s / \partial y \\ \partial t / \partial x & \partial t / \partial y \end{pmatrix}. \therefore \begin{pmatrix} s \\ t \end{pmatrix} = P \begin{pmatrix} x \\ y \end{pmatrix}.$$

If  $s = \text{constant}$ ,  $ds = s_x dx + s_y dy = 0$ , so the line  $s = \text{constant}$  is described in  $(x, y)$ -coordinate as

$$\frac{dy}{dx} = -\frac{s_x}{s_y} = -\frac{P_{11}}{P_{12}}.$$

Likewise, if  $t = \text{constant}$ ,  $dt = t_x dx + t_y dy = 0 \therefore$  so the line  $t = \text{constant}$  is described as

$$\frac{dy}{dx} = -\frac{t_x}{t_y} = -\frac{P_{21}}{P_{22}}.$$

When  $A$  is  $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ , then, using the rotation of axis, we can take

$$P = \begin{bmatrix} \cos \lambda & -\sin \lambda \\ \sin \lambda & \cos \lambda \end{bmatrix}, \quad \begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} \cos \lambda & -\sin \lambda \\ \sin \lambda & \cos \lambda \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

where  $\cot 2\lambda = \frac{c-a}{2b}$ .

### 1.3.2 Treatment of irregular boundaries(Dirichlet boundary conditions)

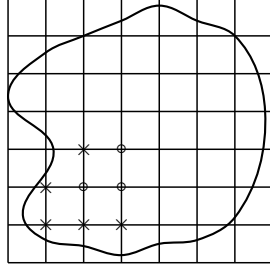


Figure 1.3:  $\Omega_h$ ,  $\circ$  regular,  $\times$  irregular

Let  $\Omega$  be a domain with grid. Let  $\Omega_h$  be the set of all grid points in  $\Omega$ .

**Definition 1.3.5.** Two points  $P, Q$  on the grid are said to be **properly adjacent** if they are adjacent and the line segment connecting  $P, Q$  belongs to  $\Omega$ . A grid point is called a **regular** point if all four adjacent points belong to  $\Omega_h$  and they are properly adjacent to  $P$ . Let  $\Omega_h^o$  be the set of all **regular** points. Define the set  $\Omega_h^* = \Omega_h - \Omega_h^o$  and the points in  $\Omega_h^*$  are called **irregular** points.

In the following, we let  $E$  be the east neighbor point of  $P$  in  $\Omega_h$  and let  $W$  be the west neighborhood point of  $P$  in  $\Omega_h$ , etc.

First order derivatives are easy to approximate, i.e, use either forward or backward difference. We can form the difference equation  $L_h U^h = f^h$  at all regular points as before. We only consider the equation at irregular points.

#### Method 1

If  $P$  is an irregular point, we let  $U^h(P) = g(Q)$ . Here  $Q$  is a point of  $\partial\Omega$  closest to  $P$ . Here  $U^h(P)$  is now known.

#### Method 2(Collatz-linear interpolation)

We form  $L_h U^h = f^h$  at all points of  $\Omega_h$  as follows: First we form  $L_h U^h = f^h$  at all regular points of  $\Omega_h$ . If  $P \in \partial\Omega_h$  is an irregular point lying near west

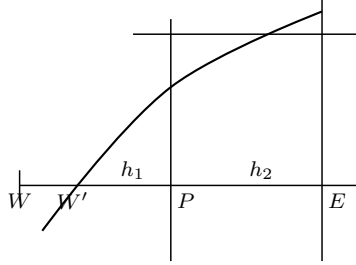


Figure 1.4: Near irregular boundary

part of  $\partial\Omega$ , take the point of intersection  $W'$  of the line segment  $EP$  with  $\partial\Omega$ . Then we let

$$U^h(P) = \frac{h_1}{h_1 + h_2} U^h(E) + \frac{h_2}{h_1 + h_2} u(W'). \quad (1.15)$$

Now append this equation to the difference equation. If  $E$  happens to belong to  $\partial\Omega$  also, then  $U^h(P)$  is completely determined, hence we do not need to append it to the difference equation.

**Remark 1.3.6.** The equation (1.15) has nothing to do with the differential equation itself, thus it may break certain properties of matrix.

### Method 3(Shortley-Weller)

For an irregular point  $P$ , we set (recall H.W. 4)

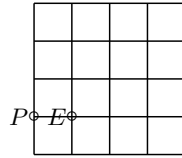
$$\frac{\partial^2 u}{\partial x^2}(P) \doteq 2 \left( \frac{U^h(E) - U^h(P)}{h_2} - \frac{U^h(P) - u(W')}{h_1} \right) / (h_1 + h_2), \quad \text{etc..}$$

This is nothing but a difference formula for nonuniform grid (see earlier example).

Advantage: This difference equation comes from the differential equation, thus preserves (hopefully) certain properties of the matrix (like positive definiteness, banded structure, diagonal dominance). But usually symmetry breaks down.

**Method 4(Fictitious point method)**

Let  $P$  be an irregular point whose west neighbor  $W$  lies outside of  $\Omega_h$ . We use extrapolation to get  $U^h(W) = \alpha U^h(W') + \beta U^h(P)$ , where  $W'$  is the point of intersection of the line segment from  $P$  with  $\partial\Omega$ . ( $\alpha = \frac{h_2}{h_1}, \beta = -\frac{h_2-h_1}{h_1}$ .  $h_2 = h$  and  $h_1$  is the distance from  $P$  to the boundary.) Finally we substitute  $U^h(W)$  into the difference equation at  $P$ . (It is called fictitious point method)

**Neumann or Robin boundary condition(regular point)**Figure 1.5:  $P$ : regular boundary point

Consider the boundary condition of type  $\frac{\partial u}{\partial n} + \gamma u = g$  on  $\partial\Omega$ . Let  $P$  be a regular boundary point(boundary point lying on the grids). If the boundary is vertical line, then use one sided difference to get

$$\frac{U^h(P) - U^h(E)}{h} + \gamma(P)U^h(P) = g(P)$$

and append it to the difference equation.

If  $P$  is an irregular boundary point(figure 1.4), use  $\frac{U^h(E)-U^h(W)}{2h} + \gamma(P)U^h(P) = g(P)$ . Now solve it for  $U^h(W)$  and substitute it into the difference equation at  $P$  to get a new equation.

$$\frac{1}{h^2}(-U^h(S) - U^h(W) + 4U^h(P) - U^h(N) - U^h(E)) = f(P).$$

If  $P$  is near corner do the same for north and south derivative.

**Neumann or Robin boundary condition(irregular point)**

We let  $C$  be a grid point not in  $\partial\Omega_h$ . Draw a normal line to  $\partial\Omega$  and let  $C'$  be the point of intersection with  $\partial\Omega$ . Now treat  $C$  as a grid point. Extend  $CC'$

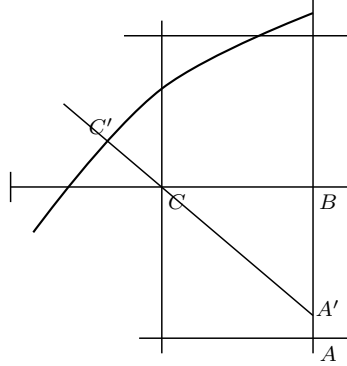


Figure 1.6: irregular boundary point

to the closest grid line consisting of  $AB$ , letting  $A'$  denote the intersection of extension and the segment  $AB$ . Use

$$\begin{aligned} \frac{\partial u}{\partial n}(C') &\doteq \frac{U^h(C) - U^h(A')}{CA'} \\ U^h(C') &\doteq (1 - \sigma)U^h(A') + \gamma U^h(C) \\ \Rightarrow \frac{U^h(C) - U^h(A')}{CA'} + \gamma(C')U^h(C') &= g(C'), \end{aligned}$$

where  $U^h(A')$  is obtained by interpolation:

$$U^h(A') = (1 - \alpha)U^h(A) + \alpha U^h(B).$$

This is an equation involving unknowns  $U^h(A)$ ,  $U^h(B)$  and  $U^h(C)$ .

**Example 1.3.7.**

$$-\nabla^2 u + (x^2 + y^2)u = 40(2 - x - y + 4xy)e^{xy}, \quad 0 \leq x, y \leq 1$$

where  $u(x, y) = [10 - 20\{(x - \frac{1}{2})^2 + (y - \frac{1}{2})^2\}]e^{xy}$ . Use this as the boundary function also. We have The error is  $O(h^2)$ .

Assuming the error is of the form  $\|U - u\|_\infty = Mh^\alpha$ , we see

$$\frac{\|U_h - u\|}{\|U_{\frac{h}{2}} - u\|} = \frac{Mh^\alpha}{M(\frac{h}{2})^\alpha} = 2^\alpha.$$

$h_x$	grids	$\ U - u\ _\infty$	ratio
0.2	$5 \times 5$	0.0506	
$0.2 \times 2^{-1}$	$10 \times 10$	0.0140	3.6148
$0.2 \times 2^{-2}$	$20 \times 20$	0.0035	4
$0.2 \times 2^{-3}$	$40 \times 40$		

Table 1.1: Example

These are computable with  $u$  replaced by  $U_{h_{min}}$  and

$$\alpha = \log \left[ \frac{\|U_h - u\|}{\|U_{\frac{h}{2}} - u\|} \right] / \log 2$$

**Theorem 1.3.8** (Maximum Principle). *Assume  $A$  is positive definite symmetric,  $c \geq 0$ . Let  $u$  is the solution of elliptic p.d.e. given by*

$$\begin{aligned} L[u] &= - \sum_i \frac{\partial}{\partial x_i} \left[ \sum_j a_{ij} \frac{\partial u}{\partial x_j} \right] + cu = -\nabla A \nabla u + cu = 0 \text{ in } \Omega \\ u &= g \quad \text{on } \partial\Omega \end{aligned}$$

Then for  $(x, y) \in \text{int } \Omega$

$$|u(x, y)| \leq \max_{(x, y) \in \partial\Omega} |u(x, y)| \quad (1.16)$$

*Proof.* Assume  $c > 0$ . There exists orthogonal matrix  $P$  such that  $P^T A P = \text{diag}\{d_1, d_2\}$  where  $d_1, d_2 > 0$ . Suppose  $u$  has a positive maximum at some interior point  $Q = (x^*, y^*)$  of  $\Omega$ . Define

$$\begin{pmatrix} s \\ t \end{pmatrix} = P^T(x^*, y^*) \begin{pmatrix} x \\ y \end{pmatrix}$$

so that  $L[u] = -\nabla_{(s,t)} P^T A P \nabla_{(s,t)} u + cu = 0$ . At  $Q$ ,  $u_s(Q) = u_t(Q) = 0$ ,  $u_{ss}(Q) \leq 0$  and  $u_{tt}(Q) \leq 0$ . Hence

$$L[u] = -(d_1 u_s)_s(Q) - (d_2 u_t)_t(Q) + c(Q)u(Q) = 0.$$

Since  $d_1 > 0$ ,  $d_2 > 0$ ,  $cu > 0$ , this is a contradiction. Thus either

$$0 \leq u(x^*, y^*) \leq \max_{(x, y) \in \partial\Omega} u(x, y)$$



or  $u(x, y) \leq u(x^*, y^*) < 0$  for all  $(x, y)$ . Similar argument shows that if  $u$  has a negative minimum in the interior, we can show (1.16) holds.

Now if  $c \geq 0$  we consider a perturbation. Choose  $\alpha$  so large that  $L[e^{\alpha x}] = -(d_1\alpha^2 + d_2\alpha^2 - c)e^{\alpha x} < 0$  and let  $v = u + Ee^{\alpha x}$ .

$$L[v] = L[u] + EL[e^{\alpha x}] < 0 \quad \text{for all } E > 0.$$

Suppose  $v$  has a pos. max. at an interior point  $Q$  of  $\Omega$ . Then  $L[v] = -d_1v_{ss}(Q) - d_2v_{tt}(Q) + c(Q)v(Q) \geq 0$ , a contradiction. Hence

$$0 \leq u + Ee^{\alpha x} \leq \max_{\partial\Omega} \{u + Ee^{\alpha x}\}.$$

Let  $E \rightarrow 0$ . Then

$$0 \leq u(x, y) \leq \max_{\partial\Omega} u.$$

Similar argument holds when  $v$  has a neg. min. at an interior point  $Q$ .  $\square$

**Remark 1.3.9.** Examining above proof we can conclude  $L[u] = 0$  can be replaced by  $L[u] \leq 0$ .

Applying maximum principle to  $u$  and  $-u$ , we see  $u$  cannot have negative minimum, i.e.,

$$u(x, y) \geq \min_{\partial\Omega} u.$$

Thus, we obtain the result.

**Corollary 1.3.10.** *If*

$$\begin{aligned} L[u] &= 0 && \text{in } \Omega \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

*then*  $u \equiv 0$ .

As a consequence we have uniqueness of solution.

**Corollary 1.3.11.** *If*  $u_1, u_2$  *satisfy*

$$\begin{aligned} L[u_i] &= f && \text{in } \Omega \\ u_i &= g && \text{on } \partial\Omega, \end{aligned}$$

*then*  $u_1 = u_2$ .

**Theorem 1.3.12** (Discrete max. principle). *Let the grid function  $U$  satisfy the finite difference equation  $L_h[U] \equiv \sum_Q A(P, Q)U(Q) = 0$ , i.e.,*

$$A(P, P)U(P) + A(P, E)U(E) + A(P, S)U(S) + A(P, N)U(N) + A(P, W)U(W) = 0$$

*for each mesh point  $P$ , where coefficient  $A(P, Q)$  are generated by finite difference method and  $A$  is pos. def. weakly diagonally dominant.*

*Then*

$$|U(P^*)| \leq \max_{P \in \partial\Omega_h} |U(P)|, \quad \text{for all } P^* \in \text{int } \Omega_h. \quad (1.17)$$

*Proof.* Solving for  $U(P)$ , we have

$$U(P) = \frac{1}{A(P, P)} \sum_{Q \neq P} -A(P, Q)U(Q).$$

Since  $|A(P, P)| \geq \sum_{Q \neq P} |A(P, Q)|$  we have

$$|U(P)| \leq \sum_{Q \neq P} \left| \frac{A(P, Q)}{A(P, P)} \right| \max_{Q \neq P} |U(Q)| \leq \max_{Q \neq P} |U(Q)|.$$

Repeat the same process until you hit the boundary. □

**Corollary 1.3.13** (Uniqueness of discrete solution).

$$\begin{aligned} L_h U &= 0 && \text{in } \Omega \\ U &= 0 && \text{on } \partial\Omega \end{aligned}$$

*implies  $U = 0$ .*

**Remark 1.3.14.** Examining the proof, one can notice a slightly weaker version of Maximum principle holds if  $L(u) \leq 0$ . Under the same condition as previous theorem, except  $L(u) \leq 0$ , we see  $u$  cannot have positive maximum in the interior.

**Theorem 1.3.15** (Discrete max. principle 2nd version). *Suppose  $L_h[U] \leq 0$ . Then  $U$  cannot have a positive maximum unless  $U$  is constant. In other words,*

$$0 < \max_{p \in \Omega_h} U(p) = \max_{\partial\Omega_h} U(p)$$

*or*

$$\max_{p \in \Omega_h} U(p) < 0.$$

*Proof.* Suppose there is a point  $P_0 \in \Omega_h$  such that  $U(P_0)$  is positive and  $U(P_0) \geq U(P)$  for all  $P \in \Omega_h$ . Then by similar argument as above,

$$U(P_0) \leq \sum_{Q \neq P_0} \left| \frac{A(P_0, Q)}{A(P_0, P_0)} \right| \max_{Q \neq P_0} U(Q) \leq \max_{Q \neq P_0} U(Q) \leq U(P_0).$$

Hence

$$U(Q) = U(P_0)$$

for all  $Q$  in the nhd of  $P_0$ . Repeating the argument for each  $Q$  in the nhd of  $P_0$  until we hit the boundary, we see  $U$  must be constant. Thus we have the desired result.  $\square$

**Note.** Minimum principle is obtained when  $L_h[U] \geq 0$ . (See exercise.)

**Theorem 1.3.16** (Convergence of FDM - More general case). *Let  $u$  be the solution of  $L[u] = -\nabla^T A \nabla u + cu = 0$  in  $\Omega$  and  $u = g$  on  $\partial\Omega$  and let  $U$  be the finite sequence of grid functions satisfying  $L_h U = 0$ , where  $L_h(u) = \mathcal{O}(h^\alpha)$ ,  $\alpha > 0$  (truncation error). If  $A$  is constant, diagonally dominant, positive definite, then  $\|U - u\|_\infty = \mathcal{O}(h^\alpha)$  as  $h \rightarrow 0$ .*

*Proof.* Let  $w = U - u$ , then  $L_h[w] = L_h[U] - L_h[u] = -L_h[u]$  and  $w = 0$  on  $\partial\Omega_h$ . Let  $s(x, y) \equiv r^2 - (x - x_0)^2 - (y - y_0)^2$  with  $(x_0, y_0) \in \text{int } \Omega$ ,  $r$  chosen so large that the circle  $s = 0$  contains  $\Omega$ . Then  $L_h[s] = L[s]$  because  $s$  is quadratic. (compute it)

$$\begin{aligned} L[s] &= -\nabla^T A \nabla s + cs = -a_{11}s_{xx} - (a_{12} + a_{21})s_{xy} - a_{22}s_{yy} + cs \\ &= 2(a_{11} + a_{22}) + cs \geq 2(a_{11} + a_{22}). \end{aligned}$$

There exist  $M > 0$  such that  $|L_h[u]| \leq Mh^\alpha$  by the hypothesis  $L_h[u] = \mathcal{O}(h^\alpha)$ .

We see that

$$L_h \left[ \frac{Mh^\alpha s(x, y)}{2(a_{11} + a_{22})} \right] \geq Mh^\alpha \geq |L_h[u]|.$$

Also

$$L_h \left[ \pm w - \frac{Mh^\alpha s(x, y)}{2(a_{11} + a_{22})} \right] = \pm L_h[w] - L_h \left[ \frac{Mh^\alpha s(x, y)}{2(a_{11} + a_{22})} \right] \leq \mp L_h[u] - Mh^\alpha \leq 0.$$

(Recall  $w = U - u$  and  $L_h[w] = -L_h[u]$  and  $w = 0$  on  $\partial\Omega$ )

Now by the discrete maximum principle-2nd version (where  $L_h(U) = 0$  is

replaced by  $L_h[U] \leq 0$ ,

$$\max_{P \in \Omega_h} \left[ \pm w - \frac{Mh^\alpha s}{2(a_{11} + a_{22})} \right] \leq \max_{P \in \partial\Omega_h} \left[ \pm w - \frac{Mh^\alpha s}{2(a_{11} + a_{22})} \right] = -\frac{Mh^\alpha}{2(a_{11} + a_{22})} \min_{\partial\Omega_h} s \leq 0.$$

Thus  $|w| \leq \frac{Mh^\alpha s}{2(a_{11} + a_{22})}$  and

$$\|u - U\|_{h,\infty} = \|w\|_{h,\infty} \leq \frac{Mh^\alpha}{2(a_{11} + a_{22})} \max_{\Omega_h} s \leq \frac{Mh^\alpha r^2}{2(a_{11} + a_{22})}.$$

□

**Exercise 1.3.17.** (1) Write down stencil for the fictitious point method and determine if it is symmetric!

(2) State and prove a version of minimum principle.

### 1.3.3 Convection -diffusion equation

Consider another type of differential equation, namely a special case of convection diffusion equation.

$$-\epsilon u_{xx} + au_x = f$$

with the BC.  $u(0) = u_0$ ,  $u(1) = u_1$  or Neumann condition  $u'(1) = 0$  with  $a(1) > 0$ . Here, we assume  $0 < \epsilon \ll 1$ . If we use the central difference scheme for the first order derivative, we get

$$\epsilon \left( \frac{-U_{i-1}^h + 2U_i^h - U_{i+1}^h}{h^2} \right) + a \frac{U_{i+1}^h - U_{i-1}^h}{2h} = f_i^h \quad (1.18)$$

$$-\left( \frac{\epsilon}{h^2} + \frac{a}{2h} \right) U_{i-1}^h + \frac{2\epsilon}{h^2} U_i^h - \left( \frac{\epsilon}{h^2} - \frac{a}{2h} \right) U_{i+1}^h = f_i^h \quad (1.19)$$

Thus the sum of off diagonal elements is

$$\sum_{j \neq i} |a_{ij}| = \left| \frac{\epsilon}{h^2} + \frac{a}{2h} \right| + \left| \frac{\epsilon}{h^2} - \frac{a}{2h} \right|.$$

If  $a, h$  is fixed and  $\epsilon \rightarrow 0$ , it becomes  $a/h$ , while  $a_{ii} = 2\epsilon/h^2 \rightarrow 0$ . Thus the resulting matrix is not diagonally dominant and it causes a lot of problems. For example, the resulting linear system is not positive definite and hence it may be more difficult to solve. But, most importantly, the resulting discretization

does not yield an accurate approximation to the problem. One way to fix this situation is to keep the Peclet number :  $\frac{ah}{\epsilon} < 2$  so that the sum of off diagonal elements is less than or equal to  $2\epsilon/h^2 = a_{ii}$ . The disadvantage of this scheme is that small  $h$  enlarges the size of discrete equation.

### Numerical Difficulties

- (1) The solution may exhibit oscillation which are physically unrealistic
- (2) Taking small mesh size means large problem size which take more time to solve.
- (3) The iterative method may fail to converge

### Upwind difference scheme

An alternative way to avoid this difficulty is to use backward difference for  $u_x$  for  $a > 0$  and forward difference for  $u_x$  for  $a < 0$ . This method of choosing difference scheme is called **upwind difference scheme**. For  $a > 0$

$$\begin{aligned} \epsilon \left( \frac{-U_{i-1}^h + 2U_i^h - U_{i+1}^h}{h^2} \right) + a \frac{U_i^h - U_{i-1}^h}{h} &= f_i^h \\ - \left( \frac{\epsilon}{h^2} + \frac{a}{h} \right) U_{i-1}^h + \left( \frac{2\epsilon}{h^2} + \frac{a}{h} \right) U_i^h - \left( \frac{\epsilon}{h^2} \right) U_{i+1}^h &= f_i^h \end{aligned}$$

The resulting system is irreducibly diagonally dominant, thus it is an  $M$ -matrix. Error analysis can be carried out to show optimal order of convergence. Also, for the solver part, simple Jacobi method works.

**Example 1.3.18.** [boundary layer]

$$-\epsilon u'' + u' = 0 \text{ in } (0, 1)$$

with  $u(0) = a, u(1) = b$  has the unique solution  $u(x) = C_1 e^{\frac{x}{\epsilon}} + C_2$ , where  $C_1 = \frac{b-a}{e^{1/\epsilon}-1}$  and  $C_2 = \frac{ae^{1/\epsilon}-b}{e^{1/\epsilon}-1}$ .

The following examples are taken from (K.W. Morton-Numerical Solution of convection-Diffusion problems, Chapman Hall, 1996)

**Example 1.3.19.**

$$-\epsilon \Delta u + \mathbf{b} \cdot \nabla u = 0 \text{ on } (0, 1) \times (0, 1)$$

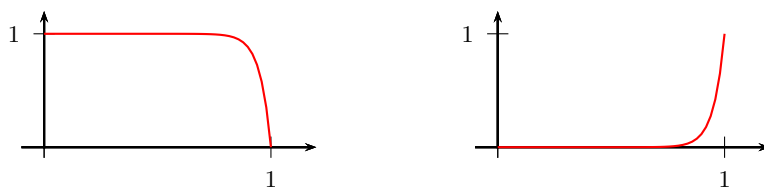


Figure 1.7:  $u(x) = C_1 e^{x/\epsilon} + C_2$ ,  $a = 1, b = 0$  and  $a = 0, b = 1$

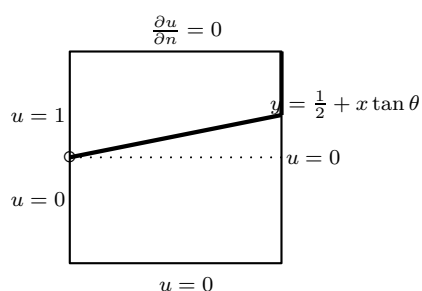


Figure 1.8: internal and boundary layer

with

$$\mathbf{b} = (\cos \theta, \sin \theta)$$

for  $0 \leq \theta < \frac{\pi}{2}$  and discontinuous inflow boundary condition

$$u(0, y) = \begin{cases} 0, & y \in [0, \frac{1}{2}) \\ 1, & y \in (\frac{1}{2}, 1] \end{cases}$$

and  $\frac{\partial u}{\partial y} = 0$  on  $y = 1$ , and  $u = 0$  on  $y = 0$  or  $x = 1$ . This leads to an internal layer along  $y = \frac{1}{2} + x \tan \theta$  and a boundary layer at  $x = 1$  for  $y > \frac{1}{2} + \tan \theta$  when  $\tan \theta < \frac{1}{2}$ . Draw the graph.

**Example 1.3.20** (Heat equation).

$$\begin{aligned} \frac{\partial u}{\partial t} + \mathbf{b} \cdot \nabla u &= \epsilon \Delta u \text{ on } (-1, 1) \times (-1, 1) \times (0, T) \\ u(x, y, 0) &= u_0(x, y) \end{aligned}$$

where  $u_0(x, y)$  is a circular cone type centered at  $(1, 0)$  with

$$\mathbf{b} = b(wy, -wx)$$

with exact inflow boundary where needed.

**Example 1.3.21.** (1) A 2-D example exhibiting some (boundary) layer.

$$u = (e^{\frac{x}{\epsilon}} - 1)(e^{\frac{y}{\epsilon}} - 1)$$

(2) The fundamental solution (shifted) of the heat equation:

$$u = \frac{1}{\sqrt{4\pi(t+1)}} e^{-\frac{x^2}{4(t+1)}}.$$

It is nonseparable.

**Example 1.3.22** (Jeon Y.). <sup>1</sup> We consider a convection dominated diffusion equation:

$$-\epsilon\Delta u + \mathbf{b} \cdot \nabla u = 0 \text{ on } (0, 1)^2$$

with

$$u(x) = \begin{cases} 1, & x = 0 \\ 1, & y = 0 \text{ and } 0 \leq x \leq 1/3 \\ 0, & \text{elsewhere} \end{cases}$$

Here  $\epsilon = 10^{-9}$  and  $\mathbf{b} = (1, 3)/\sqrt{10}$ . This problem has both interior and boundary layer.

The same equation with different boundary condition.

$$u(x) = \begin{cases} 1, & x = 0 \\ 1, & y = 0 \\ 0, & \text{elsewhere} \end{cases}$$

**Example 1.3.23** (R. Lin-Numer 09-DGLSFEM for singularly perturbed problems with nonsmooth data). We consider a convection dominated diffusion

---

<sup>1</sup>Analysis of the cell boundary element methods for convection dominated convection-diffusion equations, / JCAM 234 (2010)

equation:

$$-\epsilon^2 \Delta u + c(x)u = f(x) \text{ on } (0, 1)$$

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

$$c(x) = \begin{cases} 1, & x < 0.5 \\ 2 - x, & x > 0.5 \end{cases}$$

and with exact sol.

$$u(x) = \begin{cases} -1 + \frac{e^{-x/\epsilon} + e^{(2x-1)/(2\epsilon)}}{e^{-1/(2\epsilon)} + 1} & x < 0.5 \\ 1 - \frac{e^{(x-1)/\epsilon} + e^{(1-2x)/(2\epsilon)}}{e^{-1/(2\epsilon)} + 1} & x > 0.5 \end{cases}$$

$u$  has boundary layers at  $x = 0$  and  $x = 1$ , and an interior layer at  $x = 0.5$  when  $\epsilon \ll 1$ . Use  $\epsilon = 10^{-5}$

**Example 1.3.24** (R. Lin 2.). We consider

$$\begin{aligned} -\epsilon^2 \Delta u(x, y) + c(x, y)u &= f(x, y) \text{ on } (0, 1)^2 \\ u &= 0 \text{ on } \partial\Omega \end{aligned}$$

with exact sol

$$u(x, y) = \frac{1}{4} u^*(x, y) (\sin 4\pi x + 2) \left(1 - e^{-x/\epsilon}\right) \left(1 - e^{(x-1)/\epsilon}\right) \left(1 - e^{-y/\epsilon}\right) \left(1 - e^{(y-1)/\epsilon}\right)$$

where

$$c(x, y) = \begin{cases} 2, & y < 0.5 \\ 1, & y > 0.5 \end{cases}$$

$$u^*(x, y) = \begin{cases} 3 - e^{(2y-1)/(2\epsilon)}, & y < 0.5 \\ 1 + e^{(1-2y)/(2\epsilon)}, & y > 0.5 \end{cases}$$

$\Omega_1 = (0, 1) \times (0, 0.5)$  and  $\Omega_2 = (0, 1) \times (0.5, 1)$ , and  $\Gamma_0 = (0, 1) \times \{0.5\}$ .

**Example 1.3.25** (R. Lin 3. L-shaped domain). We consider a convection dominated diffusion equation:

$$\begin{aligned} -\epsilon^2 \Delta u + u &= f \text{ on } (-1, 1)^2 \setminus [-1, 0]^2 \\ u &= 0 \text{ on } \partial\Omega \end{aligned}$$



with exact sol

$$u(x) = \tilde{u}(r, \theta) \left(1 - e^{(x-1)/\epsilon}\right) \left(1 - e^{-(x+1)/\epsilon}\right) \left(1 - e^{(y-1)/\epsilon}\right) \left(1 - e^{-(y+1)/\epsilon}\right)$$

$$\tilde{u}(r, \theta) = r^{2/3} \sin\left(\frac{2\theta + \pi}{3}\right)$$

Let  $\Gamma_1 = \{(x, y) : -1 \leq x \leq 0\}$  and  $y = 0$  or  $x = 0$  and  $-1 \leq y \leq 0$  and  $\Gamma_2 = \Gamma - \Gamma_1$ . The solution has a corner singularity at the origin. When when  $\epsilon \ll 1$ , the solution has boundary layers near  $\Gamma_2$ .

**Exercise 1.3.26.** (1) Show that if  $h$  is chosen so that the Peclet number ( $ah/\epsilon < 2$ ), the scheme (1.19) results in a positive definite system (irreducibly Diagonally dominant)

(2) Solve the PDE in Example 1.3.18 with  $\epsilon = 10^{-2}, \dots$  BC changed to  $u(0) = 1, u(1) = 0$  by a) central difference, b) upwind scheme,  $h = 2^{-k}, k = 4, 5, 6, \dots$  Draw graph.

(3) Solve the PDE in Example 1.3.19 with  $\epsilon = 10^{-2}, \dots$  by upwind scheme,  $h = \times 2^{-k}, k = 4, 5, 6, \dots$  Draw graph.

## 1.4 Parabolic p.d.e's

Consider a heat equation on a bar.

$$u_t = u_{xx}, \quad 0 < x < 1, \quad 0 < t \leq T, \quad \text{plus IC., BC..}$$

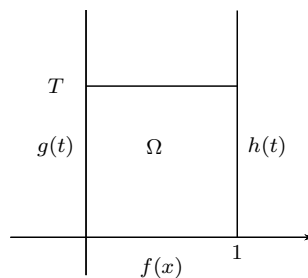


Figure 1.9: Domain

**Theorem 1.4.1** (Maximum principle). *If  $u$  satisfies the heat equation for  $0 < t \leq T$ , then*

$$\min_{\partial\Omega} \{f, g, h\} = m \leq \min_{0 \leq x \leq 1, 0 \leq t \leq T} u \leq \max_{0 \leq x \leq 1, 0 \leq t \leq T} u \leq M = \max_{\partial\Omega} \{f, g, h\}$$

*Proof.* Put  $v = u + Ex^2$ ,  $E > 0$

$$\frac{\partial v}{\partial t} - \frac{\partial^2 v}{\partial x^2} = -2E < 0.$$

If  $v$  attains a maximum at  $Q \in \text{int } \Omega$ , then

$$\begin{aligned} v_t(Q) &= 0, \\ v_{xx}(Q) &\leq 0. \end{aligned}$$

Thus  $(v_t - v_{xx})(Q) \geq 0$ , a contradiction. Hence  $v$  has maximum at a boundary point of  $\Omega$ . For any  $(x, t) \in \Omega$

$$u(x, t) \leq v(x, t) \leq \max v(x, t) \leq M + E.$$

Since  $E$  was arbitrary, the proof is complete. For minimum, use  $-E$  instead of  $E$ .  $\square$

More general parabolic p.d.e.

$$u_t = Au_{xx} + Du_x + Fu + G$$

$$\text{F.D.M} \begin{cases} \text{Explicit} \cdots \text{write down the values of grid function} \\ \text{Implicit} \cdots \text{variables implicitly representing the value} \end{cases}$$

Let the grid be given by

$$\begin{aligned} 0 &= x_0 < x_1 < x_2 < \cdots < x_{N+1} = 1, & x_i &= ih, & \text{uniform grid} \\ 0 &= t_0 < t_1 < \cdots, & t_j &= jk \end{aligned}$$

**Explicit method**

$$\begin{cases} \frac{U_{i,j+1} - U_{i,j}}{k} \doteq u_t \\ \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} \doteq u_{xx} \end{cases}$$

$$\therefore U_{i,j+1} = \lambda U_{i-1,j} + (1 - 2\lambda)U_{i,j} + \lambda U_{i+1,j}$$

where  $\lambda = k/h^2$ .

Recall: Stability means **solution remains bounded** as time goes on.

**Theorem 1.4.2.** *If  $u$  is sufficiently smooth, then*

$$\left| u_{xx} - \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} \right| = O(h^2) \quad \text{as } h \rightarrow 0$$

and

$$\left| u_t - \frac{u(x,t+k) - u(x,t)}{k} \right| = O(k) \quad \text{as } k \rightarrow 0$$

**Theorem 1.4.3.** *Suppose  $u$  is sufficiently smooth, and satisfies*

$$\begin{aligned} u_t &= u_{xx} & 0 < x < 1, \quad t > 0 \\ u(x,0) &= f(x) \\ u(0,t) &= g(t) \\ u(1,t) &= h(t). \end{aligned}$$

If  $U_{i,j}$  is the solution of the explicit finite difference scheme, then for  $0 < \lambda \leq \frac{1}{2}$ ,

$$\max_{i,j} |u_{i,j} - U_{i,j}| \doteq O(h^2 + k) \quad \text{as } h, k \rightarrow 0,$$

*i.e., finite difference solution converges to the true solution.*

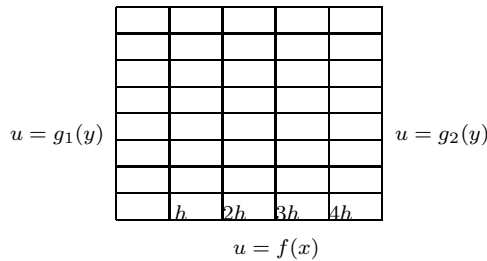


Figure 1.10:

*Proof.* Put  $u_{ij} \equiv u(x_i, t_j)$ . Then from

$$(1) \frac{u_{i,j+1} - u_{i,j}}{k} = u_t + O(k)$$

$$(2) \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} = u_{xx} + O(h^2)$$

we get

$$u_{i,j+1} = u_{i,j} + \frac{k}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + k(O(k) + O(h^2)).$$

Hence

$$u_{i,j+1} = \lambda u_{i-1,j} + (1 - 2\lambda)u_{i,j} + \lambda u_{i+1,j} + Ck(k + h^2).$$

Let the discretization error be  $w_{i,j} = u_{i,j} - U_{i,j}$  so that

$$w_{i,j+1} = \lambda w_{i-1,j} + (1 - 2\lambda)w_{i,j} + \lambda w_{i+1,j} + O(k^2 + kh^2).$$

Since  $0 < \lambda \leq \frac{1}{2}$ ,  $0 \leq 1 - 2\lambda < 1$ , three coefficient are positive and their sum is 1. (convex combination) We see

$$|w_{i,j+1}| \leq \lambda|w_{i-1,j}| + (1-2\lambda)|w_{i,j}| + \lambda|w_{i+1,j}| + M(k^2 + kh^2) \quad \text{for some } M > 0.$$

If we define  $\|w_j\| = \max_{1 \leq i \leq N} |w_{i,j}|$ , then

$$\begin{aligned} \|w_{j+1}\| &\leq \|w_j\| + M(k^2 + kh^2) \\ &\leq \|w_{j-1}\| + 2M(k^2 + kh^2) \leq \dots \leq \|w_0\| + (j+1)M(k^2 + kh^2). \end{aligned}$$

Since  $\|w_0\| = 0$ ,

$$\|w_{j+1}\| \leq (j+1)kM(k + h^2) \leq TM(k + h^2), \quad (j+1)k \leq T.$$

In fact,

$$M = \max_{0 \leq x \leq 1, 0 \leq t \leq T} \left( \frac{1}{2}|u_{tt}| + \frac{1}{12}|u_{xxxx}| \right).$$

□

**Remark 1.4.4.** If  $\lambda > \frac{1}{2}$ , the solution may not converge.

**Exercise 1.4.5.** Prove the formula is unstable for  $\lambda > \frac{1}{2}$ . Let

$$u(x, 0) = \begin{cases} \varepsilon, & x = \frac{1}{2} \\ 0, & x \neq \frac{1}{2} \end{cases} \quad \text{with } g = h = 0$$

$$\begin{aligned} U_{i,j+1} &= \lambda U_{i-1,j} + (1 - 2\lambda)U_{i,j} + \lambda U_{i+1,j}, & \lambda &= k/h^2 \\ |U_{i,j+1}| &= \lambda |U_{i-1,j}| + (2\lambda - 1)|U_{i,j}| + \lambda |U_{i+1,j}|, & 1 \leq i &\leq N - 1. \end{aligned}$$

Here the equality holds because the sign of  $U_i$  alternates, so the three terms have the same sign. Hence

$$\sum_{i=1}^{N-1} |U_{i,j+1}| = \lambda \sum_{i=1}^{N-1} |U_{i-1,j}| + (2\lambda - 1) \sum_{i=1}^{N-1} |U_{i,j}| + \lambda \sum_{i=1}^{N-1} |U_{i+1,j}|.$$

Since  $U(x_i, t) = 0, i = 1, N$ , we can add them and

$$\sum_{i=1}^{N-1} |U_{i,j+1}| = \lambda \sum_{i=0}^{N-2} |U_{i,j}| + (2\lambda - 1) \sum_{i=1}^{N-1} |U_{i,j}| + \lambda \sum_{i=2}^N |U_{i,j}|.$$

Let  $S(t_j) = \sum_{i=1}^N |U(i, j)|$ . Then since the number of nonzero  $U_{i,j}$  for each  $j$  is  $2j + 1$  (Check the numerical scheme, you will see solution is alternating along  $x$ -direction dispersing both direction),  $U(1, j) = 0, U(N - 1, j) = 0$  for  $2j + 1 < N$ , we see

$$S(t_{j+1}) = (4\lambda - 1)S(t_j) = (4\lambda - 1)^2 S(t_{j-1}) = \dots = (4\lambda - 1)^{j+1} S(0) = (4\lambda - 1)^{j+1} \varepsilon.$$

By the same reason, there is a point  $(x_p, t_j)$  such that

$$|U(x_p, t_j)| \geq \frac{1}{2j + 1} S(t_j) = \frac{1}{2j + 1} (4\lambda - 1)^j \cdot \varepsilon$$

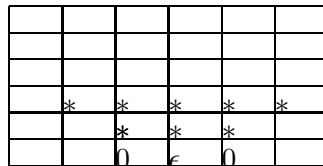


Figure 1.11: Nonzero point

which diverges as  $j \rightarrow \infty$  since  $4\lambda - 1 > 1$ .

Considering the alternating sign, one can see the solution alternates: For  $j = 1$ , we see

$$U_{i,1} = (1 - 2\lambda)\epsilon, \quad U_{i-1,1} = \lambda\epsilon, \quad U_{i+1,1} = \lambda\epsilon$$

$$U_{i,2} = 2\lambda^2\epsilon + (1 - 2\lambda)^2\epsilon, \quad U_{i-1,2} = (1 - 2\lambda)\epsilon + (1 - 2\lambda)\epsilon = 3\lambda\epsilon(1 - 2\lambda) < 0.$$

### Stability of linear system

$$\begin{pmatrix} U_{1,j+1} \\ \vdots \\ U_{N-1,j+1} \end{pmatrix} = \begin{pmatrix} 1-2\lambda & \lambda & \dots & 0 \\ \lambda & 1-2\lambda & \ddots & \\ 0 & & \ddots & \lambda \\ & & \lambda & 1-2\lambda \end{pmatrix} \begin{pmatrix} U_{1,j} \\ \vdots \\ U_{N-1,j} \end{pmatrix} + \begin{pmatrix} g(t_j) \\ 0 \\ \vdots \\ 0 \\ h(t_j) \end{pmatrix}$$

In vector form,  $\mathbf{U}_{j+1} = \mathbf{A}\mathbf{U}_j + \mathbf{G}_j$ . Assume  $\mathbf{G}_j = \mathbf{0}$ ,  $j = 1, 2, \dots$ . Let  $\mu$  be an eigenvalue of  $A$ . Then by G-disk theorem,

$$\begin{aligned} |1 - 2\lambda - \mu| &\leq 2\lambda \\ -2\lambda &\leq 1 - 2\lambda - \mu \leq 2\lambda \\ -2\lambda &\leq -1 + 2\lambda + \mu \leq 2\lambda \\ 1 - 4\lambda &\leq \mu \leq 1. \end{aligned}$$

If  $0 < \lambda \leq \frac{1}{2}$ , then  $-1 \leq \mu \leq 1$ , hence stable. If  $\lambda > \frac{1}{2}$ , then  $|\mu| > 1$  is possible. So the scheme may be unstable. The following example show it is actually unstable.

**Example 1.4.6** (Issacson, Keller). Let  $v(x, t) = \operatorname{Re}(e^{i\alpha x - wt}) = \cos \alpha x \cdot e^{-wt}$ .

$$\begin{aligned}
 v_t - v_{xx} &\doteq \frac{v(x, t + \Delta t) - v(x, t)}{\Delta t} - \frac{v(x + \Delta x, t) - 2v(x, t) + v(x - \Delta x, t)}{\Delta x^2} \\
 &= v(x, t) \left( \frac{e^{-w\Delta t} - 1}{\Delta t} \right) - \frac{\cos(\alpha x + \alpha \Delta x) - 2 \cos \alpha x + \cos(\alpha x - \alpha \Delta x)}{\Delta x^2} e^{-wt} \\
 &= v(x, t) \left( \frac{e^{-w\Delta t} - 1}{\Delta t} - \frac{2 \cos \alpha \Delta x - 2}{\Delta x^2} \right) \\
 &= v(x, t) \frac{1}{\Delta t} \{ e^{-w\Delta t} - [(1 - 2\lambda) + 2\lambda \cos \alpha \Delta x] \} \\
 &= v(x, t) \frac{1}{\Delta t} \left[ e^{-w\Delta t} - \left( 1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2} \right) \right]
 \end{aligned}$$

Thus  $v$  is a solution of the difference equation provided  $w$  and  $\alpha$  satisfy  $e^{-w\Delta t} = 1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2}$ .

With I.C.  $v(x, 0) = \cos \alpha x$ , solution becomes

$$v(x, t) = \cos \alpha x e^{-wt} = \cos \alpha x \left( 1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2} \right)^{\frac{t}{\Delta t}}$$

Clearly,  $|v(x, t)| \leq 1$ , for all  $\lambda \leq \frac{1}{2}$ . However, if  $\lambda > \frac{1}{2}$ , then we have  $|1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2}| > 1$ , for some  $\Delta x$ . So  $v(x, t)$  becomes arbitrarily large for sufficiently large  $t/\Delta t$ . Since every even function has a cosine series, we may express any even function  $f(x)$  in the form  $f(x) = \sum_n \alpha_n \cos(\alpha \pi x)$  to get an unstable problem. (See book for details)

### Implicit Finite Difference Method.

Given a heat equation

$$\begin{aligned}
 u_t &= u_{xx} \\
 u(0, t) &= g(t), \quad t > 0 \\
 u(1, t) &= h(t) \\
 u(x, 0) &= f(x), \quad 0 \leq x \leq 1.
 \end{aligned}$$

We discretize it by implicit difference method. Fix a time level  $j$ . Then we have

$$\frac{U_{i,j+1} - U_{i,j}}{\Delta t} = \frac{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}}{\Delta x^2} \quad i = 1, \dots, N-1.$$

Multiply by  $\Delta t$ , then with  $\lambda = \Delta t / \Delta x^2$ , we have

$$\begin{aligned} U_{i,j+1} - U_{i,j} &= \lambda U_{i+1,j+1} - 2\lambda U_{i,j+1} + \lambda U_{i-1,j+1} \\ -U_{i,j} &= \lambda U_{i+1,j+1} - (1 + 2\lambda)U_{i,j+1} + \lambda U_{i-1,j+1}. \end{aligned}$$

This yields a system of equations in  $N - 1$  unknowns  $\{U_{i,j+1}\}_{i=1}^{N-1}$  :

$$A\mathbf{U}_{j+1} = \mathbf{U}_j + \mathbf{G}_j, \quad (1.20)$$

where

$$A = - \begin{bmatrix} (1 + 2\lambda) & -\lambda & & & \\ -\lambda & (1 + 2\lambda) & -\lambda & & \\ & \ddots & \ddots & -\lambda & \\ & & 0 & -\lambda & (1 + 2\lambda) \end{bmatrix}, \quad \mathbf{U}_j = \begin{bmatrix} U_{1,j} \\ \vdots \\ U_{N-1,j} \end{bmatrix}, \quad \mathbf{G}_j = \begin{bmatrix} -\lambda U_{0,j+1} \\ 0 \\ \vdots \\ 0 \\ -\lambda U_{N,j+1} \end{bmatrix}$$

**Theorem 1.4.7.** *The implicit finite difference scheme is stable for all  $\lambda = \Delta t / \Delta x^2$ . (The solution remains bounded).*

*Proof.* Fix a  $j$  and let  $U_{k(j),j}$  be chosen so that  $|U_{k(j),j}| \geq |U_{i,j}|$ ,  $i = 1, \dots, N - 1$ . We choose  $i_0 = k(j + 1)$  in the following relation.

$$U_{i,j+1} = U_{i,j} + \lambda\{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}\}.$$

Then

$$(1 + 2\lambda)U_{i_0,j+1} = U_{i_0,j} + \lambda\{U_{i_0+1,j+1} + U_{i_0-1,j+1}\}.$$

Taking absolute values,

$$(1 + 2\lambda)|U_{i_0,j+1}| \leq |U_{i_0,j}| + \lambda(|U_{i_0+1,j+1}| + |U_{i_0-1,j+1}|) \leq |U_{i_0,j}| + 2\lambda|U_{i_0,j+1}|.$$

Thus  $|U_{i_0,j+1}| \leq |U_{i_0,j}| \leq |U_{k(j),j}|$  and hence  $|U_{i,j+1}| \leq |U_{i_0,j+1}| \leq |U_{k(j),j}|$  for  $1 \leq i \leq N - 1$ . Repeat the same procedure until  $j = 0$ .

$$|U_{i,j+1}| \leq |U_{k(j),j}| \leq \dots \leq |U_{k(0),0}| \leq M = \max(f, g, h), \quad \text{for } 1 \leq i \leq N - 1.$$

This relation also holds for  $i = 0$  or  $N$  also.  $\square$



### Matrix formulation

We check the eigenvalues of the system

$$A\mathbf{U}_{j+1} = \mathbf{U}_j + \mathbf{G}_j.$$

Eigenvalue of  $A$  satisfies  $|\mu + (1 + 2\lambda)| \leq 2\lambda$  by  $G$ -disk theorem. From this, we see  $|\mu| \geq 1$  and hence the eigenvalues of  $A^{-1}$  is less than one in absolute value. Thus

$$\mathbf{U}_{j+1} = \mathbf{A}^{-1}(\mathbf{U}_j + \mathbf{G}_j) = \cdots = \mathbf{A}^{-j-1}\mathbf{U}_0 + \mathbf{A}^{-j-1}\mathbf{G}_0 + \mathbf{A}^{-j-2}\mathbf{G}_1 + \cdots + \mathbf{A}^{-1}\mathbf{G}_j.$$

$$\|\mathbf{U}_{j+1}\| \leq \|\mathbf{A}^{-j-1}\| \|\mathbf{U}_0\| + \|\mathbf{A}^{-1}\| \cdot \frac{1}{1 - \|\mathbf{A}^{-1}\|} \max_j \|\mathbf{G}_j\|$$

remain bounded.

**Note.**  $A$  does not have  $-1$  as eigenvalues and all the eigenvalues are positive real.

**Theorem 1.4.8.** *For sufficiently smooth  $u$ , we have*

$$|u_{ij} - U_{ij}| = \mathcal{O}(h^2 + k) \text{ as } h \text{ and } k \rightarrow 0 \text{ (for all } \lambda).$$

*Proof.* Let  $u_{ij} = u(x_i, t_j)$  be the true solution. Then we have

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{1}{h^2} \{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}\} + \mathcal{O}(h^2 + k).$$

Let  $w_{i,j} = u_{i,j} - U_{i,j}$  be the discretization error. Then

$$\begin{aligned} w_{i,j+1} &= w_{i,j} + \lambda \{w_{i+1,j+1} - 2w_{i,j+1} + w_{i-1,j+1}\} + \mathcal{O}(kh^2 + k^2) \\ (1 + 2\lambda)w_{i,j+1} &= w_{i,j} + \lambda w_{i+1,j+1} + \lambda w_{i-1,j+1} + \mathcal{O}(kh^2 + k^2). \end{aligned}$$

Let  $\|w_j\| = \max_i |w_{i,j}|$ . Then for all  $i$

$$(1 + 2\lambda)|w_{i,j+1}| \leq \|w_j\| + 2\lambda\|w_{j+1}\| + \mathcal{O}(kh^2 + k^2)$$

and so

$$(1 + 2\lambda)\|w_{j+1}\| \leq \|w_j\| + 2\lambda\|w_{j+1}\| + \mathcal{O}(kh^2 + k^2).$$

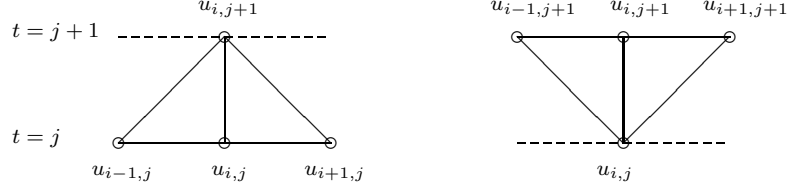


Figure 1.12: Stencil for forward, backward Euler method

Thus

$$\begin{aligned}
 \|w_{j+1}\| &\leq \|w_j\| + C(kh^2 + k^2) \\
 &\leq \cdots \leq \|w_0\| + C(j+1)k(k+h^2) \\
 &\leq \|w_0\| + CT(k+h^2) = CT(k+h^2)
 \end{aligned}$$

for  $t = (j+1)k \leq T$ . □

#### 1.4.1 Discretization of parabolic p.d.e, General Case

Consider

$$\begin{aligned}
 c \frac{\partial u}{\partial t} &= \frac{\partial}{\partial x} \left( p \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left( p \frac{\partial u}{\partial y} \right) - \gamma u + f \quad \text{in } \Omega \times (0, T) \\
 \text{I.C. } u(x, y, 0) &= h(x, y) \text{ in } \Omega \\
 \text{B.C. } u(x, y, t) &= g(x, y, t) \text{ for } (x, y) \in \partial\Omega.
 \end{aligned}$$

where  $c, p, \gamma, f$  are functions of  $x, y$  and  $t$ . Assume

$$\begin{aligned}
 0 < p_0 &\leq p(x, y, t) \leq p_1 \\
 0 &\leq \gamma(x, y, t) \leq \gamma \\
 0 < c_0 &\leq c(x, y, t) \leq c_1.
 \end{aligned}$$

Use central difference for  $pu_x$  at  $(i \pm 1/2, j)$  and  $pu_y$  at  $(i, j \pm 1/2)$ . With  $U_{i,j}^n = U(x_i, y_j, t_n)$  we let

$$\begin{aligned}
 M_h U_{i,j}^n &:= \frac{1}{\Delta x^2} \left( p_{i+1/2,j}^n U_{i+1,j}^n + p_{i-1/2,j}^n U_{i-1,j}^n - (p_{i+1/2,j}^n + p_{i-1/2,j}^n) U_{i,j}^n \right) \\
 &+ \frac{1}{\Delta y^2} \left( p_{i,j+1/2}^n U_{i,j+1}^n + p_{i,j-1/2}^n U_{i,j-1}^n - (p_{i,j+1/2}^n + p_{i,j-1/2}^n) U_{i,j}^n \right) - \gamma_{i,j}^n U_{i,j}^n
 \end{aligned}$$

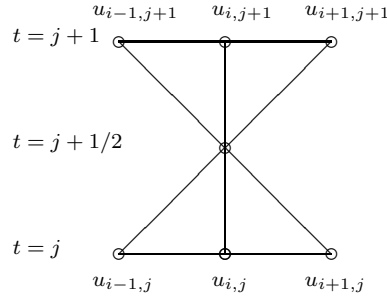


Figure 1.13: Stencil for Crank-Nicolson method

Now we consider explicit and implicit scheme at one stroke. For  $0 \leq \theta \leq 1$ , we let

$$[\theta c_{ij}^{n+1} + (1-\theta)c_{ij}^n] \frac{U_{i,j}^{n+1} - U_{i,j}^n}{\Delta t} = \theta M_h U_{i,j}^{n+1} + (1-\theta)M_h U_{i,j}^n + \theta f_{ij}^{n+1} + (1-\theta)f_{ij}^n$$

For  $\theta = 0$ , we have

$$c_{ij}^n \frac{U_{i,j}^{n+1} - U_{i,j}^n}{\Delta t} = M_h U_{i,j}^n + f_{ij}^n \quad \text{forward Euler}$$

For  $\theta = 1$ , we have

$$c_{ij}^{n+1} \frac{U_{i,j}^{n+1} - U_{i,j}^n}{\Delta t} = M_h U_{i,j}^{n+1} + f_{ij}^{n+1} \quad \text{backward Euler}$$

For  $\theta = \frac{1}{2}$ , we have Crank-Nicolson.

### Matrix formulations

For  $\theta = 1$ ,

$$\left( \left( \frac{c_{ij}^{n+1}}{\Delta t} \right) I - M_h \right) \vec{U}^{n+1} = \left( \frac{c_{ij}^{n+1}}{\Delta t} \right) \vec{U}^n + \vec{F}^n$$

For  $\theta = \frac{1}{2}$ ,

$$\left( \left( \frac{\bar{c}_{ij}}{\Delta t} \right) I - \frac{1}{2} M_h \right) \vec{U}^{n+1} = \left( \frac{\bar{c}_{ij}}{\Delta t} + \frac{1}{2} M_h \right) \vec{U}^n + \frac{1}{2} (\vec{F}^n + \vec{F}^{n+1}),$$

where  $\bar{c}_{ij} = \frac{1}{2}(c_{ij}^{n+1} + c_{ij}^n)$ .

**Exercise 1.4.9.** (1) Show the Crank-Nicolson scheme is stable for all  $\lambda$ .

(2) Show that with  $\theta = 1/2$ , the truncation error is  $O(\Delta t^2 + \Delta x^2)$ .

(3) Consider a heat equation

$$\begin{aligned} u_t &= \kappa u_{xx} & 0 < x < 1, \quad t > 0 \\ u(x, 0) &= f(x) \\ u(0, t) &= g(t) \\ u(1, t) &= h(t), \end{aligned}$$

where  $\kappa > 0$  is constant, say  $\kappa = 1$ . If  $f(x) = \cos \pi x$ ,  $g(t) = e^{-\pi^2 \kappa t}$ ,  $h(t) = -e^{-\pi^2 \kappa t}$  it has solution  $u = e^{-\pi^2 \kappa t} \cos \pi x$ . (For nonseparable example, just add linear function in  $x$  or use (shifted) fundamental solution of heat equation) Use the following method to compute numerical solution up to  $T = 1.0$ . Check  $\|u - U\|_2$  (or  $\|u - U\|_\infty$ ).

- (a) Explicit FDM with  $h = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \dots$ . Choose time step  $\Delta t$  so that  $\lambda = k/h^2 = 0.4$  and  $0.6$ .
- (b) Implicit FDM with  $h = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \dots$ . Choose time step  $\Delta t$  so that  $\lambda = k/h^2 = 0.4, 0.6, 0.8$  and  $1.6$ .
- (c) Crank-Nicholson scheme with  $k = Ch, C = 1, 2, 3, \dots 10$ .

You can use either Gauss-Seidel type of iteration method or LU-decomposition to solve the system of equations arising in the implicit method.

### Truncation error - 1 D.

Assume  $c = 1$ . If we use central difference at  $(x_i, t_{j+\frac{1}{2}})$  then we get

$$\frac{u_{i,j+1} - u_{i,j}}{k} = u_t(x_i, t_{j+\frac{1}{2}}) + O(k^2). \quad (1.21)$$

and

$$\begin{aligned} M_h u_{i,j+1} &= u_{xx}(x_i, t_{j+1}) + O(h^2) = u_{xx}(x_i, t_{j+\frac{1}{2}}) + \frac{k}{2} u_{xxt}(x_i, t_{j+\frac{1}{2}}) + O(k^2 + h^2) \\ M_h u_{i,j} &= u_{xx}(x_i, t_j) + O(h^2) = u_{xx}(x_i, t_{j+\frac{1}{2}}) - \frac{k}{2} u_{xxt}(x_i, t_{j+\frac{1}{2}}) + O(k^2 + h^2) \end{aligned}$$

we see

$$\theta M_h u_{i,j+1} + (1 - \theta) M_h u_{i,j} \quad (1.22)$$

$$= u_{xx}(x_i, t_{j+\frac{1}{2}}) + \frac{k}{2} (2\theta - 1) u_{xxt}(x_i, t_{j+\frac{1}{2}}) + O(k^2 + h^2) \quad (1.23)$$

By (1.21), (1.22) we have

$$\frac{u_{i,j+1} - u_{i,j}}{k} - \theta M_h u_{i,j} - (1 - \theta) M_h u_{i,j+1} \quad (1.24)$$

$$\begin{aligned} &= u_t(x_i, t_{j+\frac{1}{2}}) - u_{xx}(x_i, t_{j+\frac{1}{2}}) - \frac{k}{2}(2\theta - 1)u_{xxt}(x_i, t_{j+\frac{1}{2}}) + O(k^2 + h^2) \\ &= -\frac{k}{2}(2\theta - 1)u_{xxt}(x_i, t_{j+\frac{1}{2}}) + O(k^2 + h^2) \end{aligned} \quad (1.25)$$

Hence the truncation error satisfies

$$\tau(x, t) = \frac{k}{2}(2\theta - 1)u_{xxt}(x_i, t_{j+\frac{1}{2}}) + O(k^2 + h^2), \quad (1.26)$$

where  $O = O(u_{tt}, u_{xxx})$ .

## 1.5 Finite element method for parabolic problems

Let  $I = (0, T)$ .

$$\frac{\partial u}{\partial t} - \Delta u = f \text{ in } \Omega \times I \quad (1.27)$$

$$u = 0 \text{ in } \Gamma \times I \quad (1.28)$$

$$u(x, 0) = u^0 \text{ in } \Omega. \quad (1.29)$$

We shall study two methods: Semi-discretization(discretization in space only) and Full-discretization(discretization in time and space).

### 1.5.1 One dimensional model problem

$$\frac{\partial u}{\partial t} - \alpha^2 \frac{\partial^2 u}{\partial x^2} = f, \quad (x, t) \in (0, L) \times I \quad (1.30)$$

$$u(0, t) = u(\pi, t) = 0, \quad t \in I \quad (1.31)$$

$$u(x, 0) = u^0(x), \quad x \in (0, L). \quad (1.32)$$

For simplicity, assume  $\alpha = 1$ ,  $L = \pi$  and  $f = 0$ . Using the periodic B.C., let us express  $u$  in terms of its Fourier sine series:

$$u(x, t) = \sum_{j=1}^{\infty} c_j e^{-j^2 t} \sin(jx),$$

where  $c_j = \sqrt{\frac{2}{\pi}} \int_0^\pi u^0(x) \sin(jx) dx$ . Generally, the solution is given by

$$u(x, t) = \sum_{j=1}^{\infty} c_j e^{-j^2 \pi \alpha^2 t / L} \sin\left(\frac{j\pi x}{L}\right),$$

$c_j = \sqrt{\frac{2}{L}} \int_0^L u^0(x) \sin\left(\frac{j\pi x}{L}\right) dx$ . We see that  $u$  is a linear combination of sine waves with amplitude  $c_j e^{-j^2 t}$ . If  $j^2 t$  is large,  $e^{-j^2 t} \approx 0$ , (i.e, high frequency component quickly damps out and  $u(x, t)$  becomes smoother) which happens if  $j$  is large or  $t$  is large. This phenomena is consistent with the nature of diffusion process such as heat conduction. However, when  $t$  is close to 0,  $u$  is not smooth. It is known that

$$\|\dot{u}\|_{L^2(0, \pi)} = O(t^{-s}).$$

The nature is like this: The smoother the initial function is, the more rapidly  $c_j$  decays as  $j \rightarrow \infty$ . An initial phase when  $\dot{u}$  is large, is called an initial transient. This will affect on choosing the time step size. Basic Stability.

$$\|u(\cdot, t)\| \leq \|u^0\|, t \in I \quad (1.33)$$

$$\|\dot{u}(\cdot, t)\| \leq \frac{C}{t} \|u^0\|, t \in I. \quad (1.34)$$

## 1.6 Semi discretization in space

Let  $V = H_0^1(\Omega)$ . Multiply the equation by  $v$

$$(\dot{u}(t), v) + a(u(t), v) = (f(t), v), \quad v \in V \quad (1.35)$$

$$(u(0), v) = (u^0, v). \quad (1.36)$$

Let  $V_h = \text{Span}\{\phi_1, \dots, \phi_N\}$ . Then the finite element formulation is

$$(\dot{u}_h(t), v) + a(u_h(t), v) = (f(t), v) \quad (1.37)$$

$$(u_h(0), v) = (u^0, v), \quad v \in V_h. \quad (1.38)$$

With  $u_h(x, t) = \sum_{i=1}^N \xi_i(t) \phi_i(x)$ ,

$$\begin{aligned} \sum_i \xi_i'(t) (\phi_i, \phi_j) + \sum_i \xi_i(t) a(\phi_i, \phi_j) &= (f(t), \phi_j) \\ \sum_i \xi_i(0) (\phi_i, \phi_j) &= (u^0, \phi_j), \quad j = 1, \dots, N. \end{aligned}$$

In matrix form

$$\begin{aligned} B\dot{\xi}(t) + A\xi(t) &= F(t) \\ B\xi(0) &= U^0 \end{aligned}$$

where  $B = (b_{ij})$ ,  $A = (a_{ij})$ ,  $F = (F_j)$ ,  $\xi = (\xi_i)$ ,  $U^0 = (U_j^0)$ .  $b_{ij} = (\phi_j, \phi_i)$ ,  $A_{ij} = a(\phi_j, \phi_i)$ , etc.

Mass matrix  $B$  and stiffness matrix  $A$  is SPD.  $\kappa(B) = O(1)$ ,  $\kappa(A) = O(h^{-2})$ . With Cholesky decomposition  $B = L^T L$ ,  $\eta = L\xi$ , we can reduce it as

$$\dot{\eta} + \bar{A}\eta(t) = g(t) \quad (1.39)$$

$$\eta(0) = \eta^0 \quad (1.40)$$

where  $\bar{A} = L^{-T} A L^{-1}$  is also SPD.  $g = L^{-T} F$ ,  $\eta^0 = L^{-T} U^0$ . The solution is given by

$$\eta(t) = e^{-\bar{A}t} \eta^0 + \int_0^t e^{-\bar{A}(t-s)} g(s) ds.$$

Stability: Take  $v = u_h(x, t) := u_h(t)$  in (1.37) with  $f = 0$

$$\begin{aligned} (\dot{u}_h(t), u_h(t)) + a(u_h(t), u_h(t)) &= 0 \\ \frac{1}{2} \frac{d}{dt} \|u_h(t)\|^2 + a(u_h(t), u_h(t)) &= 0 \\ \|u_h(t)\|^2 + 2 \int_0^t a(u_h(s), u_h(s)) ds &= \|u_h(0)\|^2 \leq \|u^0\|^2. \end{aligned}$$

In particular,

$$\|u_h(t)\| \leq \|u_h(0)\| \leq \|u^0\|, \quad t > 0.$$

**Theorem 1.6.1.** *c There is a constant  $C$  such that*

$$\max_{t \in I} \|u(t) - u_h(t)\| \leq C \left( 1 + \ln \frac{T}{h^2} \right) \max_{t \in I} h^2 \|u(t)\|_{H^2}. \quad (1.41)$$

## 1.7 Fully discrete Scheme

We shall now discretize the scheme in time also. Here, we use finite difference method to discretize along time while we maintain finite element discretization along space. First consider a related problem (1.37).

To see the behavior of  $\eta(t)$ , we write

$$\eta(t) = \sum_{i=1}^N (\eta^0, \chi_i) e^{-\mu_n t} \chi_i,$$

where  $\{\chi_i\}$  is an orthonormal eigenvectors of  $\bar{A}$ ,

$$\mu_1 \leq \dots \leq \mu_M, \quad \mu_1 = O(1), \quad \mu_M = O(h^{-2}).$$

This again has an initial transient. For accuracy, we need to take small time step, or use implicit method.

$$0 = t_0 < t_1, \dots, < t_N = T, \quad \Delta t_n = t_n - t_{n-1}.$$

Forward Euler method:

$$\begin{aligned} \left( \frac{u_h^{n+1} - u_h^n}{\Delta t_{n+1}}, v \right) + a(u_h^n, v) &= (f(t_n), v), \quad v \in V_h, n = 1, 2, \dots \\ (u_h^0, v) &= (u^0, v). \end{aligned}$$

Backward Euler method:

$$\begin{aligned} \left( \frac{u_h^{n+1} - u_h^n}{\Delta t_{n+1}}, v \right) + a(u_h^{n+1}, v) &= (f(t_{n+1}), v), \quad v \in V_h, n = 1, 2, \dots \quad (1.42) \\ (u_h^0, v) &= (u^0, v). \end{aligned}$$

Discretization error is  $O(\Delta t_n)$ . Let  $u_h(x, t) = \sum_{i=1}^M \xi_i(t) \phi_i(x)$ . Then taking  $v = \phi_j$

$$\left( \sum_i \xi_i^{n+1} \phi_j, \phi_j \right) + \Delta t_{n+1} a \left( \sum_i \xi_i^{n+1} \phi_j, \phi_j \right) = \left( \sum_i \xi_i^n \phi_j, \phi_j \right) + \Delta t_{n+1} (f(t_{n+1}), \phi_j) \quad (1.43)$$

Or

$$(B + \Delta t_{n+1} A) \xi^{n+1} = B \xi^n + \Delta t_{n+1} F(t_{n+1}). \quad (1.44)$$



For stability with  $f = 0$ , take  $v = u_h^{n+1}$  in (1.42)

$$(u_h^{n+1}, u_h^{n+1}) - (u_h^{n+1}, u_h^n) + \Delta t_{n+1} a(u_h^{n+1}, u_h^{n+1}) = 0.$$

Using arithmetic-geometric inequality  $(u, v) \leq \frac{1}{2}(\epsilon \|u\|^2 + \frac{\|v\|^2}{\epsilon})$

$$\frac{1}{2}(\|u_h^{n+1}\|^2 - \|u_h^n\|^2) + \Delta t_{n+1} a(u_h^{n+1}, u_h^{n+1}) \leq 0.$$

Summing up

$$\|u_h^{n+1}\|^2 + 2\Delta t_{n+1} a(u_h^{n+1}, u_h^{n+1}) \leq \|u_h^0\|^2 \leq \|u^0\|^2.$$

In particular

$$\|u_h^{n+1}\| \leq \|u_h^0\| \leq \|u^0\|, n = 1, 2, \dots \quad (1.45)$$

Now Crank-Nicholson. We use combination of forward and backward Euler scheme.

$$\begin{aligned} \left( \frac{u_h^{n+1} - u_h^n}{\Delta t_{n+1}}, v \right) + a \left( \frac{u_h^{n+1} + u_h^n}{2}, v \right) &= \left( \frac{f(t_{n+1}) + f(t_n)}{2}, v \right), \\ (u_h^0, v) &= (u^0, v), \quad v \in V_h, n = 1, 2, \dots \end{aligned} \quad (1.46)$$

Discretization error is  $O(\Delta t_n^2)$

Taking  $v = (u_h^{n+1} + u_h^n)/2$  we obtain the stability as before. In matrix form

$$\left( B + \frac{\Delta t_{n+1}}{2} A \right) \xi^{n+1} = \left( B - \frac{\Delta t_{n+1}}{2} A \right) \xi^n + \Delta t_{n+1} \frac{\bar{F}(t_{n+1}) + \bar{F}(t_n)}{2} \quad (1.47)$$

As in (1.39) we use  $\eta = L\xi$  associated with Cholesky-decomposition, then the transformed equation becomes

$$\frac{\eta^{n+1} - \eta^n}{\Delta t_{n+1}} + \frac{1}{2} \bar{A} (\eta^{n+1} + \eta^n) = \frac{1}{2} (g(t_{n+1}) + g(t_n)). \quad (1.48)$$

In case  $g = 0$ , we have

$$\left( I + \frac{1}{2} \Delta t_{n+1} \bar{A} \right) \eta^{n+1} = \left( I - \frac{1}{2} \Delta t_{n+1} \bar{A} \right) \eta^n.$$

and

$$\left\| \left( I + \frac{1}{2} \Delta t_{n+1} \bar{A} \right)^{-1} \left( I - \frac{1}{2} \Delta t_{n+1} \bar{A} \right) \right\| = \max_j \frac{|1 - \frac{1}{2} \Delta t_{n+1} \lambda_j|}{1 + \frac{1}{2} \Delta t_{n+1} \lambda_j} < 1.$$

Thus the scheme is stable for all time step  $\Delta t_n$  (unconditionally stable). However, for backward-Euler or Crank-Nicholson method, one has to solve a system of linear equation for each time step, which is costly.

For the forward Euler method, one can compute  $\eta^{n+1}$  directly without solving any system, but for stability, one has to take small time step. In fact, one can show that

$$\eta^{n+1} = (I - \Delta t_{n+1} \bar{A}) \eta^n$$

and

$$\|I - \Delta t_{n+1} \bar{A}\| = \max_j |1 - \Delta t_{n+1} \lambda_j| \leq 1$$

only if  $\Delta t_{n+1} \lambda_M \leq 2$  or  $\Delta t_{n+1} = O(h^2)$  (conditionally stable). Hence for the forward Euler method the stability is guaranteed only when time step is very small even for moderate  $h$ .

## 1.8 Hyperbolic Equation

Consider a first order system of hyperbolic equations:

$$\begin{cases} u_t + a_{11}u_x + a_{12}v_x = b_1 \\ v_t + a_{21}u_x + a_{22}v_x = b_2 \end{cases} \quad (1.49)$$

where  $a_{i,j}$  and  $u, v$  and function of  $(x, t)$ .

$$\begin{aligned} u(x, 0) &= f(x) \\ v(x, 0) &= g(x), \quad -\infty < x < \infty. \end{aligned}$$

Let  $\mathbf{u} = [u, v]^T$ ,  $\mathbf{b} = [b_1, b_2]^T$ ,  $A = (a_{ij})$ . Then the D.E. is of the form

$$\mathbf{u}_t + A\mathbf{u}_x = \mathbf{b}.$$

Equation (1.49) is called hyperbolic, if there exist a  $P$  such that  $P^{-1}AP = \text{diag}\{\lambda_1, \lambda_2\}$  where  $\lambda_i(x, t)$  are real and distinct. Let  $\mathbf{z} = [z_1, z_2]^T$  be defined

by  $\mathbf{u} = P\mathbf{z}$ , then

$$\begin{aligned} P_t\mathbf{z} + P\mathbf{z}_t + A(P_x\mathbf{z} + P\mathbf{z}_x) &= \mathbf{b} \\ P\mathbf{z}_t + AP\mathbf{z}_x &= \mathbf{b} - (P_t + AP_x)\mathbf{z} \\ \therefore \mathbf{z}_t + P^{-1}AP\mathbf{z}_x &= P^{-1}\{\mathbf{b} - (P_t + AP_x)\mathbf{z}\} = \beta(x, t, \mathbf{z}). \end{aligned}$$

Componentwise,

$$(z_i)_t + \lambda_i(z_i)_x = \beta_i, \quad i = 1, 2.$$

Let  $x_i(t), i = 1, 2$  be the solution of the o.d.e.

$$\frac{dx_i}{dt} = \lambda_i(x_i, t) \quad \text{such that} \quad x_i(t^*) = x^*.$$

Let  $z_i(t) \equiv z_i(x_i(t), t)$  be defined along the curve  $\frac{dx_i}{dt} = \lambda_i(x_i, t)$  (called characteristics). Then

$$\frac{dz_i}{dt} = \frac{\partial z_i}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z_i}{\partial t} = \lambda_i(x_i(t), t) \frac{\partial z_i}{\partial x} + \frac{\partial z_i}{\partial t} = \beta_i(x_i, t, \mathbf{z})$$

Thus  $z_i(x_i(t), t)$  solve the o.d.e (p.d.e on the characteristics) with

$$z_i(0) = z_i(x_i(0), 0) = (P^{-1}\mathbf{u})_i(x_i(0), 0).$$

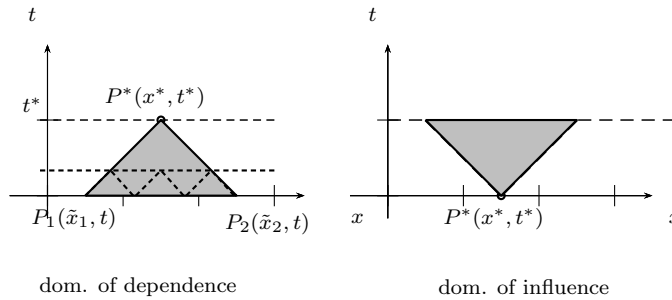


Figure 1.14: Domain of depend/influence

The shaded part is called the “Domain of dependence” of  $(x^*, t^*)$  and its base is called the “interval of dependence”.

A necessary condition for convergence: The numerical domain of dependence must contain the analytic domain of dependence. If the grid point is only in the inner region of the domain of dependence, then changing  $f$  by

$f + \delta, g$  by  $d + \delta$  near the boundary yields the same (numerical) solution.

**Example 1.8.1.** Plucked string of wave

$$\begin{cases} u_t = cv_x \\ v_t = cu_x \end{cases}$$

$$\begin{aligned} u(x, 0) &= f(x) \\ v(x, 0) &= \frac{1}{c} \int_0^x G(\sigma) d\sigma = g(x) \end{aligned}$$

$$\begin{cases} u_{tt} - c^2 u_{xx} = 0 \\ u(x, 0) = f(x) \\ u_t(x, 0) = cv_x(x, 0) = G(x) \end{cases}$$

$$\begin{bmatrix} u \\ v \end{bmatrix}_t + \begin{bmatrix} 0 & -c \\ -c & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}_x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Eigenvalues of  $A$  are  $\pm c$ . Corresponding to the eigenvector  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  Therefore,  $P = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ .

$$P^{-1}AP = \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix} \quad \text{and} \quad P^{-1} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

If  $\mathbf{u} = P\mathbf{z}$  ( $\mathbf{u} = (u(x, t), v(x, t))$ ,  $\mathbf{z} = (z_1(x, t), z_2(x, t))$ ), then

$$\mathbf{z}_t + \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix} \mathbf{z}_x = 0 \Rightarrow \begin{aligned} \frac{\partial z_1}{\partial t} + c \frac{\partial z_1}{\partial x} &= 0 \\ \frac{\partial z_2}{\partial t} - c \frac{\partial z_2}{\partial x} &= 0 \end{aligned}$$

We check the total derivative  $\frac{Dz_i}{Dt}$  along  $\frac{dx_1}{dt} = c$ , and  $\frac{dx_2}{dt} = -c$ . For example,

$$\frac{Dz_1}{dt} = \frac{\partial z_1}{\partial t} + \frac{dx}{dt} \frac{\partial z_1}{\partial x} = 0 \quad \text{along} \quad \frac{dx_1}{dt} = c \quad (1.50)$$

$$\frac{Dz_2}{dt} = \frac{\partial z_2}{\partial t} + \frac{dx}{dt} \frac{\partial z_2}{\partial x} = 0 \quad \text{along} \quad \frac{dx_2}{dt} = -c. \quad (1.51)$$

Thus if the curve passes a point  $(x^*, t^*)$  in figure 1.14, then solving the charac.

equations (1.50) and (1.51) with I.C.  $x(t^*) = x^*$ , we obtain the characteristics  $x_1(t) = ct + x^* - ct^*$  and  $x_2(t) = -ct + x^* + ct^*$ . Hence along  $x_1(t) = ct + x^* - ct^*$

$$z_1(x(t), t) = z_1(ct + x^* - ct^*, t) = z_1(x^* - ct^*, 0) \text{ since } z_1 \text{ is constant along } x_1(t)$$

Similarly along  $x_2(t) = -ct + x^* + ct^*$

$$z_2(x(t), t) = z_2(-ct + x^* + ct^*, t) = z_2(x^* + ct^*, 0) \text{ since } z_2 \text{ is constant along } x_2(t).$$

$$\mathbf{z} = P^{-1}\mathbf{u} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}$$

$$\begin{aligned} z_1(x, t) &= \frac{1}{2}[u(x, t) - v(x, t)] = \frac{1}{2}[f(x^* - ct^*) - g(x^* - ct^*)] \\ z_2(x, t) &= \frac{1}{2}[u(x, t) + v(x, t)] = \frac{1}{2}[f(x^* + ct^*) + g(x^* + ct^*)] \end{aligned}$$

Here the points  $(x, t)$  lies on the characteristics passing  $(x^*, t^*)$ . Hence

$$\begin{aligned} \begin{bmatrix} u \\ v \end{bmatrix}_{(x^*, t^*)} &= P \cdot \mathbf{z} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} f(x^* - ct^*) - g(x^* - ct^*) \\ f(x^* + ct^*) + g(x^* + ct^*) \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} f(x^* - ct^*) - g(x^* - ct^*) + f(x^* + ct^*) + g(x^* + ct^*) \\ -f(x^* - ct^*) + g(x^* - ct^*) + f(x^* + ct^*) + g(x^* + ct^*) \end{bmatrix} \end{aligned}$$

These are called D'Alembert solutions.

### 1.8.1 Method of Characteristics

Numerical procedure "See R.S. Varga" or "Y. Gregory" Ch16.

Assume we have equations for the characteristics and obtained transformed the pde along the characteristics:

$$\frac{dz_i}{dt} = \beta_i(x_i, t, z_1, z_2) \text{ along } \frac{dx_i}{dt} = \lambda_i(x_i, t), \quad i = 1, 2.$$

Assume  $z_i(t, x)$  is known at  $t$ -th level at all points (say by interpolation).

1st step: Find  $P_1(\tilde{x}_1, t)$ ,  $P_2(\tilde{x}_2, t)$  by

$$\frac{x^* - \tilde{x}_i}{\Delta t} = \lambda_i(x^*, t^*), \quad i = 1, 2 \quad (\text{Backward})$$

2nd step:

$$\frac{z_i(P^*) - z_i(P_i)}{\Delta t} = \beta_i(P_i, z_1(P_i), z_2(P_i)) \quad (\text{Forward})$$

solve for  $z_i(P^*)$ ,  $i = 1, 2$ .

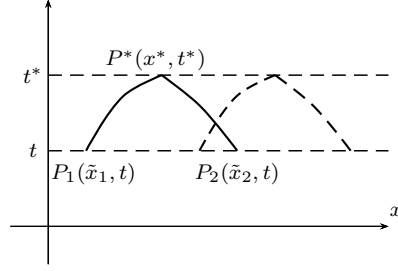


Figure 1.15: Find  $\tilde{x}_1, \tilde{x}_2$  using characteristics

## 1.8.2 FDM for Hyperbolic equations

Given pure initial value problem

$$u_{xx} = u_{tt}, \quad u(x, 0) = f(x), \quad u_t(x, 0) = g(x).$$

For  $j = 0, 1, \dots$

$$\frac{U_{i,j-1} - 2U_{i,j} + U_{i,j+1}}{\Delta t^2} = \frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{\Delta x^2} \quad i = 1, \dots, N-1.$$

$$U_{i,j+1} = m^2(U_{i-1,j} + U_{i+1,j}) + 2(1 - m^2)U_{i,j} - U_{i,j-1}, \quad m = \frac{k}{h} = \frac{\Delta t}{\Delta x} \quad (1.52)$$

I.C. First condition is easy to implement,  $U_{i,0} = f(x_i)$ . For second condition, use

$$g(x_i) = u_t|_{t=0} = \frac{U_{i,1} - U_{i,-1}}{2\Delta t} + O(k^2) \quad (1.53)$$

Thus

$$U_{i,1} - U_{i,-1} = 2kg_i. \quad (1.54)$$

For  $j = 0$

$$U_{i,1} = m^2(U_{i-1,0} + U_{i+1,0}) + 2(1 - m^2)U_{i,0} - U_{i,-1}.$$

Replacing  $U_{i,-1} = U_{i,1} - 2kg_i$ , we have

$$U_{i,1} = \frac{1}{2}m^2(f_{i-1} + f_{i+1}) + (1 - m^2)f_i + kg_i. \quad (1.55)$$

### Discussion of convergence

We first consider stability. From the consideration of characteristics, we have to assume  $|m| \leq 1$ . (Slope of char.) Let  $z_{i,j} = u_{i,j} - U_{i,j}$ . Then

$$z_{i,j+1} = m^2(z_{i-1,j} + z_{i+1,j}) + 2(1 - m^2)z_{i,j} - z_{i,j-1} + O(k^4) + O(k^2h^2). \quad (1.56)$$

If we use (1.53) in the first time step,

$$z_{i,1} = O(k^3). \quad (1.57)$$

To investigate stability, we try to see the effect of a single term  $\exp(\sqrt{-1}\beta x)$ . I.C becomes

$$z_{i,0} = \exp(\sqrt{-1}\beta ih). \quad (1.58)$$

Attempt a solution by separation of variables

$$z_{i,j} = \exp(\alpha jk) \exp(\sqrt{-1}\beta ih). \quad (1.59)$$

Substituting into (1.56) and dropping the truncation error,

$$e^{\alpha k} + e^{-\alpha k} = 2 - 4m^2 \sin^2\left(\frac{1}{2}\beta h\right)$$

which is

$$(e^{\alpha k})^2 - 2(1 - 2m^2 \sin^2\left(\frac{1}{2}\beta h\right))e^{\alpha k} + 1 = 0. \quad (1.60)$$

To avoid increasing solution as  $j \rightarrow \infty$ , it is necessary that  $|e^{\alpha k}| \leq 1$  for all real  $\beta$ . But the product of two solutions of the quadratic equation is 1, hence one of the solutions must exceed 1 unless both are equal to 1 in magnitude. Thus the discriminant must be less than 0,

$$(1 - 2m^2 \sin^2\left(\frac{1}{2}\beta h\right))^2 \leq 1.$$

$$m^2 \leq \frac{1}{\sin^2\left(\frac{1}{2}\beta h\right)}$$

This is always true if

$$m = \Delta t / \Delta x \leq 1. \quad (1.61)$$

A more careful analysis shows (Assume  $m = 1$ )

$$\|z_j\| \leq jBh^3 + \frac{1}{2}j(j-1)Ah^4, \quad (1.62)$$

where  $\|z_j\| = \max_i |z_{i,j}|$ . Since  $t = jh$

$$\|z_j\| \leq tBh^2 + \frac{1}{2}t^2Ah^2. \quad (1.63)$$

where

### 1.8.3 Implicit method for second order hyperbolic equations

Use average of two second central differences:

$$\begin{aligned} U_{i,j+1} - 2U_{i,j} + U_{i,j-1} = & \frac{1}{2}m^2 \{ (U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}) \\ & + (U_{i+1,j-1} - 2U_{i,j-1} + U_{i-1,j-1}) \} \end{aligned} \quad (1.64)$$

$$\begin{aligned} & -m^2U_{i+1,j+1} + 2(1-m^2)U_{i,j+1} - m^2U_{i-1,j+1} \\ & = 4U_{i,j} + m^2U_{i+1,j-1} - 2(1+m^2)U_{i,j-1} + m^2U_{i-1,j-1} \end{aligned} \quad (1.65)$$

Note that this method is applicable for problems with finite domain only.