Math 6008 Numerical PDEs–Lecture 6
FDM for elliptic equations

Instructor: Lei Li, INS, Shanghai Jiao Tong University;
Email: leili2010@sjtu.edu.cn

In this lecture, we continue to investigate the FDM for elliptic equations.

1 Neumann boundary condition

Previously, we considered the Dirichlet boundary condition. Here, look at another type of popular boundary condition, the Neumann boundary condition.

$$-u''(x) = f, u'(0) = \sigma_0, u'(1) = \sigma_1.$$

1.1 The discretization of the boundary condition

How do we approximate $u'(0) = \sigma_0$?

One option is to use the simple one-sided difference:

$$D_+ u = \frac{1}{h}(u_1 - u_0) = \sigma_0.$$

This works but it's only of first order accuracy.

There are many approaches to obtain a second order approximation. Here, we introduce a common way: introducing the ghost point.

Suppose we have a value at $x_{-1} = -h$ as u_{-1}. Then,

$$\frac{u_1 - u_{-1}}{2h} = \sigma_0$$
$$-\frac{u_1 - 2u_0 + u_{-1}}{h^2} = f(x_0)$$

Hence,

$$-\frac{2u_1 - 2u_0 - 2h\sigma_0}{h^2} = f(x_0).$$

This is $D_+ u + \frac{h}{2} f(x_0) = \sigma_0$.This is can be obtained by Taylor expansion as well: $u_1 \approx u_0 + hu'(x_0) + \frac{h^2}{2} u''(x_0)$.

The system of equations

$$\begin{aligned}
\frac{2u_1 - 2u_0}{h^2} &= \frac{2\sigma_0}{h} - f(x_0),\\
-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} &= f(x_j), \ \ j = 1, \dots, m\\
\frac{2u_m - 2u_{m+1}}{h^2} &= -\frac{2\sigma_1}{h} - f(x_{m+1})
\end{aligned}$$

1.2 Solvability and nonuniqueness

We have $A_1 U = F_1$. The matrix obtained now is not invertible. In fact, it has a null vector $e = [1, 1, \dots, 1]^T$. Moreover,

$$\ker(A^T) = \operatorname{span}((\frac{1}{2}, 1, 1, \dots, 1, \frac{1}{2})^T).$$

The discretized system is solvable if and only if $F_1^T(\frac{1}{2}, 1, 1, \dots, 1, \frac{1}{2})^T = 0$. This is the analogy of the condition

$$\int_0^1 f(x)\, dx = \sigma_1 - \sigma_0.$$

Why does this happen? Is this due to the numerical method? Actually, the issue is from the differential equation itself: if we integrate $\int_0^1 f dx = \sigma_1 - \sigma_0$. This means f cannot be imposed arbitrarily. Further, when it is solvable, the solution cannot be determined since there may be a constant.

To solve the numerical solution uniquely, we must impose an extra condition. One common way is $\sum_i u_i = 0$.

Remark 1. *For small scale problems, adding this constraint is fine. However, if the scale is large, this simple imposement makes the matrix nonsquare and the linear system is not so convenient to solve.*

For rectangular region, a more convenient option will be to apply the Discrete Cosine Transform or modified Conjugate Gradient method. The discrete cosine transform applies because the discrete cosine functions form a basis for the functions defined on the grid with numerical Neumann condition.

2 Variable coeffcient

Consider the problem

$$-(\kappa(x)u')' = f.$$

How do we discretize this problem?

The first way is $-\kappa u'' - \kappa' u' = f$ and

$$-\kappa_j \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} - \kappa_j \frac{u_{j+1} - u_{j-1}}{2h} = f(x_j).$$

This is not good since the coefficient matrix is not symmetric!

If we use the form given and do the finite difference directly, we may obtain:

$$-\frac{1}{h^2}\delta(\kappa\delta u)_i = f(x_i), \quad \delta u_i = \hat{D}_0 u_i = \frac{u(x+h/2) - u(x-h/2)}{h}.$$

More explicitly, one has

$$-\frac{(\kappa u')_{j+1/2} - (\kappa u')_{j-1/2}}{h} = -\frac{\kappa_{j+1/2}(u_{j+1} - u_j) - \kappa_{j-1/2}(u_j - u_{j-1})}{h^2} = f(x_j).$$

It turns out that this yields a symmetric matrix and it is also of second order accuracy!

It is usually better to discretize the conservative form directly. Such a discretization can also be obtained by the integration way as in *finite volume method*.

3 2D elliptic equations

Some typical elliptic equations in 2D include

$$-a_1 u_{xx} - a_2 u_{xy} - a_3 u_{yy} + a_4 u_x + a_5 y_y + a_6 u = f, \quad a_2^2 - 4a_1 a_3 < 0$$

and

$$-(a(x,y)u_x)_x - (b(x,y)u_y)_y + c(x,y)u = f, \quad a > 0, \quad b > 0.$$

Note that we have an extra negative sign compared with the book. This makes no difference since we can redefine f but this convention is preferred in literature since the so-defined elliptic operator is positive.

3.1 The 5-point scheme

Consider the simplest 2D poisson equation

$$\begin{aligned} -\Delta u = f,\ D = (0,a)\times(0,b), \\ u = g,\quad \partial D. \end{aligned}$$

Consider taking the step sizes

$$h = \frac{a}{I+1}, \quad k = \frac{b}{J+1}.$$

The boundary points are

$$\partial D_h = \{(x_i, y_j): \quad i = 0, I+1,\ \text{or}\ j = 0, J+1\}.$$

Applying the centered difference in both dimensions, one obtains the **5-point stencil/scheme** (五点差分格式).

$$-\Delta_h u_{ij} = -\frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} - \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{k^2} = f_{ij} = f(x_i, y_j).$$

This is called 5-point stencil since we only used five points.

The **local truncation error** can be easily obtained to be

$$\begin{aligned} T_{ij} = &-\frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} \\ &- \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2} - f(x_i, y_j) = O(h^2 + k^2). \end{aligned}$$

Hence, it is of second order accuracy.

3.2 The construction of the coefficient matrix

Consider the Dirichlet boundary condition. The linear system of equations can also be written as

$$AU = F$$

and here the matrix is of size $IJ \times IJ$.

- If one does not want to construct the matrix A for the linear system of equations, the system can be solved using some iterative methods, like the Gauss-Seidel method.
- Below, we take a look at the case when $h = k$ and how one may set up the matrix.

We must order the points. The most straightforward way is to order them as follows:

$$u = (u_{11}, u_{21}, \dots, u_{m1}, u_{12}, u_{22}, \dots, u_{1m}, \dots, u_{mm}).$$

This is convenient in Matlab since this actually corresponds to reshaping by columns of matrices. Other ordering may be possible to make the matrix even sparser. Introduce

$$\vec{u}_j = (u_{1j}, u_{2j}, \dots, u_{mj}).$$

Consider

```
e = ones(m,1);
A1 = spdiags( [e -2*e e], -1:1, m, m);
```

Then, we have

$$-\frac{1}{h^2}A_1\vec{u}_j - \frac{1}{h^2}(\vec{u}_{j+1} - 2\vec{u}_j + \vec{u}_{j-1}) = \vec{f}_j$$

$f_{1j} = f(x_1, y_j) + \frac{1}{h^2}g(0, y_j)$, $f_{mj} = f(x_m, y_j) + \frac{1}{h^2}g(1, y_j)$ and $f_{ij} = f(x_i, y_j)$ for others. From this equation, it's clear that the big matrix M has $m * m$ blocks. The (p, p) block is $-\frac{1}{h^2}(A_1 - 2I)$ and the $(p, p-1)$ and $(p, p+1)$ blocks are $-\frac{1}{h^2}I$. Note that $\vec{u}_0$ and $\vec{u}_{m+1}$ can be determined by the boundary values and can be moved to right hand side.

The big matrix can be constructed using

```
A1=spdiags(ones(m,1)*[1 -2 1], -1:1, m, m);
M=-(kron(A1, speye(m))+kron(speye(m), A1))/h^2;
```

Code presentation Consider the case $g = 0$ and $f = 2\pi^2\sin(\pi x)\sin(\pi y)$.

4 The nine-point scheme

A natural question is whether one can use more points to get higher accuracy. A natural idea is to use the nine points $(x_i + z_1 h, y_j + z_2 h)$ where $z_i = 0, \pm 1$.

For simplicity, assume

$$h = k.$$

Introduce the operators

$$\xi = h\frac{d}{dx}, \quad \eta = h\frac{d}{dy}.$$

By Taylor expansion, one may find that for example, $u(x+h, y) = e^{\xi}u(x,y)$ and $u(x+h, y-h) = e^{\xi-\eta}u(x,y)$ (noting that ξ and η commute).

By the symmetry of Laplace operator, one may consider constructing a method using the following symmetric sums:

$$\begin{aligned}
S_1 &:= u(x+h,y) + u(x,y+h) + u(x-h,y) + u(x,y-h),\\
S_2 &:= u(x+h,y+h) + u(x-h,y+h) + u(x+h,y-h) + u(x-h,y-h).
\end{aligned}$$

By the Taylor expansion,

$$\begin{aligned}
S_1 &= (e^{\xi} + e^{\eta} + e^{-\xi} + e^{-\eta})u(x,y) = (4 + \xi^2 + \eta^2 + \frac{\xi^4}{12} + \frac{\eta^4}{12})u(x,y) + O(h^6),\\
S_2 &= (e^{\xi+\eta} + e^{-\xi+\eta} + e^{\xi-\eta} + e^{-\xi-\eta})u = (4 + 2\xi^2 + 2\eta^2 + \xi^2\eta^2 + \frac{1}{6}(\xi^4 + \eta^4))u + O(h^6).
\end{aligned}$$

Note that $\xi^2 + \eta^2 = h^2\Delta$. Hence, one would like to keep $\xi^2 + \eta^2$. Noting $\xi^4 + \eta^4 = (\xi^2 + \eta^2) - 2\xi^2\eta^2$. One can find that $4S_1 + S_2$ would contain $(\xi^2 + \eta^2)$ and $(\xi^2 + \eta^2)^2$ only:

$$4S_1 + S_2 = (20 + 6(\xi^2 + \eta^2) + \frac{1}{2}(\xi^2 + \eta^2)^2)u + O(h^6).$$

This tells us that

$$\Delta u = \frac{4S_1 + S_2 - 20u}{6h^2} - \frac{1}{12}h^2\Delta^2 u + O(h^4).$$

Hence, one finds that the nine point approximation

$$\Delta_9 u_{ij} = \frac{1}{6h^2}(4u_{i+1,j}+4u_{i-1,j}+4u_{i,j+1}+4u_{i,j-1}+u_{i+1,j+1}+u_{i+1,j-1}+u_{i-1,j+1}+u_{i-1,j-1}-20u_{ij})$$

is a second order accurate approximation for the Laplacian operator. However, if we do

$$-\Delta_9 u_{ij} = f_{ij} + \frac{1}{12}\Delta_h f_{ij},$$

the accuracy is $O(h^4)$.

Exercise: Think about constructing the matrix in Matlab.

5 FDM using the polar coordinates/cylindrical coordinate

The Laplacian operator under polar coordinates (r, θ) with $r = \sqrt{x^2+y^2}$ and $\tan\theta = y/x$ is given by

$$\Delta u = \frac{1}{r}\partial_r(r\partial_r u) + \frac{1}{r^2}\partial_{\theta\theta}u.$$

In 3D, under the cylindrical coordinates (r, θ, z), the Laplacian operator is given by

$$\Delta u = \frac{1}{r}\partial_r(r\partial_r u) + \frac{1}{r^2}\partial_{\theta\theta}u + \partial_{zz}u.$$

Side notes*

There are several ways to derive these formulas:

- Using change of variables, one may change the derivatives $\partial_{xx}, \partial_{yy}$ etc into the partial derivatives on r, θ.
- We use the geometric significance of the nabla operator $df = d\vec{r}\cdot\nabla f$ and may derive the gradient vector ∇f in the corresponding coordinates. Then, $\Delta = \nabla\cdot\nabla$ and one may get the desired formula.

- Using the theory of differential forms, one may check that for a curvilinear coordinates in 3D (t^1, t^2, t^3),

  $$ds^2 = E_1(dt^1)^2 + E_2(dt^2)^2 + E_3(dt^3),$$

  one will have

  $$\nabla f = \sum_i \frac{1}{\sqrt{E_i}} \partial_{t_i} f \boldsymbol{e}_i$$

  The divergence operator can also be given. See section 14.1.5 of the book "Mathematical Analaysis" by Zorich or any textbook on Riemannian geometry.

In fact, the above three are essentially the same. For example, if one considers the parametrization $\boldsymbol{r}(t_1, t_2, t_3)$, one would have $E_i = |\partial_{t_i}\boldsymbol{r}|^2$.

Here, I will use the second way to illustrate the one for cylindrical coordinates: the position vector can be written as

$$\boldsymbol{r}(r, \theta, z) = (r\cos\theta, r\sin\theta, z)^T.$$

Here, the three basic vector is given by

$$\hat{r} = \partial_r\boldsymbol{r}/|\partial_r\boldsymbol{r}| = (\cos\theta, \sin\theta, 0)^T, \quad \hat{\theta} = \partial_\theta\boldsymbol{r}/|\partial_\theta\boldsymbol{r}| = (-\sin\theta, \cos\theta, 0)^T, \quad \hat{z} = \partial_z\boldsymbol{r}/|\partial_z\boldsymbol{r}| = (0, 0, 1)^T.$$

It is then clearly that

$$\boldsymbol{r}(r, \theta, z) = r\hat{r} + z\hat{z}.$$

Then,

$$d\boldsymbol{r} = \partial_r\boldsymbol{r}dr + \partial_\theta\boldsymbol{r}d\theta + \partial_z\boldsymbol{r}dz = \hat{r}dr + r\hat{\theta}d\theta + \hat{z}dz.$$

[In fact, this gives the metric in the curvilinear curve already: $ds^2 = d\boldsymbol{r}\cdot d\boldsymbol{r} = \cdots$].

For a function $f(r, \theta, z)$, one has

$$df = \partial_r f dr + \partial_\theta f d\theta + \partial_z f dz = d\boldsymbol{r} \cdot \nabla f.$$

If one set $\nabla f = A_1\hat{r} + A_2\hat{\theta} + A_3\hat{z}$, one then has

$$A_1 = \partial_r f, \quad A_2 r = \partial_\theta f, \quad A_3 = \partial_z f.$$

Hence,

$$\nabla = \hat{r}\partial_r + \frac{\hat{\theta}}{r}\partial_\theta + \hat{z}\partial_z.$$

It follows that

$$\Delta f = \nabla \cdot \nabla f = (\hat{r}\partial_r + \frac{\hat{\theta}}{r}\partial_\theta + \hat{z}\partial_z) \cdot (\hat{r}\partial_r f + \frac{\hat{\theta}}{r}\partial_\theta f + \hat{z}\partial_z f)$$

Here, using the orthonormal facts of the basis, one finds

$$\partial_{rr} f + \frac{\hat{\theta}}{r} \cdot (\partial_\theta \hat{r}\partial_r f + \frac{\hat{\theta}}{r}\partial_{\theta\theta} f) + \partial_{zz} f = \partial_{rr} f + \frac{1}{r}\partial_r f + \frac{1}{r^2}\partial_{\theta\theta} f + \partial_{zz} f.$$

5.1 The discretization

By the nondegeneracy, one needs $\lim_{r\to 0^+} r\partial_r u = 0$

The usual way to choose the grid is

$$r_i = (i + 1/2)\Delta r, \quad \theta_j = j\Delta\theta.$$

This choice of r_i is due to the discretization, where one wants to avoid the value at $r = 0$, where degeneracy happens. Another reason is that one wants $r_{i-1/2}$ for the variable coefficient.

In fact, using the technique for variable coefficient, one has the following discretization for $i \geq 1$:

$$\frac{1}{r_i}\frac{r_{i+1/2}u_{i+1/2,j} - (r_{i+1/2} + r_{i-1/2})u_{ij} + r_{i-1/2}u_{i-1,j}}{\Delta r^2} + \frac{1}{r_i^2}\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\Delta\theta^2} = f(r_i, \theta_j).$$

This is not complete. One must treat the one for $i = 0$ or $r_0 = \frac{1}{2}\Delta r$. This can be done by the integration method for the region $0 \leq r \leq \Delta r, \theta_{j-1/2} \leq \theta \leq \theta_{j+1/2}$. The resulted formula is

$$\frac{2}{\Delta r}\frac{u_{1,j} - u_{0,j}}{\Delta r} + \frac{4}{\Delta r^2}\frac{u_{0,j+1} - 2u_{0j} + u_{0,j-1}}{\Delta\theta^2} = f(r_0, \theta_j).$$

Remark 2. *In some sense, this equation can be understood as the one obtained by the ghost point technique: using the fact* $\lim_{r\to 0} r\partial_r u = 0$*, we may have*

$$r_{0-1/2}\frac{u_{0,j} - u_{-1,j}}{\Delta r} = 0$$

Moreover, $r_{0+1/2} = \Delta r$. *Using these, the equation can be obtained formally, though not very rigorous.*