Finite Element Method

Xiaopeng Zhang

Chapter 1

Laplace Equation

1.1 Derivation of Canonical form

Any second order linear PDE in two variables can be represented as the form

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f \tag{1.1}$$

We could derive the canonical form thus simplify the equation by canceling the cross term or muting the coefficients of two second order terms(which are equivalent by simple variable substitution).

Consider variable substitution $\xi = \xi(x, y)$, $\eta = \eta(x, y)$. We assume $\frac{D(\xi,\eta)}{D(x,y)} = \begin{vmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{vmatrix} \neq 0$ in the neighborhood of some (x_0, y_0), thus the variable substitution is invertible according the implicit function theorem.

We have

$$au_{\xi\xi} + 2bu_{\xi\eta} + cu_{\eta\eta} + du_\xi + eu_\eta + fu = g \tag{1.2}$$

with the following relations.

$$\begin{cases} u_x = u_\xi \xi_x + u_\eta \eta_x \\ u_y = u_\xi \xi_y + u_\eta \eta_y \\ u_{xx} = u_{\xi\xi}\xi_x^2 + 2u_{\xi\eta}\xi_x\eta_x + u_{\eta\eta}\eta_x^2 + u_\xi \xi_{xx} + u_\eta \eta_{xx} \\ u_{xy} = u_{\xi\xi}\xi_x\xi_y + u_{\xi\eta}(\xi_x\eta_y + \xi_y\eta_x) + u_{\eta\eta}\eta_x\eta_y + u_\xi \xi_{xy} + u_\eta \eta xy \\ u_{yy} = u_{\xi\xi}\xi_y^2 + 2u_{\xi\eta}\xi_y\eta_y + u_{\eta\eta}\eta_y^2 + u_\xi \xi_{yy} + u_\eta \eta_{yy} \end{cases}$$

One might think of writing for example u_ξ as some other v_ξ since when substituting variables, the function $u(x, y)$ itself has changed to another $v(\xi, \eta)$, but in the field of PDE, we simply write $v(\xi, \eta)$ still as $u(\xi, \eta)$ to imply that they are nothing different but variable substitution.

The coefficients a, b and c can be determined from the relations as

$$\begin{cases} a = A\xi_x^2 + 2B\xi_x\xi_y + C\xi_y^2 \\ b = A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + C\xi_y\eta_x \\ c = A\eta_x^2 + 2B\eta_x\eta_y + C\eta_y^2 \end{cases}$$

Note that representations of a and c are quite similar, if we could eliminate them from the equation, the simplification is done followed by one more simple variable substitution.

Consider first-order linear PDE

$$A\varphi_x^2 + 2B\varphi_x\varphi_y + C\varphi_y^2 = 0 \tag{1.3}$$

If there exist two independent solution $\varphi = \varphi_1(x, y)$, $\varphi = \varphi_2(x, y)$, we can eliminate a and c by applying $\begin{cases} \xi = \varphi_1(x, y) \\ \eta = \varphi_2(x, y) \end{cases}$

A common way to solve first-order linear PDE is characteristic method. Our aim is to derive $\varphi(x, y)$, but instead of directly solving the equation, we consider a special curve on the xOy plane

$$\Gamma : \varphi(x, y) = 0 \tag{1.4}$$

If we combine (1.3) and (1.4) as a system, we could find the relation between x and y by the implicit function theorem thus $\varphi(x, y)$ can be found. Since $\frac{dy}{dx} = -\frac{\varphi_x}{\varphi_y}$, plug it into (1.4), we have

$$A(\frac{dy}{dx})^2 - 2B\frac{dy}{dx} + C = 0 \tag{1.5}$$

There exist the following three circumstances.

(1). $\Delta = B^2 - AC > 0$, then two curves satisfy the aforementioned equation system, $(B + \Delta)x - Ay = 0$ and $(B - \Delta)x - Ay = 0$, which are the two independent solutions of (1.4). Therefore, the variable substitution are $\xi = (B + \Delta)x - Ay$ and $\eta = (B - \Delta)x - Ay$

Now equation (1.2) is changed to be

$$u_{\xi\eta} = A_1u_\xi + B_1u_\eta + C_1u + D \tag{1.6}$$

With another variable substitution

$$\xi = \tfrac{1}{2}(s + t),\ \eta = \tfrac{1}{2}(s - t)$$

(1.6) becomes $u_{ss} - u_{tt} = A_2 u_s + B_2 u_t + C_1 u + D_1$

PDEs in this form are called hyperbolic PDEs. This is because the quadratic form $Q(x, y) = Ax^2 + 2Bxy + Cy^2 = 1$ is a hyperbola when $B^2 - AC > 0$

(2). $\Delta = B^2 - AC < 0$, then (1.5) has only complex solutions thus we cannot find two independent real solutions of φ.(the variable substitution of $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ urges the solutions to be real)

We should find yet another method. Let's assume that we have found the variable substitution $\xi = \xi(x, y)$ and $\eta = \eta(x, y)$ such that for example $a = c$ and $b = 0$.

Consider function $\phi = \xi + i\eta$. Since $\phi_x = \xi_x + i\eta_x$, $\phi_y = \xi_y + i\eta_y$, we have

$$\begin{aligned} A\phi_x^2 + 2B\phi_x\phi_y + C\phi_y^2 &= A(\xi_x^2 - \eta_x^2) + 2B(\xi_x\xi_y - \eta_x\eta_y) + C(\xi_y^2 - \eta_y^2) \\ &\quad + i(2A\xi_x\eta_x + 2B(\xi_x\eta_y + \xi_y\eta_x) + 2C\xi_y\eta_y \\ &= a - c + i \cdot 2b \\ &= 0 \end{aligned}$$

This means whenever we find a complex solution $\phi(x, y)$ that satisfy (1.3) , we can recover ξ and η with the transformation $\xi = Re\phi$ and $\eta = Im\phi$. Since ξ and η satisfy $a - c = 0$ and $b = 0$, so with a change of variables from x and y to ξ and η will transform the PDE (1.1) into canonical form

$$u_{\xi\xi} + u_{\eta\eta} + (\text{lower-order terms}) \tag{1.7}$$

PDEs as the form of (1.7) are called elliptic PDEs. Consider quadratic form $Q(x, y) = Ax^2 + 2Bxy + Cy^2 = 1$, with simple calculation we can find that it is centered around the origin and any line passing through the origin will have two intersections with the curve of the quadratic form. In fact it is a ellipse so we call this type of PDEs as elliptic PDEs.

(3). $\Delta = B^2 - AC = 0$, we can write (1.3) as $(\sqrt{A}\xi_x + \sqrt{C}\xi_y)^2 = 0$. Therefore we have only one bundle of characteristics, name it as $\varphi_1(x, y) = c_1$. We set $\xi = \varphi_1(x, y)$ so $a = 0$. Furthermore, since $\Delta = 0$, we have

$$\begin{aligned} b &= A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + C\xi_y\eta_y \\ &= (\sqrt{A}\xi_x + \sqrt{C}\xi_y)(\sqrt{A}\eta_x + \sqrt{C}\eta_y) = 0 \end{aligned}$$

Since both a and c are 0, we choose any $\eta = \varphi_2(x, y)$ that independent from φ_1, then with simple calculation, (1.1) can be transformed as canonical form

$$u_{\eta\eta} = A_2 u_\xi + B_2 u_\eta + C_2 u + D_2 \tag{1.8}$$

PDEs as the form of (1.8) are called parabolic PDEs. The name parabola is used because the assumption on the coefficients are the same as the condition for analytic geometry equation $Ax^2 + 2Bxy + Cy^2 + Dx + Ey + F = 0$ to define a planar parabola.

1.2 Discretization of the Laplace Equation

The Laplace equation is a elliptic PDE. In physical applications, it could be used to illustrate the equilibrium temperature or concentration of some chemical. In two dimensions, it is represented as

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \tag{1.9}$$

The function f is called the source term or inhomogeneous term or forcing function. (1.1) is supposed to be satisfied for every point in the interior of domain Ω. Here for simplicity, we set Ω to be $\{(x, y)|0 < x < s, 0 < y < t\}$

In terms of boundary conditions, we treat only the Dirichlet boundary, which is given by

$$u(x, y) = g(x, y) \quad \text{on } \Gamma \text{ or } \partial\Omega \tag{1.10}$$

Discretization means we think of the equation holding at finite discrete points on domain Ω and only solve finite number of equations, unlike the original problem, where the equation holds on infinite number of interior points and the number of values to be solved are also infinite.

Finite difference refers to approximation to derivatives, such as

$$v'(x) \approx \frac{v(x + \Delta x) - v(x)}{\Delta x}$$

$$v''(x) \approx \frac{v(x + \Delta x) - 2v(x) + v(x - \Delta x)}{\Delta x^2}$$

Applying such operation to the Laplace equation yields

$$\begin{aligned} f = \Delta u(x, y) \approx & \frac{u(x + \Delta x, y) - 2u(x, y) + u(x - \Delta x, y)}{\Delta x^2} \\ & + \frac{u(x, y + \Delta y) - 2u(x, y) + u(x, y - \Delta y)}{\Delta y^2} \end{aligned} \tag{1.11}$$

We create a grid of points in the interior with uniform spacing Δx and Δy

$$0 = x_0 < x_1 < x_2 < ... < x_n < x_{n+1} = s, \quad x_{i+1} - x_i = \Delta x, \quad \forall 0 < i < n$$

$$0 = y_0 < y_1 < y_2 < ... < y_n < y_{n+1} = t, \quad y_{j+1} - y_j = \Delta y, \quad \forall 0 < j < n$$

There is a grid of n^2 points where the finite difference approximation can be applied. We write $U_{jk} \approx u(x_j, y_k)$ as the to be solved approximate values of underlying values $u(x_j, y_k)$. Similar to (1.11) where u should satisfy (1.9) at grid points but using the finite different approximation, we give the following equation as approximation.

$$\frac{U_{j+1,k} - 2U_{jk} + U_{j-1,k}}{\Delta x^2} + \frac{U_{j,k+1} - 2U_{j,k} + U_{j,k-1}}{\Delta y^2} = f_{j,k} \tag{1.12}$$

If we write the above equation as the form

$$(\frac{2}{\Delta x^2} + \frac{2}{\Delta y^2})U_{j,k} - (\frac{U_{j+1,k} + U_{j-1,k}}{\Delta x^2} + \frac{U_{j,k+1} + U_{j,k-1}}{\Delta y^2}) = -f_{j,k} \tag{1.13}$$

This implies the Laplace equation is roughly about the average since if $f = 0$, $U_{j,k}$ depicts the average.

The boundary values are given their exact known values.

$$\begin{aligned} U_{0k} &= u(0, y_k) = g(0, y_k) \\ U_{n+1,k} &= u(x_{n+1}, y_k) = g(1, y_k) \\ U_{j,0} &= u(0, y_k) = g(0, y_k) \\ U_{j,n+1} &= u(x_{n+1}, y_k) = g(1, y_k) \end{aligned}$$

There are n^2 equations (1.12), one for each interior grid point. There are also n^2 unknowns $U_{j,k}$. Thus the finite difference equations (1.12) are a system of linear equations. If the corresponding matrix is non-singular, then there is a unique solution U, which is the finite difference approximation to u.

Without loss of generality, we will consider $\Delta x = \Delta y$ in the following sections.

Theorem 1.1. *The coefficient matrix A derived from* (1.12) *is s.p.d.*

Proof. Consider a quadratic form

$$F(U) = \frac{1}{2}\sum_{j=0}^{n}\sum_{k=0}^{n}[(U_{j+1,k} - U_{jk})^2 + (U_{j,k+1} - U_{jk})^2]$$

Note that

$$\frac{\partial F(U)}{\partial U_{lm}} = 4U_{lm} - (U_{l-1,m} + U_{l+1,m} + U_{l,m-1} + U_{l,m+1}) = (AU)_{lm}$$

Therefore, we have $\frac{\partial F(U)}{\partial U} = AU$, which also means $F(U) = U^T AU$.

Since $F(U) \geq 0$, we know that A is a s.p.d matrix. □

1.3 Variational principle

A variational principle is a minimization problem so that u or U is the minimizer.

1.3.1 For Laplace Equation

The Dirichlet variational principle for the inhomogeneous Laplace equation is

$$\min_{u=g \text{ on } \Gamma} \frac{1}{2}\int_\Omega |\nabla u(x,y)|^2 dxdy + \int_\Omega f(x,y)u(x,y)dxdy \tag{1.14}$$

This is a weak form of Laplace equation which means that a solution of Laplace equation is also a solution of (1.14) but inverse only holds in most cases.

Dirichlet gave a simple argument and a rigorous argument was given by Hermann Weyl after almost a century.

Define a functional as

$$F(w) = \frac{1}{2}\int_\Omega |\nabla w(x,y)|^2 dxdy + \int_\Omega f(x,y)w(x,y)dxdy$$

Suppose $F(w)$ reaches its minimum at $w = u$, adding some perturbation$u \leftarrow u + \epsilon v$ onto u will lead to lower the energy as long as $v = 0$ on Γ. That is $F(u + \epsilon v) > F(u)$

$$\begin{aligned} F(u+\epsilon v) =& F(u) + \epsilon \cdot [\int_\Omega \nabla u(x,y)\cdot\nabla v(x,y)dxdy + \int_\Omega f(x,y)v(x,y)dxdy] \\ &+ \epsilon^2 \cdot \frac{1}{2}\int_\Omega |\nabla v(x,y)|^2 dxdy \end{aligned}$$

If we define $G(\epsilon) = F(u + \epsilon v)$, note that $G(\epsilon)$ reaches its minimum at $\epsilon = 0$ thus we have $G^{'}(0) = 0$, then

$$\int_{\Omega} \nabla u(x,y) \cdot \nabla v(x,y) dxdy + \int_{\Omega} f(x,y)v(x,y)dxdy = 0$$

By Gauss-Green theorem, if $v = 0$ on Γ, then

$$\int_{\Omega} \nabla u(x,y) \cdot \nabla v(x,y) dxdy = -\int_{\Omega} \Delta u(x,y) \cdot v(x,y)dxdy$$

This yields another equivalent weak form of the Laplace equation, for every $v(x,y)$ that satisfies the Dirichlet boundary condition, we have

$$\int_{\Omega} [\Delta u(x,y) - f(x,y)]v(x,y)dxdy = 0 \tag{1.15}$$

The relationship of the Laplace equation and 2 weak forms are:

$$(D) \rightarrow (1.14) \leftrightarrow (1.15)$$

1.3.2 For discrete Laplace equation

A natural discrete approximation of the integral fu is

$$\int_{\Omega} f(x,y)u(x,y)dxdy \approx \Delta x^2 \sum_{j=1}^{n}\sum_{k=1}^{n} f_{jk}U_{jk} \tag{1.16}$$

Note that j and k goes from 1 to n instead of from 0 to $n+1$ because of the zero boundary condition.

The discrete approximation of the Dirichlet integral is

$$\frac{1}{2}\int_{\Omega} |\nabla u(x,y)|^2 dxdy \approx \frac{1}{2}\sum_{j=0}^{n}\sum_{k=0}^{n}[(U_{j+1,k} - U_{jk})^2 + (U_{j,k+1} - U_{jk})^2] \tag{1.17}$$

Theorem 1.2 (Discrete variational principle)**.** *Discrete Laplace equation $AU = F$ holds if and only if*

$$U = \min_{U} \frac{1}{2}\sum_{j=0}^{n}\sum_{k=0}^{n}[(U_{j+1,k} - U_{jk})^2 + (U_{j,k+1} - U_{jk})^2] - \Delta x^2 \sum_{j=1}^{n}\sum_{k=1}^{n} f_{jk}U_{jk}$$

Proof. Define $F(U)$ as the finite sum to be minimized. Consider the overall derivative

$$\frac{\partial F(U)}{\partial U_{lm}} = 4U_{lm} - (U_{l-1,m} + U_{l+1,m} + U_{l,m-1} + U_{l,m+1}) - \Delta x^2 \cdot f_{jk} = 0$$

This is exactly the discrete version of Laplace equation. □

Remark1. In DVP, the latter term is with a minus sign while in variational principle of Laplace equation, the integral of fv is with a positive sign.

Remark2. If we change j and k in the discrete approximation of Dirichlet integral to start from 1, the derivative at the boundary will not coincide with $AU = F$ then. Do be careful with the index during coding.

1.4 Stability

Poincare inequality is basically saying that you cannot get a large function without a large gradient.

Theorem 1.3 (Poincare inequality)**.** *Let Ω be a bounded set and consider L^p norm. There exists a constant $C_{\Omega,p}$, depending only on Ω and p. For any function u of the Sobolev space $W_0^{1,p}(\Omega)$,*

$$\|u\|_{L^p(\Omega)} \leq C_{\Omega,p} \|\nabla u\|_{L^p(\Omega)}$$

Remark.

$$\|u\|_{W_0^{1,p}(\Omega)} = (\sum_{|\alpha| \leq k} \int_\Omega |D^\alpha u(x)|^p dx)^{1/p}$$

with zero condition on the boundary.

$$\int_\Omega u(x,y)^2 dxdy \leq C \int_\Omega |\nabla u(x,y)|^2 dxdy$$

Theorem 1.4 (Discrete Poincare inequality)**.**

$$\Delta x^2 \sum_{j,k} U_{jk}^2 \leqslant C_\Omega \sum_{j,k} [(U_{j+1,k} - U_{j,k})^2 + (U_{j,k+1} - U_{j,k})^2]$$

Proof. omit □

1.5 Gauss Seidel Method

Gauss Seidel method computes a sequence of iterates, $U^{(1)}, U^{(2)}, ..., U^{(k)}, ...$ It's derived from fixed point iteration $u_{n+1} = (I - Q^{-1}A)u_n + Q^{-1}F$, where Q is selected to be $Q = D + L$ so that the iteration becomes $(D + L)u_{n+1} = A_u \cdot u_n + F$. A_n is the upper triangular part of A. By component-wise,

$$\sum_{i<j} a_{ji} U_i^{(k+1)} + a_{jj} U_j^{(k+1)} = b_j - \sum_{i>j} a_{ji} U_i^{(k)}$$

To compute jth component of $U^{(k+1)}$, we have

$$U_j^{(k+1)} = \frac{1}{a_{jj}} (F_j - \sum_{i<j} a_{ji} U_i^{(k+1)} - \sum_{i>j} a_{ji} U_i^{(k)}) \tag{1.18}$$

The first sum involves the components of $U^{(k+1)}$ that have already been computed. The second sum involves components of $U^{(k)}$ that have not yet been updates

Chapter 2

Finite Element Methods

Finite element methods are commonly used to solve elliptic PDEs for their geometric flexibility and ability to handle PDEs. The differences between FEM and FDM are :

- The most attractive feature of FEM is its ability to handle complicated geometries(boundaries). FDM is restricted to handle rectangular shapes and simple alterations.
- The most attractive feature of FDM is that it is straightforward to implement.
- One could consider FDM a particular case of FEM approach.

We start with introducing some common definitions in FEM.

A *triangle* consists of the *interior*, the *edges* and the *vertices*. An edge connects each pair of vertices. The edge does not contain the vertices themselves, which we indicate by writing (a, b) instead of $[a, b]$. This convention has the consequence that if $d \in e$ is any point on the edge, then d is not a vertex of the triangle. The interior is all the triangle except the vertices and the edges, in other words, the face of the triangle.

A *triangulation* of Ω, denoted by τ, is a set of triangles so that the interiors are disjoint and every point is in the closure of some triangle. A valid triangulation is one without slave nodes, which is defined as a vertex of one triangle that is on an edge of another triangle.

A *piecewise linear* C_0 function is a continuous function $\mathcal{R}^2 \to \mathcal{R}$ defined on Ω, that is affine(linear) when restricted to any $T \in \tau$. Here by affine we mean there are constants so that $u(x, y) = \alpha x + \beta y + \gamma$. And by piecewise affine we mean that for every $T \in \tau$, there exist α_T, β_T, γ_T such that

$u(x, y) = \alpha_T x + \beta_T y + \gamma_T, \forall (x, y) \in T$. The PLT space, denoted by $\mathcal{S}_\tau$, consists of all piecewise linear functions on τ that satisfy the Dirichlet boundary conditions.