Lecture Notes on
Computational and Applied Mathematics [*]

May 2, 2011

Prepared by
I-Liang CHERN [1]
and
Jun ZOU [2]

Mainly based on the textbook
“ Introduction to Applied Mathematics ”
by Professor Gilbert Strang
Wellesley-Cambridge Press (1986)
(**Chapters 1, 2, 3, 4 and 5**)
Note: University library call number QA37.2.S88

[*] We have prepared the lecture notes purely for the convenience of my teaching. Students taking this course may use the notes as part of their reading and reference materials. Students are advised to find more details about the background, motivation and exercises in the textbooks suggested.

[1] Department of Mathematics, National Taiwan University, Taipei, Taiwan

[2] Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Contents

1 Mathematical modeling — discrete versus continuous models

Physical systems are usually modeled by disrete systems or continuous systems. For instance, one kind of discrete system consist of nodes connected by bonds (or edges) representing interactions among nodes. A spring-mass system is one of such discrete systems. The masses are the nodes, whereas the springs are the bonds. In the solid mechanics, we may also image the atoms are the nodes and the chemical bonds connecting them are the bonds. Continuous systems can be viewed as continuous limits of discrete systems. Solid mechanics, fluid mechanics are such systems.

If the systems do not vary in time, we call them in equilibria. In this situation, discrete systems are usually modeled by algebraic equations, whereas continuous systems are usually formulated by differential equations.

Below, I shall use the spring-mass system and elastic bar to explain how to do mathematical modeling for discrete and continuous systems.

1.1 Modeling spring-mass systems

Consider a spring-mass system which consists of n masses placed vertically between two walls. The n masses and the two end walls are connected by $n+1$ springs. If all masses are zeros, the springs are “at rest” states. When the masses are greater than zeros, the springs are elongated due to the gravitation force. The mass m_i moves down u_i distance, called the displacement. The goal is to find the discplacements u_i of the masses m_i, $i = 1, ..., n$.

In this model, the nodes are the masses m_i. We may treat the end walls are the fixed masses, and call them m_0 and m_{n+1}, respectively. The edges (or the bonds) are the springs. Let us call the spring connecting m_i and m_{i+1} by edge (or spring) i, $i = 1, ..., n+1$. Suppose the spring i has spring constant c_i. Let us call the downward direction the positive direction.

Let me start from the simplest case: $n = 1$ and no bottom wall. The mass m_1 elongates the spring 1 by a displacement u_1. The elongated spring has a *restoration force* $-c_1 u_1$ acting on m_1.[3] This force must be balanced with the gravitational force on

[3] The minus sign is due to the direction of force is upward.

m_1.[4] Thus, we have

$$-c_1u_1 + f_1 = 0,$$

where $f_1 = m_1g$, the gravitation force on m_1, and g is the gravitation constant. From this, we get

$$u_1 = \frac{f_1}{c_1}.$$

Next, let us consider the case where there is a bottom wall. In this case, both springs 1 and 2 exert forces upward to m_1. The balance law becomes

$$-c_1u_1 - c_2u_1 + f_1 = 0.$$

This results $u_1 = f_1/(c_1 + c_2)$.

Let us jump to a slightly more complicated case, say $n = 3$. The displacements

$$u_0 = 0,\ u_4 = 0, \tag{1.1}$$

due to the walls are fixed. The displacements u_1, u_2, u_3 cause elongations of the springs:

$$e_i = u_i - u_{i-1}, i = 1, 2, 3, 4. \tag{1.2}$$

The restoration force of spring i is

$$w_i = c_ie_i. \tag{1.3}$$

The force exerted to m_i by spring i is $-w_i = -c_ie_i$. In fact, when $e_i < 0$, the spring is shortened and it pushes downward to mass m_i (the sign is positive), hence the force is $-c_ie_i > 0$. On the other hand, when $e_i > 0$, the spring is elongated and it pull m_i upward. We still get the force $-w_i = -c_ie_i < 0$. Similarly, the force exerted to m_i by spring $i+1$ is $w_{i+1} = c_{i+1}e_{i+1}$. When $e_{i+1} > 0$, the spring $i+1$ is elongated and it pulls m_i downward, the force is $w_{i+1} = c_{i+1}e_{i+1} > 0$. When $e_{i+1} < 0$, it pushes m_i upward, and the force $w_{i+1} = c_{i+1}e_{i+1} < 0$. In both cases, the force extertеd to m_i by spring $i+1$ is w_{i+1}.

Thus, the force balance law on m_i is

$$w_{i+1} - w_i + f_i = 0, i = 1, 2, 3. \tag{1.4}$$

There are three algebraic equations for three unknowns u_1, u_2, u_3. In principle, we can solve it.

[4] The mass m_1 is in equilibrium.

Let us express the above equations in matrix form. First, the elongation:

$$e = Au, \text{ or } \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix} = \begin{pmatrix} 1 & & \\ -1 & 1 & \\ & -1 & 1 \\ & & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

the restoration force:

$$w = Ce, \text{ or } \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} c_1 & & & \\ & c_2 & & \\ & & c_3 & \\ & & & c_4 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

the force balance laws:

$$A^t w = f, \text{ or } \begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & 1 & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

where A^t is the transpose of A.

We can write the above equations in block matrix form as

$$\begin{pmatrix} C^{-1} & A \\ A^t & 0 \end{pmatrix} \begin{pmatrix} -w \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ -f \end{pmatrix}. \tag{1.5}$$

This kind of block matrix appears commonly in many other physical systems, for instance, network flows, fluid flows. In fact, any optimization system with constraint can be written in this form. Here, the constraint part is the second equation. We shall come back to this point in the next section.

One way to solve the above block matrix system is to eliminate the variable w and get

$$Ku := A^t CAu = f. \tag{1.6}$$

The matrix $K := A^t CA$ is a symmetric positive definite matrix. It is called the *stiffness matrix*. For $n = 4$, we get

$$K := A^t CA = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 + c_4 \end{pmatrix}$$

Below, we use the Gaussian elimination method to solve this system.

1.2 Solving the spring-mass system — Gaussian elimination

In this subsection, we shall solve $Ax = b$ by the Gaussian elimination method This method is equivalent to factor A into LU, a product of a lower triangular matrix and an upper triangular matrix, where L is normalied with unity diagonal entries. The problem of solving $Ax = b$ is decomposed into two steps:

$$Lc = b, \; Ux = c.$$

The former can be solved by forward substitution, whereas the latter can be solved by backward substitution. We give detail discription below. In the case when K is sysmetric, we can factor A into $A = LDL^t$, where D is a diagonal matrix. The equation $Ax = b$ can be splitted into

$$Ld = c, \; Dd = c, \; L^t x = d.$$

They can be solved by forward substitution, direct inversion and backward substitution, respectively.

1.2.1 Gaussian elimination

Consider solving the system of equations of the following form

$$A\,x = b\,,$$

where A is a non-singular $n \times n$ matrix and $b \in \mathbb{R}^n$ is a vector.

First, let us review some basic properties of lower triangular matrices.

Properties of lower triangular matrices.

- **Understanding row operation by a lower triangular matrix:**

$$\begin{aligned} L_2 A &\equiv \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & l_{32} & 1 & 0 \\ 0 & l_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} + l_{32}a_{21} & a_{32} + l_{32}a_{22} & a_{33} + l_{32}a_{23} & a_{34} + l_{32}a_{24} \\ a_{41} + l_{42}a_{21} & a_{42} + l_{42}a_{22} & a_{43} + l_{42}a_{23} & a_{44} + l_{42}a_{24} \end{bmatrix}. \end{aligned}$$

If you look at the above operations carefully, you will come to the conclusion:

the actions of adding row 2 multiplied by l_{32} and l_{42} respectively to row 3 and row 4 respectively, are equal to the operation L_2A.

- **The product of two elementary lower triangular matrices** Now consider the following two matrices:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & 0 & 1 & 0 \\ l_{41} & 0 & 0 & 1 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & l_{32} & 1 & 0 \\ 0 & l_{42} & 0 & 1 \end{bmatrix},$$

we can directly check

$$L_1L_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & 0 & 1 \end{bmatrix}.$$

Please **check** the product L_2L_1 !

You may also define

$$L_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & l_{43} & 1 \end{bmatrix},$$

Find the products, $L2L1$ and $L_1L_2L_3$.

- **A lower triangular system can be solved by forward substitution.** Consider the linear equation

$$\begin{aligned} c_1 &= b_1 \\ l_{21}c_1 + c_2 &= b_2, \\ l_{31}c_1 + l_{32}c_2 + c_3 &= b_3, \\ l_{41}c_1 + l_{42}c_2 + l_{43}c_3 + c_4 &= b_4 \end{aligned}$$

You can solve c_1 first, then substitute into the second equation to get c_2 and so on. This is called *method of forward substitution*. You can check that

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -l_{21} & 1 & 0 & 0 \\ -l_{31} & 0 & 1 & 0 \\ -l_{41} & 0 & 0 & 1 \end{bmatrix} .$$

Similar results are true for other L_i matrices and any $n \times n$ matrix A.

Gaussian elimination and LU factorization Gaussian elimination is a process to reduce a full $n \times n$ system of equations

$$\begin{aligned} & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ & \cdots\cdots \\ & a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{aligned}$$

into a upper diagonal system of equations

$$\begin{aligned} & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ & 0 \quad x_1 + \tilde{a}_{22}x_2 + \cdots + \tilde{a}_{2n}x_n = \tilde{b}_2 \\ & \cdots\cdots \\ & 0 \quad x_1 + 0 \quad x_2 + \cdots + \tilde{a}_{nn}x_n = \tilde{b}_n \, . \end{aligned}$$

This is equivalent to a process of reducing the full matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

into a upper triangular matrix

$$\tilde{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & \tilde{a}_{22} & \cdots & \tilde{a}_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \tilde{a}_{nn} \end{bmatrix} .$$

Next, we will explain the Gaussian elimination and the LU factorization for two simple examples: one is a 2×2 system of equations, the other is a 3×3 system. If you can understand these two simple examples, then you may easily carry out the Gaussian elimination and LU factorization for more general $n \times n$ systems.

- **Gaussian elimination for a 2×2 matrix**

  Consider the following 2×2 system:

$$Ax \equiv \begin{bmatrix} 2 & 4 \\ 4 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \equiv b\,. \tag{1.7}$$

  Eliminating x_1 in the 2nd equation needs to add row 1 multiplied by -2 to row 2, that equals

$$\tilde{L}A \equiv \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 4 & 11 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 0 & 3 \end{bmatrix} \equiv U.$$

  Using the property of the matrix $\tilde{L}$, the above equation gives

$$A = \begin{bmatrix} 2 & 4 \\ 4 & 11 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 0 & 3 \end{bmatrix} \equiv L\,U.$$

  This implies a LU factorization of A.

  Using the above factorization, solving the system

$$Ax = b$$

  is equivalent to solving the system

$$LUx = b,$$

  which can be done as follows:

$$L\,c = b, \quad U\,x = c.$$

  Applying this process to equation (1.7), we have

$$L\,c = b \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

  which gives

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}.$$

Then

$$U\,x = c \quad \Longleftrightarrow \quad \begin{bmatrix} 2 & 4 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix},$$

which gives the solution of equation (1.7):

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

- **Gaussian elimination for a 3×3 matrix**

  Let us consider one more simple example for the Gaussian elimination:

$$Ax \equiv \begin{bmatrix} 1 & 1 & 1 \\ 3 & 6 & 4 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1/3 \end{bmatrix} \equiv b\,. \tag{1.8}$$

  Eliminating x_1 in the 2nd equation needs to add row 1 multiplied by -3 to row 2, and eliminating x_1 in the 3rd equation needs to add row 1 multiplied by -1 to row 3, that equals

$$\tilde{L}_1 A \equiv \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 3 & 6 & 4 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 1 & 0 \end{bmatrix} \equiv U_1.$$

  Now eliminating x_2 in the 3rd equation needs to add row 2 multiplied by $-1/3$ to row 3, that equals

$$\tilde{L}_2 \tilde{L}_1 A \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 3 & 6 & 4 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & -1/3 \end{bmatrix} \equiv U.$$

  Using the property of the matrix $\tilde{L}_1$ and $\tilde{L}_2$, the above equation gives

$$\begin{aligned} A &= \begin{bmatrix} 1 & 1 & 1 \\ 3 & 6 & 4 \\ 1 & 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & -1/3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & -1/3 \end{bmatrix} \\ &\equiv L\,U. \end{aligned}$$

This completes a LU factorization of A.

Using this factorization, solving the system

$$Ax = b$$

is equivalent to solving the system

$$LUx = b,$$

which can be done as follows:

$$L\,c = b, \quad U\,x = c.$$

Applying this process to equation (1.8), we have

$$L\,c = b \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1/3 \end{bmatrix},$$

which gives

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}.$$

Then

$$U\,x = c \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & -1/3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix},$$

which gives the solution of equation (1.8):

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -8/3 \\ -1/3 \\ 3 \end{bmatrix}.$$

1.2.2 *LU* factorization for a $n \times n$ matrix

The previous LU factorization can be carried out for more general $n \times n$ matrices. In general, if all the main submatrices of A is non-singular, then we have

$$A = L\,U$$

where L is a lower triangular matrix with 1 as its diagonal entries, and U is a upper triangular matrix. Let

$$D = \mathrm{diag}(U),$$

then we can further factorize A as follows:

$$A = L\,D\,U$$

where L and U are lower and upper triangular matrices respectively, both matrices with 1 as their diagonal entries, and D is a diagonal matrix.

Remark 1.1. *Please refer to the Tutorial Notes and Assignments for more details about how to find the factorization of the form $A = L\,D\,U$.*

1.2.3 *LU* factorization of a symmetric positive definite matrix

Let A be a symmetric and positive definite matrix. A has a unique decomposition:

$$A = L\,D\,U\,.$$

Since A is symmetric, so

$$A = L\,D\,U = U^T\,D\,L^T.$$

By the uniqueness, we have $U = L^T$, that is,

$$A = L\,D\,L^T.$$

If we write

$$D^{1/2} = \mathrm{diag}(\sqrt{d_{ii}}),$$

then

$$A = (L\,D^{1/2})\,(L\,D^{1/2})^T.$$

This indicates that for any symmetric and positive definite matrix A, we have the factorization of the form

$$A = \tilde{L}\,\tilde{L}^T$$

where $\tilde{L}$ is a lower triangular matrix. This is called the *Cholesky factorization* of a positive definite matrix.

Note that the diagonal entries of L in the Cholesky factorization is not necessary to be 1, **not like** the entries of L in the LU factorization.

Remark 1.2. *Please check the Tutorial Notes for more details about how to find the factorization of the form $A = L\,L^T$.*

1.2.4 Solving the spring-mass system

Now, we go back to the spring-mass system:

$$Ku = A^tCAu = f \tag{1.9}$$

Due to the symmetry of our matrix K, we can indeed factor K into

$$K = LDL^t$$

where L is a lower triangular matrix

$$L = \begin{pmatrix} 1 & & \\ l_{21} & 1 & \\ 0 & l_{32} & 1 \end{pmatrix}$$

and $D =$ diag$(\lambda_1, \lambda_2, \lambda_3)$ is a diagonal matrix. Solving $Ku = f$ is now decomposed into three steps:

$$Ly = f, \ Dw = y, \ L^tu = w.$$

Each of them can be solved by substitution.

To find the matrices L and D, we multiply LDL^t and get the entry terms as

$$LDL^t = \begin{pmatrix} \lambda_1 & \lambda_1 l_{21} & 0 \\ \lambda_1 l_{21} & \lambda_1 l_{21}^2 + \lambda_2 & \lambda_2 l_{32} \\ 0 & \lambda_2 l_{32} & \lambda_2 l_{32}^2 + \lambda_3 \end{pmatrix}$$

By comparing the entries of A and LDL^t, we get

$$\begin{aligned} \lambda_1 &= c_1 + c_2 \\ \lambda_1 l_{21} &= -c_2 \\ \lambda_1 l_{21}^2 + \lambda_2 &= c_2 + c_3 \\ \lambda_2 l_{32} &= -c_3 \\ \lambda_2 l_{32}^2 + \lambda_3 &= c_3 + c_4 \end{aligned}$$

We can solve $\lambda_1, l_{21}, \lambda_2, l_{32}, \lambda_3$ successively from the above equations.

Remarks. The above material is mainly from pp. 40-44.

Homeworks

1. Find the matrix A^tCA for general case n.
2. pp. 46, 1.4.11,
3. pp. 46, 1.4.12.

1.3 Modeling an elastic bar

1.3.1 Strain and Stress

Consider a continuous elastic bar[5] of length 1, which is hanged vertically. (it is displaced up and down due to gravity). Set up an x-axis along the bar, so that its positive direction pointing downwards and its origin is located at the top of the elastic bar. Consider any point at x along the bar (the position is at x if no external force present), it is displaced down to $x + u(x)$ because of the action of the external force of gravity[6]. Function $u(x)$ is called the displacement. The stretching at any point is measured by the derivative $e = du/dx$, called the *strain*. If u is a constant, the elastic bar is unstretched. Otherwise the stretching of the bar produces an internal force called stress (one can experience this force easily by pulling the two ends of an elastic bar). By experiments, people find this internal force is proportional to the strain in the bar, i.e.

$$\text{(internal force)} \quad w(x) = c(x)\,\frac{du}{dx}\ ,$$

where $c(x)$ is a constant determined by the elastic material, or a function if the material is inhomogeneous.

To set up the model, we take a small piece of the bar $[x, x + \triangle x]$, its equilibrium requires all forces acted on it to be balanced. We have

$$\left(ac(x)\frac{du}{dx}\right)_{x+\triangle x} - \left(ac(x)\frac{du}{dx}\right)_x + (\rho \triangle x a)g = 0, \tag{1.10}$$

where g is the gravitational constant, a the cross-sectional area, and $\rho(x)$ the density at position x.

Dividing both sides of equation (1.10) by $a\triangle x$, then taking $\Delta x \to 0$, we get

$$-\frac{d}{dx}(c(x)\,\frac{du}{dx}) = f(x) \tag{1.11}$$

[5] You may pull back and forth an elastic bar and its length is much bigger than its size of cross-section.

[6] Some other external force may be considered.

where $f(x) = g\,\rho(x)$, external force per unit length.

The equation (1.11) must come with appropriate physical boundary conditions to ensure it is well-posed.

1.3.2 Boundary conditions

(a) Both ends of the elastic bar are fixed, so no displacements:

$$u(0) = 0, \quad u(1) = 0.$$

This is called Dirichlet boundary conditions.

(b) Top end of the elastic bar is fixed (no displacement), the other end is free (no internal force since it is in the air):

$$u(0) = 0, \quad w\big|_{x=1} = c(x)\frac{du}{dx}\Big|_{x=1} = 0\ .$$

The first is called a Dirichlet boundary condition, the second is called a Neumann boundary condition. The boundary conditions

$$u(0) = 0, \quad \text{or} \quad u(1) = 0$$

or

$$c(x)\frac{du}{dx}\Big|_{x=1} = 0$$

are all called homogeneous boundary conditions, while the boundary conditions

$$u(0) = 1, \quad \text{or} \quad u(1) = -2,$$

or

$$c(x)\frac{du}{dx}\Big|_{x=1} = -3$$

are all called non-homogeneous boundary conditions.

So the complete model for an elastic bar is :

$$-\frac{d}{dx}(c(x)\frac{du}{dx}) = f(x), \quad 0 < x < 1$$

with boundary conditions

$$u(0) = 0, \quad u(1) = 0$$

or

$$u(0) = 0, \quad c(x)\frac{du}{dx}\Big|_{x=1} = 0\ .$$

This differential equation is called a two-point boundary value problem[7].

[7] Think about why we need two boundary conditions.

1.4 Solutions of the elastic bar model

We now try to find the solution of the following boundary value problem

$$\begin{cases} -\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) = f(x), \quad 0 < x < 1 \\ u(0) = 0, \quad c(x)\frac{du}{dx}\big|_{x=1} = 0\,. \end{cases} \tag{1.12}$$

Solution. Integrating the equation (1.12) over $(x, 1)$, we obtain

$$-c(x)\frac{du}{dx}\Big|_x^1 = \int_x^1 f(t)\,dt\ ,$$

using the boundary conditions, we have

$$c(x)u'(x) = \int_x^1 f(t)dt\ ,$$

or

$$u'(x) = \frac{1}{c(x)}\int_x^1 f(t)dt\,.$$

Integrating over $(0, x)$ gives

$$u(x) = \int_0^x \frac{1}{c(x)}\int_x^1 f(t)dtdx, \tag{1.13}$$

this is the required exact solution of the problem (1.12). ♯

Example 1.1. *Find the exact solution of the problem*

$$\begin{cases} -\frac{d^2u}{dx^2} = x^2, \quad 0 < x < 1 \\ u(0) = 0, \quad \frac{du}{dx}\big|_{x=1} = 0\,. \end{cases} \tag{1.14}$$

Solution. Integrating the equation (1.14) over $(x, 1)$, we obtain

$$-\frac{du}{dx}\Big|_x^1 = \int_x^1 t^2\,dt\ ,$$

using the boundary conditions, we have

$$u'(x) = \int_x^1 t^2dt = \frac{1}{3} - \frac{1}{3}x^3.$$

Integrating over $(0, x)$ gives

$$u(x) = \int_0^x (\frac{1}{3} - \frac{1}{3}t^3)dt = \frac{1}{3}x - \frac{1}{12}x^4\,. \tag{1.15}$$

It is easy to verify that this $u(x)$ is really the solution of the system (1.14). ♯

Example 1.2. *Find the exact solution of the following problem*

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) = f(x), \quad 0 < x < 1$$

with the boundary conditions

$$u(0) = -1,\ u(1) = 1\,.$$

__**Solution**__. Write the equation as

$$-(c(x)u'(x))' = f(x),$$

then integrating over $(x, 1)$, we get

$$c(x)u'(x) = C_0 - \int_x^1 f(t)\,dt,$$

where C_0 is an integration constant. This implies

$$u'(x) = \frac{C_0}{c(x)} - \frac{1}{c(x)}\int_x^1 f(t)\,dt\,.$$

Now integrating over $(0, x)$ gives

$$u(x) = -1 + C_0\int_0^x \frac{1}{c(t)}dt - \int_0^x \frac{1}{c(x)}\int_x^1 f(t)\,dt\,.$$

Using the boundary condition $u(1) = 1$, we can find the integration constant C_0. $\sharp$

1.5 Connection between continuous model and discrete model

The analogy between the continuous model and the discrete model. We make a table to show the analogy between the spring-mass model and the elastic bar model.

We should explain why the conjugate of the operator d/dx is $-d/dx$. To see this, we consider the following spaces:

$$\begin{aligned} C^1(0,1) &= \{u : [0,1] \to \mathbb{R} | u \text{ is continuously differentiable. }\} \\ C_0^1(0,1) &= \{u : [0,1] \to \mathbb{R} | u \text{ is continuously differentiable and } u(0) = u(1) = 0\} \end{aligned}$$

These are vector spaces. We can define inner product

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

variables and relations	spring-mass	elastic bar
displacement	u_i	$u(x)$
elongation (strain)	e_j	$e(x)$
restoration force (stress)	w_j	$w(x)$
gravitation force	f_i	$f(x)$
connection relation	$e = Au$	$e = \frac{d}{dx}u$
Hook's law	$w = Ce$	$w(x) = c(x)e(x)$
Force balance law	$A^t w = f$	$-\frac{d}{dx}w = f$

We may think the differential operator d/dx maps $C[0,1]$ into $C[0,1]$. The formula of integration by parts gives

$$\int_0^1 \frac{du}{dx} v dx = -\int_0^1 u \frac{dv}{dx} dx + [uv]\Big|_{x=0}^{x=1}.$$

When $u, v \in C_0^1[0,1]$, we get

$$(\frac{d}{dx}u, v) = (u, -\frac{d}{dx}v).$$

This is why the conjugate of d/dx is $-d/dx$.

Elastic bar model is a continuous limit of the spring-mass system. In the continuous model (1.11), we divide the domain $[0,1]$ into $n+1$ subintervals uniformly, each has length $\Delta x = 1/(n+1)$. We label grid points $i\Delta x$ by x_i. We imagine there are masses m_i at x_i with springs connecting them consecutively. Each spring has length Δx while it is at rest. According to the spring-mass model, we have

$$c_i(u_i - u_{i-1}) - c_{i+1}(u_{i+1} - u_i) = m_i g. \tag{1.16}$$

where c_i is the spring constant of the spring connecting x_i to x_{i+1}. As $\Delta x \approx 0$ with $x_i \approx x$, we have

$$m_i \approx \rho(x_i)\Delta x,\ c_i \approx c(x_{i-1/2})/\Delta x.$$

Here, ρ is the density. Why the spring constant is prportitional to $1/\Delta x$? Think about the problem: Let us connect n springs with the same spring constant, what is the resulting spring constant?

Now, we this approximation, we get that for small Δx, the spring-mass system becomes

$$\frac{1}{\Delta x}\Big(c(x_{i-1/2})(u_i - u_{i-1}) - c(x_{i+1/2})(u_{i+1} - u_i)\Big) = \rho(x_i)\Delta x. \tag{1.17}$$

As we take $\Delta x \to 0$, we get the equation for the elastic bar:

$$-\frac{d}{dx}\Big(c(x)\frac{d}{dx}u(x)\Big) = f(x),$$

where $f = g\rho$.

Notice that the end displacements u_0 and u_{n+1} satisfy the fix-end boundary conditions

$$u_0 = 0,\ u_{n+1} = 0.$$

which correspond to the boundary condition of $u(\cdot)$ in the elastic bar model:

$$u(0) = 0,\ u(1) = 0.$$

Remark. This part comes from pp. 153–162.

Homeworks.

- pp. 164, 3.1.2,
- pp. 164, 3.1.4,
- pp. 164, 3.1.6,
- pp. 165, 3.1.15,
- pp. 165, 3.1.16

2 Equilibrium Equations and Variation Principles

In previous section, we have formulated the spring-mass system and an elastic bar model based on *force balance laws.* This is indeed the *Newtonian mechanics.* In solving constrained motion problem, *Lagrange* reformulate the Newtonian mechanics in variation form. This formulation solves not only the constrained problem, but has its own generality and becomes a fundation of mechanics. This section is devoted to the variation formulation of the spring-mass system and elastic bar problem.

In previous section, we have formulated the spring-mass system in equilibrium as algebraic equations. We have formulated in two ways:

- the displacement-force form

$$Ku := A^t C A u = f. \tag{2.1}$$

  where

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \ A = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ & & & -1 \end{pmatrix}, \ C = \begin{pmatrix} c_1 & & & \\ & c_2 & & \\ & & \ddots & \\ & & & c_{n+1} \end{pmatrix}$$

- the stress-displacement form:

$$\begin{pmatrix} C^{-1} & A \\ A^t & 0 \end{pmatrix} \begin{pmatrix} -w \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ -f \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_{n+1} \end{pmatrix}. \tag{2.2}$$

The first formulation is equivalent to a minimizational problem:

$$\mathrm{Min}_u \ P(u) := \frac{1}{2}(Ku, u) - (f, u).$$

The physical meaning of this quantity is the total potential energy. The first term is the potential energy stored in the springs. The second term is the gravitation potential energy of the masses. Thus, the equilibrium solution is the minimal energy solution.

The second formulation will be reformulated as a *constrained minimization problem*:

$$\mathrm{Min}_w \ Q(w) := \frac{1}{2}(C^{-1}w, w)$$

$$\text{with constraint } A^t w = f.$$

We shall show the two variational formulations are equivalent.

We will also derive the same variational formulations for the equilibrium elastic bar system.

2.1 Variation formulation for the spring-mass system

2.1.1 Mimimum principle

Consider the functional

$$P(u) := \frac{1}{2}(Ku, u) - (f, u),$$

where K is a symmetric positive definite matrix in $\mathbb{R}^n$. The directional derivative of P at u in the direction v is defined as

$$P'(u)v = \left.\frac{d}{dt}\right|_{t=0} P(u + tv)$$

$P'(u)$ is called the gradient (or the first variation) of P at u. We can compute this gradient: [8]

$$\begin{aligned} P'(u)v &= \left.\frac{d}{dt}\right|_{t=0} \frac{1}{2}(K(u + tv), u + tv) - (f, u + tv) \\ &= \frac{1}{2}\Big((Kv, u) + (Ku, v)\Big) - (f, v) \\ &= (Ku - f, v). \end{aligned}$$

Here, we have used K being symmetric. Thus,

$$P'(u) = Ku - f.$$

The second derivative is the Hessian. It is

$$P''(u) = K.$$

If u^* is a minimum of $P(v)$, then $P'(u^*) = 0$. This is called the Euler-Lagrange equation of P.

[8] Here, I use the following properties: $(f, g)' = (f', g) + (f, g')$. This is because $(f, g) = \sum_i f_i g_i$ and $(f, g)' = \sum_i \Big(f_i' g_i + f_i, g_i'\Big) = (f', g) + (f, g')$.

Conversely, If u^* satisfies the Euler-Lagrange equation $Ku^* = f$, then u^* is the minimum of $P(v)$. In fact, for any v, we compute $P(v) - P(u^*)$. We claim

$$P(v) - P(u^*) = \frac{1}{2}(K(v-u^*), v-u^*).$$

To see this, since $P(v)$ is a quadratic function of v, we can complete the squares:

$$\begin{aligned}
P(v) - P(u^*) &= \frac{1}{2}(Kv, v) - (f, v) - \frac{1}{2}(Ku^*, u^*) + (f, u^*) \\
&= \frac{1}{2}(Kv, v) - \frac{1}{2}(Ku^*, u^*) - (f, v-u^*) \\
&= \frac{1}{2}(Kv, v) - \frac{1}{2}(Ku^*, u^*) - (Ku^*, v-u^*) \\
&= \frac{1}{2}(Kv, v) + \frac{1}{2}(Ku^*, u^*) - (Ku^*, v) \\
&= \frac{1}{2}(K(v-u^*), v-u^*) \geq 0.
\end{aligned}$$

Hence we get that u^* is a minimum. In fact, u^* is the only minimum because $P(v) = P(u^*)$ if and only if $(K(v-u^*), v-u^*) = 0$. Since K is positive definite, we get $v - u^* = 0$.

We conclude the above discussion as the follows.

Let $P(u) := \frac{1}{2}(Ku, u) - (f, u)$ and K is symmetric positive definite. The vector u^* which minimizes $P(v)$ must satisfy the Euler-Lagrange equation $P'(u^*) = Ku^* - f = 0$. The converse is also true.

The physical meaning of P is the *total potential energy* of the spring-mass system. Indeed,

$$\frac{1}{2}(CAu, Au) = \sum_{i=1}^{n} \frac{1}{2} c_i (u_i - u_{i-1})^2$$

is the sum of the *potential energy stored in the spring,* whereas the term

$$(f, u) = \sum_{i=1}^{n} f_i u_i$$

is the sum of the *works* done by the mass m_i with displacement u_i for $i = 1, ..., n$. The term $-(f, u)$ is the *gravitational potential* due to the masses m_i with displacements u_i.

2.1.2 Constrained Minimization

Next, let us study a minimization problem with constraint. We recall an important technique to solve constrained optimization problem is the technique of *Lagrange multiplier.* It converts a constrained optimization problem to an optimization problem *without* *constraint.* Consider

$$\mathrm{Min} f(x,y,z) \text{ subject to } g(x,y,z)=0. \tag{2.3}$$

This problem is equivalent to a variation problem *without constraint*: Consider

$$L(x,y,z,\lambda) := f(x,y,z) + \lambda g(x,y,z) \tag{2.4}$$

The critical solution of (2.3) is also a critical solution of (2.4), and vice versa. The critical solution of (2.4) satisfies

$$\frac{\partial L}{\partial x}=0,\ \frac{\partial L}{\partial y}=0,\ \frac{\partial L}{\partial z}=0,\ \frac{\partial L}{\partial \lambda}=0.$$

There are four equations for four unknowns (x,y,z,λ), we can solve it in principle. The last equation simply means that the gradient of f is parallel to the gradient of g.

If there are two constraints, say

$$g_1(x,y,z)=0,\ g_2(x,y,z)=0,$$

then we simply add one more Lagrange multiplier, namely, define $L = f + \lambda_1 g_1 + \lambda_2 g_2$ and look for critical solution of L:

$$\frac{\partial L}{\partial x}=0,\ \frac{\partial L}{\partial y}=0,\ \frac{\partial L}{\partial z}=0,\ \frac{\partial L}{\partial \lambda_1}=0,\ \frac{\partial L}{\partial \lambda_2}=0.$$

The last two equations means that ∇f is a linear combination of ∇g_1 and ∇g_2.

Next, for the spring-mass system, we consider the following variational problem with constraint:

$$\mathrm{Min}\ Q(w) := \frac{1}{2}(C^{-1}w,w),\ \text{subject to } A^t w = f. \tag{2.5}$$

The minimization problem of the equilibrium spring-mass system states that the stress w is determined by minimizing the potential energy it determines under a force balance constraint $A^t w = f$. [9]

[9] We have learned that the physical meaning of w is the stress, or the restoration force, or the internal force in the spring. Notice that w is a vector $(w_1,...,w_{n+1})$. So, its component w_j is the stress of the spring j. The quantity $c_1^{-1}w_1$ is the elongation of the spring 1. When we vary w_1 from 0 to w_1, the potential energy stored in the spring 1 is

$$\int_0^{w_1} c_1^{-1} w_1\, dw_1 = \frac{1}{2} c_1^{-1} w_1^2.$$

After we sum these potential energies over all springs, we get that $Q(w) = \frac{1}{2}(C^{-1}w,w)$ the total potential energy stored in all springs due to the stress w.

Since The constraint equation $A^t u - f = 0$ is in $\mathbb{R}^n$, there should be n Lagrange multipliers, say, $u_1, ..., u_n$, or, in short, the Lagrange multiplier $u \in \mathbb{R}^n$. Consider

$$L(u, w) := \frac{1}{2}(C^{-1}w, w) - (u, A^t w - f). \tag{2.6}$$

The Euler-Lagrange equation for this unconstrained variation problem is

$$\begin{aligned}\frac{\partial L}{\partial w} &= C^{-1}w - Au = 0\\ \frac{\partial L}{\partial u} &= -A^t w + f = 0,\end{aligned}$$

which is precisely the equation (2.2). In the Lagrange formulation, he introduced the Lagrange multiplier u, which is indeed the *displacements* of the masses, and from $\partial L/\partial w = 0$, we get that they are related to the stress w by $w = CAu$.

So far, we have seen that if w^* satisfies the constrained minimization problem, then there exists u^* such that (u^*, w^*) is a critical point (*unconstrained*) of $L(u, w)$.

Conversely, if (u^*, w^*) is a critical point of $L(u, w)$, we want to show that $Q(w) \geq Q(w^*)$ for all w satisfying the constraint $A^t w = f$. To see this, since Q is a quadratic function, we have

$$\begin{aligned}Q(w) - Q(w^*) &= \frac{1}{2}(C^{-1}w, w) - \frac{1}{2}(C^{-1}w^*, w^*)\\ &= (C^{-1}w^*, w - w^*) + \frac{1}{2}(C^{-1}(w - w^*), w - w^*)\end{aligned}$$

We claim the first term is zero. To see this, since both w and w^* satisfy the constraint $A^t w = f$, we get $A^t(w - w^*) = 0$. Next, from $\partial L/\partial w = 0$, $C^{-1}w^* = Au^*$. Hence,

$$(C^{-1}w^*, w - w^*) = (Au^*, w - w^*) = (u^*, A^t(w - w^*)) = 0.$$

Thus, we get

$$Q(w) - Q(w^*) = \frac{1}{2}(C^{-1}(w - w^*), w - w^*) \geq 0.$$

Further, w^* is the unique minimum among all w satisfies the constraint.

We conclude this discussion as the follows.

The constrained minimization problem:

$$\textbf{Min } Q(w) := \frac{1}{2}(C^{-1}w, w), \textbf{ subject to } A^t w = f$$

is equivalent to the unconstrained variational problem:

$$\textbf{Find the critical points of } L(u, w) := \frac{1}{2}(C^{-1}w, w) - (u, A^t w - f)$$

whose solution should satisfy the Euler-Lagrange equation:

$$\begin{aligned} \frac{\partial L}{\partial w} &= C^{-1}w - Au = 0 \\ \frac{\partial L}{\partial u} &= -A^t w + f = 0. \end{aligned}$$

Remark. The critical point (u^*, w^*) is indeed a saddle point. You may think about the case $u \in \mathbb{R}$ and $w \in \mathbb{R}$. In this case, C is a scalar, say c and A is also a scalar, say a. The corresponding $L(u, w) = \frac{1}{2}c^{-1}w^2 - auw + uf$. One can readily see that the critical point of this function L is a saddle point.

Below, we give a more detail description about this saddle point. We shall establish the following equivalence table:

(1) $P'(u) = Ku - f = 0$, where $K = A^t CA$	(2) $\begin{cases} \frac{\partial L}{\partial w} = C^{-1}w - Au = 0 \\ \frac{\partial L}{\partial u} = -A^t w + f = 0 \end{cases}$ $L(u, w) := \frac{1}{2}(C^{-1}w, w) - (u, A^t w - f)$
(3) $\text{Max}_u(-P(u)) := \frac{1}{2}(CAu, Au) - (f, u)$	(4) Min $Q(w) := \frac{1}{2}(C^{-1}w, w)$, subject to $A^t w = f$
(5) $\text{Max}_u \ \text{Min}_w L(u, w)$	(6) $\text{Min}_w \ \text{Max}_u L(u, w)$

We have established the equivalences: (1)⇔(2), (1)⇔(3), (2)⇔(4). Next, we show

(4)⇔(6). In fact,

$$\mathrm{Max}_u L(u,w) = \mathrm{Max}_u Q(w) - (u, A^t w - f) = \begin{cases} Q(w) & \text{if } A^t w - f = 0, \\ +\infty & \text{if } A^t w - f \neq 0 \end{cases}$$

Thus, solving (6) is equivalent to (4).

Next, we show (3)⇔(5). For each fixed u, we solve $\mathrm{Min}_w L(u,w)$. If $\bar{w}(u)$ is the minimum of $L(u,\cdot)$ then $\bar{w}(u)$ satisfies

$$\frac{\partial L}{\partial w} = 0,$$

which gives $C^{-1}\bar{w} - Au = 0$, or $\bar{w} = CAu$. We plug this into $L(u,w)$ to get

$$\begin{aligned} L(u,\bar{w}(u)) &= \frac{1}{2}(CAu, Au) - (u, A^t CAu) + (u,f) \\ &= -\frac{1}{2}(A^t CAu, u) + (u,f) \\ &= -P(u). \end{aligned}$$

The critical point $\bar{w}(u)$ is indeed a minimum. In fact,

$$\begin{aligned} L(u,w) - L(u,\bar{w}(u)) &= \frac{1}{2}(C^{-1}w,w) - (u, A^t w - f) - \frac{1}{2}(C^{-1}\bar{w}(u), \bar{w}(u)) + (u, A^t\bar{w}(u) - f) \\ &= \frac{1}{2}(C^{-1}w,w) - \frac{1}{2}(C^{-1}\bar{w}(u) - (u, A^t(w - \bar{w}(u))) \\ &= \frac{1}{2}(C^{-1}w,w) - \frac{1}{2}(C^{-1}\bar{w}(u) - (Au, w - \bar{w}(u)) \\ &= \frac{1}{2}(C^{-1}w,w) - \frac{1}{2}(C^{-1}\bar{w}(u) - (C^{-1}\bar{w}(u), w - \bar{w}(u)) \\ &= \frac{1}{2}(C^{-1}w,w) + \frac{1}{2}(C^{-1}\bar{w}(u) - (C^{-1}\bar{w}(u), w) \\ &= (C^{-1}(w - \bar{w}(u)), w - \bar{w}(u)) \geq 0. \end{aligned}$$

Thus,

$$\mathrm{Min}_w L(u,w) = -P(u).$$

This shows (3)⇔(5).

We call the minimization problem (4) *the primal problem:*

$$\textbf{Primal Problem:}\quad \text{Min } Q(w) := \frac{1}{2}(C^{-1}w, w), \text{ subject to } A^t w = f.$$

and the maximization problem (3) *the dual problem.*

$$\textbf{Dual Problem:}\quad \text{Max}(-P(u)) := \frac{1}{2}(CAu, Au) - (f,u) \tag{2.7}$$

The equivalence between the primal and dual problems can be stated as the following theorem.

1. **The primal problem (2.5) and the dual problem (2.7) are equivalent. This means that if w^* and u^* are respectively the solutions of (2.5), (2.7), then they are related through**

$$C^{-1}w^* = Au^*.$$

2. **The solution pair (u^*, w^*) is a saddle point of L and it satisfies**

$$P(u^*) + Q(w^*) = 0. \tag{2.8}$$

Or equivalently

$$\mathbf{Max}_u\ \mathbf{Min}_w L(u,w) = \mathbf{Min}_w\ \mathbf{Max}_u L(u,w) = L(u^*,w^*) \tag{2.9}$$

Remark. The equivalence (2.9) is called the *MiniMax principle*. For a general L, we always

$$\mathrm{Max}_u\ \mathrm{Min}_w L(u,w) \le \mathrm{Min}_w\ \mathrm{Max}_u L(u,w)$$

To see this, we have for any u_1 and w,

$$L(u_1,w) \le\ \mathrm{Max}_u L(u,w)$$

We then take minimum over w. This gives

$$\mathrm{Min}_w L(u_1,w) \le\ \mathrm{Min}_w\ \mathrm{Max}_u L(u,w)$$

The right-hand side is a number. We then take maximum over u_1. This gives

$$\mathrm{Max}_{u_1}\ \mathrm{Min}_w L(u_1,w) \le\ \mathrm{Min}_w\ \mathrm{Max}_u L(u,w).$$

2.2 A general framework for applications

In the spring-mass system, we have masses m_i connected by springs. In general, we consider a network consisting of n nodes and m directed edges. These masses m_j are the nodes, whereas the springs are the directed edges. The direction means that the positive direction is from m_{j-1} to m_j. At each node, we associate it with a displacement u_j, called the nodal variable. At each spring, we associate it with an elongation of the spring e_i and a restoration force w_i, called the edge variable. They are related by the Hook's law: $w = Ce$. We then use connectivity relation and force balance laws to derive equations for the nodal and edge variables.

Such an approach can be quite general. Let us use the following electricsl network to understand general framework.

Electrical network. Consider an electrical network which consists of wires with resistors, batteries. The nodes are those points where two wires meet. The directed edges are the wires and the direction indicates the direction of the current. Suppose there are n nodes and m directed edges. At each node, a potential x_j is associated with. This is the nodal variable. On each edge, the variable e_i (with sign), $i = 1, ..., m$ denotes the potential drop on edge i. The connectivity of the nodes and edges can be characterized by the following matrix A^0: its column represents the nodes, and its row represent the edges; its entries are defined by

$$a_{i,j} = \begin{cases} 1 & \text{means that edge } i \text{ enters node } j \\ -1 & \text{means that edge } i \text{ leaves node } j \\ 0 & \text{means that edge } i \text{ does not connect to node } j \end{cases} \tag{2.10}$$

This matrix A^0 is called connectivity matrix. Notice that the sum of each row of A^0 is zero because for each directed edge i, its starting node has coefficient -1 and end node has coefficient $+1$. This means that only the difference of x_j and x_k is mattered. Hence, we can normalize a particular node so that its nodal value $x_j = 0$. Such a x_j is called *grounded*. By eliminating this x_j, we call the reulting nodal variable x and the resulting matrix A. Notice that the remaining x_j are independent now. Notice that The column vector a_j records the the connection of edges to node x_j. The independence of x_j means that the column vectors $a_1, ..., a_n$ of A are independent. This is equivalent to say that the matrix $A^tA = ((a_i, a_j))_{n\times n}$ are symmetric positive definite matrix. This is our basic assumption on A. The node-edge relation is denoted by

$$e = -Ax.$$

If there is a battery connected to that edge, there is an additional potential drop or increase, depending on the connected direction of the battery. So in general we have

$$e = b - Ax,$$

where b represents the battery potential drop.

On each edge, we associate it with a current y_j, called the edge variables. The relation between e and y is defined by the Ohm's law: $e = Ry$, where R is the resistent:

$$R = \begin{pmatrix} R_1 & & & \\ & R_2 & & \\ & & \ddots & \\ & & & R_m \end{pmatrix}.$$

Finally, at each node, there is a Kirchhoff's current law, which requiring that the net flow at node i equals an appled source f_i. That is

$$A^t y = f.$$

General framework Consider a directed network which consists of N nodes and m directed edges. At each node, we associate it with a nodal variable x_i, $i = 1, ..., N$. At each edge, we associate with three edge variables e_j, b_j and y_j, $j = 1, ..., m$. The nodal variable x and the edge variable e are connected by the connectivity matrix A with possible extra input b. That is,

$$e = b - Ax$$

where A is defined by (2.10). The edge variable y and e are related by some physical law:

$$y = Ce.$$

At each node, a Kirchhoff current law should be satisfied

$$A^t y = f.$$

Or in matrix form:

$$\begin{pmatrix} C^{-1} & A \\ A^t & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix} \tag{2.11}$$

These are called the fundamental equilibrium equations for the directed network.

We list the table of the discrete models that we have studied.

For general directed graph case, the corresponding variations are the follows. We define

$$Q(y) = \frac{1}{2}(C^{-1}y, y) - (b, y),$$

Directed graph	spring-mass	electrical network
nodal variable: x_j nodal source: f_j	displacements: u_j gravitational force f_j	potential x_j current inject f_j
edge variables: e_i edge variables: y_i edge sources: b_i	elongation e_i restoration force w_i	potential drop e_i current y_i battery b_i
egde relation: $y = Ce$	$w = Ce$	$y = Ce$
node-edge connection: $e = b - Ax$	$e = Au$	$e = b - Ax$
edge-node connection: $A^t y = f$	$A^t w = f$	$A^t y = f$

The constraint is

$$A^t y = f.$$

The Lagragian L becomes

$$L(x, y) = Q(y) + (x, A^t y - f).$$

The dual function P is defined to be

$$P(x) := -\ \mathrm{Min}_y L(x, y) = \frac{1}{2}(C(b - Ax), b - Ax) + (x, f)$$

We have the same equivalence table.

(1) $P'(x) = A^t C(Ax - b) + f = 0$	(2) $\begin{cases} \frac{\partial L}{\partial y} &= C^{-1}y + Ax - b = 0 \\ \frac{\partial L}{\partial x} &= A^t y - f = 0 \end{cases}$ $L(x, y) := \frac{1}{2}(C^{-1}y, y) - (b, y) - (x, A^t y - f)$
(3) $\mathrm{Max}_x(-P(x))$	(4) Min $Q(y)$ subject to $A^t y = f$
(5) $\mathrm{Max}_x\ \mathrm{Min}_y L(x, y)$	(6) $\mathrm{Min}_y\ \mathrm{Max}_x L(x, y)$

A geometric intepretation To give a geometric intepretation, we first assume $C = I$. This assumption does not hurt because we can make a change of variable $y' = C^{-1/2}y$. Then we use y' instead of y. Next, we can replace

$$Q(y) = \frac{1}{2}(y, y) - (y, b)$$

by

$$Q(y) = \frac{1}{2}\|y - b\|^2.$$

because their minimization in y differes only by a constant $\|b\|^2/2$. Thirdly, we may assume either $f = 0$ or $b = 0$.

1. For the first case, we first find y_0 such that $A^t y_0 = f$. With this, the constraint becomes

$$A^t(y - y_0) = 0,$$

We then replace $y - y_0$ by y and $b - y_0$ by b. With this substitution, the function $Q(y)$ remains unchange. But, can we find such a y_0? The answer is positive. For we can solve

$$A^t A x_0 = f$$

because $A^t A$ is symmetric positive definite. Then we choose $y_0 = Ax_0$.

2. For the second case, we replace $y - b$ by y. Then the constraint becomes

$$A^t(y + b) = f$$

We then replace $f - A^t b$ by f.

Let us study geometric intepretation for the first case. The second case is the same through a translation of y.

Our primal problem is

$$\text{Min}_y \frac{1}{2}\|y - b\|^2 \text{ subject to } A^t y = 0.$$

Here, $y, b \in \mathbb{R}^m$. A is a $m \times n$ matrix with column vectors $a_1, ..., a_n$. That is, $A = (a_1, ..., a_n)$. Let us first characterize the constraint $A^t y = 0$. Let

$$W_{a_j} = \{y|(y, a_j) = 0\}.$$

It is a hyperplane in $\mathbb{R}^m$ with normal a_j. Let

$$\begin{aligned} W &= \{y|\ A^t y = 0\} = \{y|(y, a_j) = 0, j = 1, ..., n\} \\ &= W_{a_1} \cap \cdots \cap W_{a_n} \end{aligned}$$

Thus, the constraint $A^t y = 0$ means that $y \in W$ and the primal problem is to find the shortest distance from b to W. This is nothing but the orthogonal projection of b onto W. Let us call this projection y^*. For instance, let us take $m = 3$ and $n = 2$. $A = (a_1, a_2)$ with

$$a_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \ a_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

The constraint space W is the z-axis. If $b = (1, 2, 3)^t$, then $y^* = (0, 0, 3)^t$.

In general, it is not easy to find y^* directly. So we will find it through the help of the dual problem. This will explain below.

Next, let us study the the Lagragian.

$$L(x, y) = \frac{1}{2}\|y - b\|^2 + (x, A^t y) = \frac{1}{2}\|y - b\|^2 + (Ax, y)$$

We notice that $Ax = x_1 a_1 + \cdots + x_n a_n \in \mathbb{R}^m$. Let

$$V = \{a = x_1 a_1 + \cdots + x_n a_n | x \in \mathbb{R}^n\} = \mathrm{Span}\{a_1, ..., a_n\}$$

It is important to notice that $\mathbb{R}^m = V \bigoplus W$. This means that $V \perp W$ and $\mathbb{R}^m = V + W$. To see $V \perp W$, we know that any $a \in V$ is a linear combination of a_i, $i = 1, ..., n$, and any $y \in W$ is perpendicular to $a_i, i = 1, ..., n$. Hence $a \perp y$. To see $\mathbb{R}^m = V + W$, let $b \in \mathbb{R}^m$, we project b onto V by finding $a^* = \sum_j x_j^* a_j$ such that

$$(b - a^*, a_i) = 0, i = 1, ..., n.$$

This leads to the equations

$$\sum_{j=1}^{n} (a_i, a_j) x_j^* = (b, a_i), i = 1, ..., n$$

Or in matrix form

$$A^t A x^* = g$$

where $g_i = (b, a_i)$. We have known this is solvable because $A^t A$ is symmetric positive definite. With this projection a^*, the vector b is decomposed into

$$b = a^* + (b - a^*)$$

with a^* in V and $(b - a^*) \in W$.

Each $x \in \mathbb{R}^n$ can be identify a vector $a = \sum_{i=1}^n x_i a_i \in V$ and vice versa. So we write $L(x, y)$ as

$$L(a, y) = \frac{1}{2}\|b - y\|^2 + (a, y).$$

The minimum over y gives

$$-P(a) = \text{ Min}_y L(a, y).$$

This minimum occurs at

$$\frac{\partial L}{\partial y} = 0.$$

That is

$$\bar{y} - b + a = 0.$$

Hence

$$\begin{aligned} -P(a) &= L(x, \bar{y}(x)) = \frac{1}{2}\|a\|^2 + (a, b - a) \\ &= -\frac{1}{2}\|a - b\|^2 + \frac{1}{2}\|b\|^2. \end{aligned}$$

Thus, the dual problem

$$\text{Max}_{a \in V}(-P(a)) := -\frac{1}{2}\|a - b\|^2 + \frac{1}{2}\|b\|^2$$

is equivalent to find

$$\text{Min}_{a \in V} \frac{1}{2}\|b - a\|^2.$$

We can simply project orthogonally b onto V. Let us call this projection a^*, or $a^* = \sum_{j=1}^n x_j^* a_j$.

Now, y^* is the orthogonal projection of b on W whereas a^* is the orthogonal projection of b onto W. From $\mathbb{R}^m = V \bigoplus W$, we get $y^* = b - a^*$. By the pythagoras' law

$$\|b\|^2 = \|a^*\|^2 + \|y^*\|^2.$$

But this is precisely the duality theorem. It says

$$-P(a^*) = \max_{a \in V} \min_{y \in \mathbb{R}^m} L(a, y) = \min_{y \in \mathbb{R}^m} \max_{a \in V} L(a, y) = \min_{y \in W} Q(y) = Q(y^*)$$

We have seen that $-P(a^*) = -\frac{1}{2}\|a^* - b\|^2 + \frac{1}{2}\|b\|^2$, whereas $Q(y^*) = \frac{1}{2}\|b - y^*\|^2$. So, the duality theorem $-P(x^*) + Q(y^*) = 0$ is nothing but the Pythagoras' law.

Next, let us give another geometric picture. Let us rewrite L as

$$L(\lambda a, y) = \frac{1}{2}\|b - y\|^2 + \lambda(a, y)$$

with $a \in V$ and $\|a\| = 1$, $\lambda \in \mathbb{R}$. The primal problem is

$$\min_{y\in\mathbb{R}^m}\max_{x\in\mathbb{R}^n} L(x,y) = \min_{y\in\mathbb{R}^m}\max_{a\in V} L(a,y) = \min_{y\in\mathbb{R}^m}\max_{a\in V,\|a\|=1}\max_{\lambda\in\mathbb{R}} L(\lambda a, y).$$

We want to convert it to

$$\min_{y\in\mathbb{R}^m}\max_{a\in V,\|a\|=1}\max_{\lambda\in\mathbb{R}} L(\lambda a, y) = \max_{a\in V,\|a\|=1}\min_{y\in\mathbb{R}^m}\max_{\lambda\in\mathbb{R}} L(\lambda a, y).$$

It is easy to see that

$$\max_{\lambda\in\mathbb{R}} L(\lambda a, y) = \begin{cases} \frac{1}{2}\|b-y\|^2 & \text{if } (a,y)=0 \\ +\infty & \text{otherwise.} \end{cases}$$

Thus,

$$\min_{y\in\mathbb{R}^m}\max_{\lambda\in\mathbb{R}} L(\lambda a, y) = \min_{y\in W_a} \frac{1}{2}\|y-b\|^2.$$

This is equivalent to find the minimal distance to the hyperlane W_a. We notice that $a \in V$ if and only if W_a passes through W (prove in homework). The answer to this minimal distance problem is simple. The minimum occurs at

$$\bar{y}(a) = b - \lambda a, (a, \bar{y}) = 0.$$

This gives

$$\bar{y}(a) = b - (a,b)a.$$

Here, we have used $\|a\| = 1$. The minimum distance is

$$-\tilde{P}(a) = \frac{1}{2}\|b-\bar{y}\|^2 = \frac{1}{2}|(a,b)|^2.$$

The maximum of $|(a,b)|^2$ over all $a \in V$ with $\|a\| = 1$ occurs at $a^*/\|a^*\|$. This is because the orthogonal decomposition:

$$b = a^* + y^*$$

with $a^* \in V$ and $y^* \in W$. Hence $(a,b)^2 = (a,a^*)^2$ with maximum occurs when $a = a^*/\|a^*\|$. The maximum distance is $\|a^*\|$. This means that $-P(x^*) = -\tilde{P}(a^*)$ is the maximum of the distance of b to all hyperplanes passing through W.

Remark. Read pp. 87-114.

Homeworks.

- pp. 94, 2.1.8
- pp. 95, 2.1.9
- pp. 95, 2.1.10
- Show that
$$-\ \mathrm{Min}_y L(x,y) = \frac{1}{2}(C(b-Ax), b-Ax) + (x,f)$$
- Let $P(x) := \frac{1}{2}(C(b-Ax), b-Ax) + (x,f)$, show that $P'(x) = A^t C(Ax-b) + f = 0$.
- pp. 107, 2.2.3 (a), (b)
- pp. 107, 2.2.5
- pp. 108, 2.2.7
- pp. 108, 2.2.10
- If $\mathbb{R}^m = V \bigoplus W$, where $V = \mathrm{span}\{a_1, ..., a_n\}$, then hyperplane W_a containing W if and only if $a \in V$.

2.3 Variational formulations for differential equations

In previous section, we have formulated the elastic bar problem by a differential equation. We can solve it analytically. However, this analytic method does not work in general, not for high dimensional problems, not even for a slightly general problems in one space dimension, the Sturm-Liouville system:

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u(x) = f(x),\ 0 < x < 1 \tag{2.12}$$

with boundary conditions

$$u(0) = 0, \qquad c(x)\frac{du}{dx}\big|_{x=1} = 0 \tag{2.13}$$

- **Remark.** The Sturm-Liouville system, with homogeneous or non-homogeneous boundary conditions have many physical applications. For instance,

  (a) in the quantum theory, the equation is called the Schrödinger's equation.

(b) For modeling the oscillations of a drum, it is called the Bessel's equation.

One important method to study the properties of the solutions to the equations (2.12)-(2.13) is to use the integral form, often called the variational formulation.

We shall discuss how to derive the variational formulation for the differential equation (2.12)-(2.13). The same methodology can be applied to any other second order differential equations.

The derivation is standard and simple. To do so, we multiply both sides of equation (2.12) by an arbitrary test function v satisfying $v(0) = 0$ to obtain

$$-\frac{d}{dx}(c(x)\frac{du}{dx})v + q(x)uv = f(x)v \ ,$$

then integrating over $(0, 1)$ gives

$$\int_0^1 \left(-\frac{d}{dx}(c(x)\frac{du}{dx})v + q(x)uv \right) dx = \int_0^1 f(x)vdx \ . \tag{2.14}$$

Now by integration by parts and the boundary conditions (2.13), we have

$$\int_0^1 \left(c(x)\frac{du}{dx}\frac{dv}{dx} + q(x)uv \right) dx = \int_0^1 f(x)v \ dx.$$

This leads to the **variational formulation** for the equations (2.12)-(2.13):

Find the solution u such that $u(0) = 0$ and

$$a(u, v) = g(v) \qquad \text{for any } v \text{ satisfying } v(0) = 0 \tag{2.15}$$

where $a(\cdot, \cdot)$ and $g(\cdot)$ are given by

$$\begin{aligned} a(u, v) &= \int_0^1 \left(c(x)\frac{du}{dx}\frac{dv}{dx} + q(x)uv \right) dx \,, \\ g(v) &= \int_0^1 f(x)v \ dx \,. \end{aligned}$$

Remark. The advantage of this formulation is that it involves only first order derivatives of u, not second derivatives in the original differential equation formulation. Thus, it has less regularity constraint on the solution u.

One can check that $a(\cdot, \cdot)$ is linear with respect to each variable, and is symmetric, i.e., for any u and v,

$$a(u, v) = a(v, u) \,.$$

Furthermore, we know that $a(\cdot,\cdot)$ is also positive, i.e.,

$$a(v,v) > 0 \quad \forall\, v \neq 0\,.$$

Equivalence between boundary value and variational problems

In the following, we shall verify that

The boundary value problem (2.12)-(2.13) is equivalent to the variational problem (2.15).

First, we know already that the solution u of the boundary value problem (2.12)-(2.13) is also a solution to the variational equation (2.15). Next, we will confirm that any solution u of (2.15) is also a solution of the boundary value problem (2.12)-(2.13).

In fact, since u satisfies (2.15), we have

$$\int_0^1 \Big(c(x)\frac{du}{dx}\frac{dv}{dx} + q(x)uv \Big)\, dx = \int_0^1 f(x)v\ dx \quad \forall\ v \text{ with } v(0) = 0\ .$$

Using integration by parts, we obtain

$$\int_0^1 \Big(-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big)v + q(x)uv \Big)\ dx + c(x)\frac{du}{dx}v\big|_{x=0}^{x=1} = \int_0^1 f(x)v\ dx\ . \tag{2.16}$$

As the test function v is arbitrary, we can take v to be arbitrary but satisfying the boundary conditions $v(0) = v(1) = 0$, then (2.16) becomes

$$\int_0^1 \Big\{ -\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u - f \Big\} v\ dx = 0 \quad \text{ for any } v \text{ with } v(0) = v(1) = 0\ ,$$

this implies

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u = f, \quad 0 < x < 1. \tag{2.17}$$

Substituting this into (2.16), we have

$$c(1)u_x(1)v(1) = 0 \quad \text{for any } \ v \text{ with } v(0) = 0\,,$$

this indicates that u also satisfies the condition

$$c(x)\frac{du}{dx}\big|_{x=1} = 0\ . \tag{2.18}$$

(2.17) and (2.18) tell us that u is a solution of the boundary value problem (2.12)-(2.13). ♯

2.3.1 Minimum principle for differential equation

Now we investigate the relation between the boundary value problem (2.12)-(2.13) and the following potential energy functional

$$P(u) = \frac{1}{2}\int_0^1 \left(c(x)\big(\frac{du}{dx}\big)^2 + q(x)u^2\right) dx - \int_0^1 f(x)\, u(x)\, dx,$$

we are going to verify the following relations:

The function u that minimizes $P(v)$ over all v satisfying $v(0) = 0$ must be the solution of the system (2.12)-(2.13), that is, it satisfies the differential equation

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u = f(x), \quad 0 < x < 1$$

with the boundary conditions

$$u(0) = 0 \quad \textbf{and} \quad c\frac{du}{dx}\big|_{x=1} = 0\ .$$

The converse is also true.

To see this, let u minimize $P(u)$, so we have

$$P(u) \le P(v) \quad \forall\ v \text{ with } \quad v(0) = 0\ . \tag{2.19}$$

Consider a real function

$$F(t) = P(u + tv)\ .$$

Using (2.19) we know

$$F(0) \le F(t) \quad \forall\, t \in \mathbb{R}^1\ ,$$

that is, $t = 0$ is a minimizer of $F(t)$. This implies

$$F'(0) = 0\ . \tag{2.20}$$

Now by definition,

$$\begin{aligned} F(t) - F(0) =& P(u + tv) - P(u) \\ =& \frac{1}{2}\Big\{\int_0^1 \big(c(x)(u_x + tv_x)^2 + q(x)(u + tv)^2\big)\, dx - \int_0^1 f(u + tv)dx\Big\} \\ &- \frac{1}{2}\Big\{\int_0^1 \big(c(x)u_x^2 + q(x)u^2\big)\, dx - \int_0^1 fu\ dx\Big\} \\ =& t\Big\{\int_0^1 (c(x)u_xv_x + q(x)uv)dx - \int_0^1 fv\ dx\Big\} + \frac{1}{2}t^2\int_0^1 c(x)v_x^2dx\ , \end{aligned}$$

which gives

$$\begin{aligned}
F'(0) &= \int_0^1 (c(x)u_x v_x + q(x)uv)dx - \int_0^1 fvdx \\
&= \int_0^1 (c(x)u_x v)_x - (c(x)u_x)_x v + q(x)uv)\,dx - \int_0^1 fvdx \\
&= \int_0^1 (c(x)u_x v)_x - (c(x)u_x)_x v + q(x)uv)\,dx - \int_0^1 fvdx + (c(x)u_x(x)v(x))_{x=0}^{x=1} \\
&= \int_0^1 \left[-(c(x)u_x)_x + q(x)u - f(x)\right] v\,dx + c(1)u_x(1)v(1)
\end{aligned}$$

for any v with $v(0) = 0$. This with (2.20) yields

$$\int_0^1 c(x)\left(\frac{du}{dx}\frac{dv}{dx} + q(x)uv\right) dx = \int_0^1 fvdx \quad \text{for any } v \text{ with } v(0) = 0\ ,$$

namely, u is a solution of the variational problem (2.15), so it is also a solution of the boundary value problem (2.12)-(2.13).

To see the converse part, for any v such that $v(0) = 0$ we can calculate

$$\begin{aligned}
P(v) - P(u) &= \Big\{\frac{1}{2}(c\,v_x, v_x) + (q\,v, v) - (f, v)\Big\} \\
&\quad -\Big\{\frac{1}{2}(c\,u_x, u_x) + (q\,u, u) - (f, u)\Big\} \\
&= \Big\{\frac{1}{2}\Big(c\,(v-u)_x, (v-u)_x\Big) + (q\,(v-u), v-u)\Big\} \\
&\quad +\Big\{\Big(c\,u_x, (v-u)_x\Big) + (q\,u, v-u) - (f, v-u)\Big\}\,.
\end{aligned}$$

Using this relation and the equivalence between the boundary value problem (2.12)-(2.13) and the variational problem (2.15), one can easily see that if u is a solution to the boundary value problem (2.12)-(2.13), then it must a minimizer of $P(v)$. ♯

2.3.2 Variational formulation for more general boundary conditions

We now consider a bit more general boundary condition problem:

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u(x) = f(x),\ \ a < x < b \tag{2.21}$$

with boundary conditions

$$c(x)\frac{du}{dx}\Big|_{x=a} = \alpha, \qquad u(b) = \beta \tag{2.22}$$

Same as we did in the last subsection, we can derive the variational formulation for the system (2.21)-(2.22).

To do so, we multiply both sides of equation (2.21) by an arbitrary test function v satisfying $v(b) = 0$, then integrate over (a, b) to obtain

$$\int_a^b \left(-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big)v + q(x)uv \right) dx = \int_a^b f(x)vdx \ . \tag{2.23}$$

Now using integration by parts and the boundary conditions (2.22), we deduce

$$\int_a^b \left(c(x)\frac{du}{dx}\frac{dv}{dx} + q(x)uv \right) dx = \int_a^b f(x)v \ dx - \alpha\, v(a).$$

This leads to the **variational formulation** for the equations (2.21)-(2.22):

Find the solution u such that $u(b) = \beta$ and

$$a(u, v) = g(v) \quad \text{ for any } v \text{ satisfying } v(b) = 0 \tag{2.24}$$

where $a(\cdot, \cdot)$ and $g(\cdot)$ are given by

$$\begin{aligned} a(u, v) &= \int_a^b \left(c(x)\frac{du}{dx}\frac{dv}{dx} + q(x)uv \right) dx\,, \\ g(v) &= \int_a^b f(x)v \ dx - \alpha\, v(a)\,. \end{aligned}$$

Equivalence between boundary value and variational problems

The same as we did in the last subsection, we can verify that

The boundary value problem (2.21)-(2.22) is equivalent to the variational problem (2.24).

First, we know already by the derivation of the variational problem (2.24) that the solution u of the boundary value problem (2.21)-(2.22) is also a solution to the variational equation (2.24). Next, we will confirm that any solution u of (2.24) is also a solution of the boundary value problem (2.21)-(2.22).

In fact, since u satisfies (2.24), we obtain by using integration by parts that for any v satisfying $v(b) = 0$,

$$\int_a^b \left(-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big)v + q(x)uv \right) \ dx + c(x)\frac{du}{dx}v\Big|_{x=a}^{x=b} = \int_a^b f(x)v \ dx - \alpha\, v(a) \ . \tag{2.25}$$

Now taking all the test functions v which satisfy the boundary conditions $v(a) = v(b) = 0$, then (2.25) becomes

$$\int_a^b \Big\{ -\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u - f \Big\} v\, dx = 0 \quad \text{for any } v \text{ with } v(a) = v(b) = 0\ ,$$

this implies

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u = f, \quad a < x < b. \tag{2.26}$$

Substituting this into (2.25), we have

$$-c(a)u_x(a)v(a) = -\alpha\, v(a) \quad \text{for any } \ v \text{ with } v(b) = 0\,,$$

this indicates that u also satisfies the condition

$$c(x)\frac{du}{dx}\Big|_{x=a} = \alpha\ . \tag{2.27}$$

(2.26) and (2.27) tell us that u is a solution of the boundary value problem (2.21)-(2.22). $\sharp$

Equivalence between boundary value and minimization problem

Now we investigate the relation between the boundary value problem (2.21)-(2.22) and the following potential energy functional

$$P(u) = \frac{1}{2}\int_b^a \left(c(x)\big(\frac{du}{dx}\big)^2 + q(x)u^2 \right) dx - \Big\{ \int_a^b f(x)\, u(x)\, dx - \alpha\, v(a) \Big\},$$

we are going to verify the following relations:

The function u that minimizes $P(v)$ over all v satisfying $v(b) = \beta$ must be the solution of the system (2.21)-(2.22), that is, it satisfies the differential equation

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) + q(x)u = f(x), \quad a < x < b$$

with the boundary conditions

$$c\frac{du}{dx}\Big|_{x=a} = \alpha\,, \quad u(b) = \beta\,.$$

The converse is also true.

The proof of this equivalence is basically the same as we did in the last subsection. So we omit it here.

2.3.3 Complementary minimum principle for the internal force

We know from the previous discussions that the displacement u of an elastic bar satisfies the boundary value problem:

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) = f(x)\,, \quad 0 < x < 1 \tag{2.28}$$

and the boundary conditions

$$u(0) = 0, \quad c(x)\frac{du}{dx}\Big|_{x=1} = 0\ .$$

Moreover, u also solves the equivalent variational problem

$$\int_0^1 c(x)\frac{du}{dx}\frac{dv}{dx}dx = \int_0^1 f(x)v\ dx \quad \forall\ v \text{ with } v(0) = 0$$

and minimizes the potential energy functional

$$P(u) = \frac{1}{2}\int_0^1 c(x)\Big(\frac{du}{dx}\Big)^2 dx - \int_0^1 f(x)u\ dx\ .$$

Below, we shall discuss some similar results for the internal force $w(x) = c(x)\frac{du}{dx}$. From (2.28) we know w satisfies

$$-\frac{dw}{dx} = f(x), \quad 0 < x < 1 \tag{2.29}$$

and the boundary condition

$$w(1) = 0\,. \tag{2.30}$$

Corresponding to the problem (2.29)-(2.30), we define a new energy functional

$$Q(w) = \frac{1}{2}\int_0^1 \frac{1}{c(x)}w^2(x)dx$$

and consider the minimization problem

$$\min_w Q(w) \text{ subject to } -\frac{dw}{dx} = f(x), \quad w(1) = 0 \tag{2.31}$$

This is a constrained optimization problem.

To transform the constrained problem into a unconstrained problem, we introduce a *Lagrangian functional*

$$\begin{aligned} L(u,w) &= Q(w) + \int_0^1 u\big(\frac{dw}{dx} + f\big)dx \\ &= \frac{1}{2}\int_0^1 \frac{1}{c(x)}w^2(x)dx + \int_0^1 u\big(\frac{dw}{dx} + f\big)dx\ , \end{aligned}$$

where u is called a *Lagrange multiplier.* With the help of Lagrange multiplier, we can convert the constrained variation problem to a non-constrained variation problem. Namely, if w^* satisfies (2.31), then it is also a critical solution of L, i.e.

$$\frac{\partial L}{\partial u} = 0, \frac{\partial L}{\partial w} = 0.$$

Before showing this, let us define $\partial L/\partial u$ as

$$\left(\frac{\partial L}{\partial u}(u,w), \eta\right) := \left.\frac{d}{dt}\right|_{t=0} L(u+t\eta, w)$$

for all η with $\eta(0) = 0$. We find

$$L(u+t\eta, w) = Q(w) + (u+t\eta, w_x + f)$$

Therefore,

$$\frac{d}{dt}L(u+t\eta, w) = (\eta, w_x + f).$$

Hence,

$$\frac{\partial L}{\partial u} = w_x + f.$$

We find that the condition $\partial L/\partial u = 0$ recovers the constraint condition.

The partial derivative $\partial L/\partial w$ can be computed similarly. For any v with $v(1) = 0$, we have

$$\begin{aligned} L(u, w+tv) &= Q(w+tv) + (u, (w+tv)_x + f) \\ &= Q(w) + t(\frac{w}{c}, v) + t^2 Q(v) - t(u_x, v) + (u, w_x + f) \end{aligned}$$

Here, I have used $(uv)_{x=0}^{x=1} = 0$ due to $u(0) = 0$ and $v(1) = 0$. As we differentiate $L(u, t+w+tv)$ in t, we get

$$\left(\frac{\partial L}{\partial w}(u,w), v\right) := \left.\frac{d}{dt}\right|_{t=0} L(u, w+tv) = \left(\frac{w}{c} - u_x, v\right).$$

Hence,

$$\frac{\partial L}{\partial w}(u,w) = \left(\frac{w}{c} - u_x, v\right).$$

The condition

$$\frac{\partial L}{\partial w}(u,w) = 0$$

gives the Hook's law: $w = cu_x$.

(1) $\begin{cases} -\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) = f(x)\,, \quad 0<x<1 \\ u(0)=0, \quad c(x)\frac{du}{dx}\Big\vert_{x=1} = 0 \end{cases}$	(2) $\begin{cases} w - c(x)\frac{du}{dx} = 0 \\ -\frac{dw}{dx} + f = 0 \\ u(0)=0,\ w(1)=0. \end{cases}$ $L(u,w) := \frac{1}{2}\int_0^1 \frac{1}{c(x)} w^2(x)dx + \int_0^1 u\big(\frac{dw}{dx}+f\big)dx$
(3) $\text{Max}_u(-P(u))$ with $u(0)=0$	(4) $\text{Min}_w Q(w)$ subject to $-\frac{dw}{dx} = f,\ w(1)=0$
(5) $\text{Max}_u\ \text{Min}_w L(u,w), u(0)=0, w(1)=0$	(6) $\text{Min}_w\ \text{Max}_u L(u,w), u(0)=0, w(1)=0$

We shall establish the equivalence table.

We have seen (1)⇔(2), (1)⇔(3).

Now, we show (3)⇔(5). This is equivalent to say that

$$-P(u) = \text{Min}_w L(u,w) \text{ subject to } u(0)=0, w(1)=0$$

For each fixed u with $u(0)=0$, if $\bar{w}(u)$ is a minimum, then it is a critical point of L with respect to w. That is

$$\frac{\partial L}{\partial w} = 0.$$

The partial derivative can be obtained from

$$\frac{d}{dt}\Big|_{t=0} L(u, w+tv) = (\frac{\partial L}{\partial w}, v).$$

for v with $v(1)=0$. This gives

$$\frac{\partial L}{\partial w} = \frac{w}{c} - u_x.$$

Thus, the critical point $\bar{w} = cu_x$. We plug it into L to get

$$\begin{aligned} L(u,\bar{w}(u)) &= \frac{1}{2}\int_0^1 \frac{1}{c(x)}\bar{w}^2(x)dx + \int_0^1 u\big(\frac{d\bar{w}}{dx}+f\big)dx \\ &= \frac{1}{2}\int_0^1 cu_x^2\,dx + \int_0^1 (-(u_x w) + fu)\,dx \\ &= \frac{1}{2}\int_0^1 cu_x^2\,dx + \int_0^1 (-(u_x cu_x) + fu)\,dx \\ &= -P(u) \end{aligned}$$

Here, we have used $u(0) = 0$ and $w(1) = 0$ in the integration by part. Using the technique of completing square, this critical point must be a minimum. I leave you to prove it.

Next, we show (4)⇔(6). This is due to

$$\text{Max}_{\substack{u \\ u(0)=0}} \; Q(w) + (u, w_x + f) = \begin{cases} Q(w) & \text{if } w_x + f = 0 \\ +\infty & \text{otherwise} \end{cases}$$

For (2)⇒(4), it is the same as we had before for the spring-mass case. We leave this proof to students.

To show (5)⇔(6), we use an intermediate step: we show that

if w^* and u^* are functions such that $w^*(1) = 0$, $u^*(0) = 0$, and are the minimizer and the maximizer of the following problems:

$$L(u^*, w^*) = \min_w L(u^*, w)\,, \quad L(u^*, w^*) = \max_u L(u, w^*)\,, \tag{2.32}$$

then (u^*, w^*) satisfies

$L(u^*, w^*) = \mathbf{Max}_u\ \mathbf{Min}_w L(u, w) = \mathbf{Min}_w\ \mathbf{Max}_u L(u, w)$ **with** $u(0) = 0, w(1) = 0.$

Let u, w satisfy $u(0) = 0, w(1) = 0$ and (u^*, w^*) satisfies (2.32), We have

$$L(u^*, w^*) = \min_w L(u^*, w) \le \max_u(\min_w L(u, w)).$$

On the other hand,

$$L(u^*, w^*) = \max_u L(u, w^*) \ge \min_w \max_u L(u, w)$$

Thus, we get

$$\min_w \max_u L(u, w) \le \max_u(\min_w L(u, w).$$

The reverse inequality is easy to show. Because for any u_1,

$$L(u_1, w) \le \max_u L(u, w)$$

We take minimum over w to get

$$\min_w L(u_1, w) \le \min_w \max_u L(u, w)$$

The left-hand side nw is a constant. We then take maximum over u_1, then we get

$$\max_{u_1}\min_{w} L(u_1,w) \le \min_{w}\max_{u} L(u,w).$$

Conversely, if (u^*,w^*) is the saddle point of L such that

$$L(u^*,w^*) = \max_{u}\min_{w} L(u,w) = \min_{w}\max_{u} L(u,w)$$

we show that it satisfies (2.32). We have

$$L(u^*,w^*) = \max_{u}\min_{w} L(u,w) \le \min_{w} L(u^*,w) \le \min_{w}\max_{u} L(u,w) = L(u^*,w^*)$$

Similarly,

Finally, we are going to show that (u^*,w^*) satisfies (2.32) then w^* and u^* satisfy (2), i.e.

$$w^* = c(x)u^*_x\,, \quad -w^*_x = f\,. \tag{2.33}$$

And the converse is also true.

We first show that if w^* and u^* are functions such that $w^*(1)=0$, $u^*(0)=0$, and are the solutions to (2.33), then they are also the solutions to the optimization problems in (2.32). In fact, for any v, we have

$$\begin{aligned}
L(u^*,v) - L(u^*,w^*) &= \frac{1}{2}(c^{-1}v,v) + (u^*,v_x+f) - \frac{1}{2}(c^{-1}w^*,w^*) - (u^*,w^*_x+f) \\
&= \frac{1}{2}(c^{-1}(v-w^*),v-w^*) + (c^{-1}(v-w^*),w^*) + (u^*,(v-w^*)_x) \\
&= \frac{1}{2}(c^{-1}(v-w^*),v-w^*) + (v-w^*,u^*_x) + (u^*,(v-w^*)_x) \\
&= \frac{1}{2}(c^{-1}(v-w^*),v-w^*) \ge 0\,.
\end{aligned}$$

On the other hand, for any u such that $u(0)=0$, we have

$$L(u,w^*) = \frac{1}{2}(c^{-1}w^*,w^*) + (u,w^*_x+f) = \frac{1}{2}(c^{-1}w^*,w^*)\,,$$

so we know that $L(u,w^*)$ is constant with respect to u. This proves both w^* and u^* are the desired solutions to the optimization problems in (2.32).

Next we show that if w^* and u^* are functions such that $w^*(1)=0$, $u(0)=0$, and are the minimizer and the maximizer of the optimization problems in (2.32), then they are also the solutions to (2.33).

To do so, we define (for simplicity we drop the index $*$ in u^* and w^*)

$$F(t) = L(u, w + tv) \quad \text{for any } v \ .$$

As w is the minimizer of $L(u, w)$, we know

$$F(t) = L(u, w + tv) \geq L(u, w) = F(0) \quad \forall\, t \in \mathbb{R}^1,$$

so $t = 0$ is a minimizer for $F(t)$, thus

$$F'(0) = 0 \ .$$

Now by definition,

$$\begin{aligned} F(t) - F(0) =& L(u, w + tv) - L(u, w) \\ =& \frac{1}{2}\int_0^1 \frac{1}{c}(w + tv)^2 dx + \int_0^1 u\big(\frac{d(w + tv)}{dx} + f\big) dx \\ & - \Big\{\frac{1}{2}\int_0^1 \frac{1}{c} w^2 dx + \int_0^1 u\big(\frac{dw}{dx} + f\big)\Big\} dx \ , \end{aligned}$$

or

$$F(t) - F(0) = t\int_0^1 (\frac{1}{c(x)} wv + u\frac{dv}{dx}) dx + \frac{1}{2}t^2 \int_0^1 \frac{1}{c(x)} v^2 dx,$$

therefore

$$0 = F'(0) = \int_0^1 \Big(\frac{1}{c(x)} wv + u\frac{dv}{dx}\Big) dx \ .$$

Integration by parts gives

$$\int_0^1 \Big(\frac{1}{c(x)} w - \frac{du}{dx}\Big) v dx + uv(x)\Big|_{x=0}^{x=1} = 0 \quad \forall v$$

which implies

$$\frac{1}{c(x)} w = \frac{du}{dx} \quad \text{or} \quad w = c(x)\frac{du}{dx} \ . \tag{2.34}$$

Thus we get back the original physical law $w = c(x)\frac{du}{dx}$.

On the other hand, we can find the maximizer of u, that gives the condition:

$$-\frac{dw}{dx} = f(x) \, . \tag{2.35}$$

This proves the desired results.

From the equations (2.34)-(2.35, we see that u satisfies

$$-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big) = f(x) \, , \quad 0 < x < 1.$$

This means that the Lagrange multiplier u is actually the displacement function.

Remark. Read pp. 166-172.

Homeworks.

- pp. 180, 3.2.2
- pp. 180, 3.2.3
- pp. 180, 3.2.4
- pp. 180, 3.2.7
- pp. 180, 3.2.8
- pp. 180, 3.2.9

3 Numerical methods for differential equations

In this section, we shall introduce several popular numerical methods for solving differential equations. The first class is the *finite difference methods.* The second class is the *finite element method.*

3.1 Finite Difference Approximation

Consider the Sturm-Liouville equation

$$-(c(x)u')' + q(x)u = f, x \in (a,b), \tag{3.1}$$

$$u(a) = 0,\ u(b) = 0. \tag{3.2}$$

The finite difference method treat the approximate solution defined on grid points: $x_0 < \cdots < x_{n+1}$. For simplicity, we choose uniform grids. That is, we define the mesh size $h = (b-a)/n+1$ and defind $x_i = a + ih$, $i = 0, ..., n+1$. The derivative of u can be approximated by

- forward differencing: $u'(x) = (u(x+h) - u(x))/h + O(h)$
- backward differencing: $u'(x) = (u(x) - u(x-h))/h + O(h)$
- centered differencing: $u'(x) = (u(x+h) - u(x-h))/2h + O(h^2)$

These can be proved by Taylor expansion. With this, we approximate (3.1) first by the centered differencing at x_i by using data at $x_{i+1/2}$ and $x_{i-1/2}$: If u is a solution of (3.1), then

$$\frac{1}{h}\Big(c(x_{i-1/2})u'(x_{i-1/2}) - c(x_{i+1/2})u'(x_{i+1/2})\Big) + q(x_i)u(x_i) = f(x_i) + O(h^2).$$

Here $x_{i+1/2} := a + (i + 1/2)h$. Next, we approximate

$$u'(x_{i+1/2}) = \frac{1}{h}\Big(u(x_{i+1}) - u(x_i)\Big) + O(h^2).$$

Then the solution of (3.1) can be approximated by

$$\frac{1}{h^2}\Big(c(x_{i-1/2})(u(x_i) - u(x_{i-1})) - c(x_{i+1/2})(u(x_{i+1}) - u(x_i))\Big) + q(x_i)u(x_i) = f(x_i) + O(h^2). \tag{3.3}$$

The finite difference approximation to (3.1) and (3.2) is use grid function U_i to approximate the solution $u(x_i)$, $i = 0, ..., n+1$ with U satisfying

$$\frac{1}{h^2}\Big(c(x_{i-1/2})(U_i - U_{i-1})) - c(x_{i+1/2})(U_{i+1} - U_i)\Big) + q(x_i)U_i = f(x_i), i = 1, ..., n \quad (3.4)$$

$$U_0 = 0,\ u_{n+1} = 0. \quad (3.5)$$

There are n equations for the n unknowns $U_1, ..., U_n$. As we have seen in the spring-mass system, this equation has the form

$$KU + QU = F$$

where K is the stiffness matrix and Q is a diagonal matrix. In principle, we can solve this linear equation by Gaussian elimination.

Homeworks.

- Suppose we discretize the domain $[0, 1]$ by by $\{x_i\}_{i=0}^n$. They are not necessary uniform. Let $h_i = x_i - x_{i-1}$. Given a function $u \in C^4$ and let $u_i = u(x_i)$. We can approximate u by the following finite difference method:

  $$u''(x_i) \sim \frac{\frac{u_{i+1}-u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}}{\frac{h_i + h_{i+1}}{2}}$$

  What is the error of this finite difference approximation? (Exress in terms of h_i and h_{i+1}.) What is the error if the grid is uniform?

3.2 Finite Element Methods—Galerkin's approach

In the above finite difference approach, we have assumed u and the coefficient c are smooth. In some applications, the coefficient may not be smooth and the corresponding solution may not be in C^2. In this case, a weak formulation is favored. It required less regularity of u and the coefficients.

To solve a given differential equation, the Galerkin method starts with the variational formulation of the differential equation, then construct the finite dimensional space, and finally solve the discrete problem.

3.2.1 Weak formulation

Let us take the following two-point boundary value problem as an example:

$$\begin{cases} -\big(c(x)u'(x)\big)' + q(x)u(x) = f(x), \qquad a < x < b\,, \\ c(x)u'(x)|_{x=a} = \alpha\ , \quad u(b) = \beta \end{cases} \tag{3.6}$$

<u>**Derive the variational problem**</u>. Multiply (3.6) by a test function $v(x)$ satisfying the conditions $v(b) = 0$ (**why ?**), then integrate over (a, b) to obtain the following variational formulation of the system (3.6):

$$\begin{cases} \text{Find } u(x) \text{ such that } u(b) = \beta \text{ and} \\ a(u, v) = g(v) \quad \forall \;\; v(x) \text{ satisfying } v(b) = 0 \end{cases} \tag{3.7}$$

where $a(u, v)$ and $g(v)$ are two integrals given by

$$\begin{aligned} a(u, v) &= \int_a^b \big(a(x)u'\,v' + c(x)uv\big)dx, \\ g(v) &= \int_a^b f(x)\,v(x)dx - \alpha\,v(a). \end{aligned}$$

3.2.2 Approximation of functions—Trial functions

We want to approximate solution u by some basis functions called trial functions. That is,

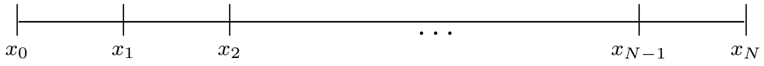
$$u \approx u_h := \sum_i u_i \phi_i(x).$$

The functions ϕ_i are the basis functions. They are finite many. Then the problem of differential equation is reduced to find u_i, a finite dimensional problem.

A simple basis functions is the follows.

We divide the interval $[a, b]$ into N subintervals using the grid points

$$T^h: \;\; a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b\ .$$



The points x_0, x_1, $\cdots$, x_N are called the nodal points or grid points. We shall denote the length of the subinterval $[x_{i-1}, x_i]$ by h_i.

Then we can construct one basis function $\phi_i(x)$ at each grid point x_i , $i = 0, 1, 2, \cdots, N$ such that

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1\,, & j = i \\ 0\,, & j \neq i \end{cases}. \tag{3.27}$$

This generates the finite element space

$$V_h = \text{ span } \{\phi_0, \phi_1, \cdots, \phi_N\}.$$

There are many choices for the piecewise polynomial basis functions ϕ_k which satisfy the conditions (3.27). Next, we discuss the simplest case with piecewise linear basis functions.

Piecewise linear finite element spaces. At each interior grid point $x = x_i$, $i = 1, 2, \cdots, N-1$, we construct a basis function as follows:

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i}\,, & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{h_{i+1}}\,, & x \in [x_i, x_{i+1}] \\ \quad 0\,, & x \notin [x_{i-i}, x_{i+1}] \end{cases}$$

At the boundary grid point $x = x_0$:

$$\phi_0(x) = \begin{cases} \frac{x_1 - x}{h_1}\,, & x \in [x_0, x_1] \\ \quad 0\,, & x \not\in [x_0, x_1] \end{cases},$$

while at the boundary point $x = x_N$:

$$\phi_N(x) = \begin{cases} \frac{x - x_{N-1}}{h_N}\,, & x \in [x_{N-1}, x_N] \\ \quad 0\,, & x \not\in [x_{N-1}, x_N] \end{cases}.$$

Clearly, it is easy to see that the above constructed basis functions $\{\phi_i\}_{i=0}^N$ satisfy the conditions (3.27). And these basis functions define a piecewise linear finite element space:

$$V_h = \text{span}\,\{\phi_0, \phi_1, \cdots, \phi_N\}\,.$$

And for any $v \in V_h$, we can prove that

$$v(x) = \sum_{i=0}^{N} v(x_i)\phi_i(x) \quad \text{for} \;\; a \leq x \leq b.$$

3.2.3 Projection of the equation—test functions

Galerkin's approach . Suppose there are a set of basis functions given by

$$V_h = \Big\{\phi_0(x), \phi_1(x), \cdots, \phi_N(x)\Big\},$$

then we can approximate the variational formulation (3.7) as follows:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that } u_h(b) = \beta \text{ and} \\ a(u_h, v_h) = g(v_h) \quad \forall\, u_h \in V_h^0 \end{cases} \tag{3.28}$$

where

$$V_h^0 = \big\{v_h \in V_h;\ v_h(b) = 0\big\}.$$

These v_h rae called test functions. In the Galerkin's approach, the trial functions and the test functions are identical. This is natural because it will result in the number of unknowns equals the number of equations.

System of linear algebraic equations. Let us see how to solve the Galerkin problem (3.28). For convenience, we will construct the basis functions such that

$$\phi_0(b) = \phi_1(b) = \cdots = \phi_{N-1}(b) = 0\,, \quad \phi_N(b) = 1\,.$$

As $u_h \in V_h$, we can express it as follows:

$$u_h(x) = \sum_{i=0}^{N} u_i\phi_i(x) = \beta\,\phi_N(x) + \sum_{i=0}^{N-1} u_i\phi_i(x) \quad \text{with} \quad u_i \in \mathbb{R}^1\,.$$

Using (3.28), we know

$$a(u_h, \phi_j) = g(\phi_j), \quad j = 0, 1, 2, \cdots, N-1 \tag{3.29}$$

Substituting u_h into (3.29) gives

$$\sum_{i=0}^{N-1} u_i a(\phi_i, \phi_j) = g(\phi_j) - \beta\, a(\phi_N, \phi_j)\,, \quad j = 0, 1, 2, \cdots, N-1\,,$$

which can be written as

$$\mathcal{A}\, u = b \tag{3.30}$$

where

$$u = \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-1} \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0,N-1} \\ a_{10} & a_{11} & \cdots & a_{1,N-1} \\ \cdots & \cdots & \cdots & \cdots \\ a_{N-1,0} & a_{N-1,1} & \cdots & a_{N-1,N-1} \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{N-1} \end{pmatrix}$$

where a_{ij} and b_j are given by

$$a_{ij} = a(\phi_j, \phi_i), \quad b_j = g(\phi_j) - \beta\, a(\phi_N, \phi_j)\,.$$

The system of algebraic equations (3.30) can be solved either by the Gauss elimination or by some iterative methods, such as Jacobi method and Gauss-Seidel method.

Remarks. Read 428-433.

Homeworks.

- Find the matrix $\mathcal{A}$ for the equation $-u'' = f$ with ϕ_i being the linear nodal functions at grids x_i.
- Verify that the matrix $\mathcal{A}$ in (3.30) is symmetric positive definite.
- Derive the finite element method for solving the following boundary value problems:

$$\begin{cases} -\big(a(x)u'(x)\big)' + c(x)u(x) = f(x), & 1 < x < 3\,, \\ u(1) = 1\,, \quad u(3) = -1 \end{cases} \tag{3.31}$$

$$\begin{cases} -\big(a(x)u'(x)\big)' + c(x)u(x) = f(x), & -1 < x < 2\,, \\ u(-1) = 1\,, \quad u'(2) = -1 \end{cases} \tag{3.32}$$

$$\begin{cases} -\big(a(x)u'(x)\big)' + c(x)u(x) = f(x), & -2 < x < 1\,, \\ a(-2)u'(-2) = 10\,, \quad u(1) = -10 \end{cases} \tag{3.33}$$

4 Vector calculus in Euclidean space $\mathbb{R}^3$ with Applications

In this section, I shall first review some basic concepts about the vector calculus, which are widely used in engineering, physics and other areas requiring mathematics. Then I shall derive some important models in two or three dimensions including the heat conduction model, the potential flow, electrostatics, and the Maxwell equations for the elecromagnetism.

Consider the Euclidean space $\mathbb{R}^n$, with a rectangular coordinate system formed by the x_1-, x_2-, $\cdots$ and x_n-coordinate axes. The vector calculus is about some basic operations of a vector-valued function in $\mathbb{R}^n$. We shall consider $n = 2$ or $n = 3$.

4.1 Review of Vector Calculus

4.1.1 Some differential operators for scalar and vector fields

Scalar field and vector field

- **scalar field**. A scalar field u is a real-valued function in $\mathbb{R}^n$. Concrete examples are gravitational potential $u(\mathbf{x}) = -1/r$, where $r = |x|$.
- **vector field**. A vector field $\mathbf{v}$ is a vector-valued function $\mathbf{v}(\mathbf{x})$ in $\mathbb{R}^n$ with n components, we shall write $\mathbf{v}(\mathbf{x})$ as

$$\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), v_2(\mathbf{x}), \cdots, v_n(\mathbf{x}))^T \quad \text{with} \quad \mathbf{x} = (x_1, x_2, \cdots, x_n)^T.$$

  I list some concrete examples of vecor fields: wind field, heat flux, electrical field, magnetic field, gravitational force field, displacement vector field of an elastic material.

Some differential operators

- **Gradient.** Given a scalar function $u(\mathbf{x})$ in $\mathbb{R}^n$, define

$$\text{grad } u = \nabla u := \Big(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \cdots, \frac{\partial u}{\partial x_n}\Big)^T$$

- **<u>Curl.</u>** For a vector field in $\mathbb{R}^3$, define

$$\text{curl } \mathbf{v} \equiv \nabla \times \mathbf{v} = \begin{vmatrix} i & j & k \\ \partial_{x_1} & \partial_{x_2} & \partial_{x_3} \\ v_1 & v_2 & v_3 \end{vmatrix} = \begin{pmatrix} \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1} \\ \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \end{pmatrix}$$

  In the two dimensional case, the vector field is expressed as $\mathbf{v}(x, y) = (0, 0, \psi(x, y))$ for some scalar field $\psi(x, y)$, the curl of $\mathbf{v}$ becomes

$$\nabla \times \mathbf{v} = (\partial_y \psi, -\partial_x \psi, 0)^T.$$

  Sometimes, we denote

$$\mathbf{curl}\, \psi = \nabla^{\perp} \psi \ = (\frac{\partial \psi}{\partial y}, -\frac{\partial \psi}{\partial x})^T \, .$$

  Notice that ∇ and $\nabla^{\perp}$ are orthogonal in the sense that

$$\nabla \psi \cdot \nabla^{\perp} \psi = 0 \, .$$

  One can also define another curl operation in two dimensions. For any vector-valued function $\mathbf{v}(x, y)$, we define

$$\text{curl}\, \mathbf{v} = \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \, .$$

  This maps a vector-valued function into a scalar function.

- **<u>Divergence.</u>** Given a vector field $\mathbf{v} in \mathbb{R}^n$, define

$$\text{div } \mathbf{v} = \equiv \nabla \cdot \mathbf{v} := \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \cdots + \frac{\partial v_n}{\partial x_n}$$

- **<u>Laplacian.</u>** For a scalar field u, define

$$\triangle u = \text{div grad } u = \nabla \cdot \nabla u = \nabla^2 u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} \, .$$

  The operator $\triangle$ is called the **Laplacian operator**.

Examples

1. Find the gradient of u for $u(x_1, x_2, x_3) = -1/r$, where $r = \sqrt{x_1^2 + x_2^2 + x_3^2}$.
2. In two dimensions, find the gradient of $\theta := \tan^{-1}(y, x)$.
3. In two dimensions, let the vector field $\mathbf{v} = (-y, x)$, find its curl. This is a rigid body rotation.
4. In two dimensions, find the divergence and curl of $\mathbf{v} = (-x, y)$. This velocity is called a jet.

Homeworks.

- In two dimensions, find the gradient of $u(x, y) = -\log r$, where $r = \sqrt{x^2 + y^2}$.
- Show that $\nabla \times \nabla u = 0$ for any scalar function u.
- Show that $\nabla \cdot (\nabla \times \mathbf{v}) = 0$ for any vector field $\mathbf{v}$ in $\mathbb{R}^3$.
- Verify the following relation
$$\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \Delta \mathbf{v}\,.$$
for any vector field $\mathbf{v}$ in $\mathbb{R}^3$.
- For a scalar field u and a vector field $\mathbf{v}$ in $\mathbb{R}^n$, show
$$\nabla \cdot (u\,\mathbf{v}) = \nabla u \cdot \mathbf{v} + u(\nabla \cdot \mathbf{v}).$$
- Let $\mathbf{v} = (x, y, z)/r^3$, where $r = \sqrt{x^2 + y^2 + z^2}$. Let Σ be a closed surface enclose the origin. Find the integral $\int_\Sigma \mathbf{v} \cdot \mathbf{n}\, dS$.

4.1.2 Fundamental Theorems of Vector Calculus

Theorem 4.1. *Let u be a scalar field in $\mathbb{R}^n$. Then the path integral of ∇u equals $u(\mathbf{x}_1) - u(\mathbf{x}_0)$, where $\mathbf{x}_0$ and $\mathbf{x}_1$ are the initial and final positions of the path C:*

$$\int_C \nabla u \, d\mathbf{s} = u(\mathbf{x}_1) - u(\mathbf{x}_0).$$

Theorem 4.2 (Stokes Theorem)**.** *Let* $\mathbf{v}$ *be a vector field in* $\mathbb{R}^3$ *and* Σ *be a surface in* $\mathbb{R}^3$ *with boundary* C *which is a closed simple curve. Then*

$$\int_{\Sigma} \nabla \times \mathbf{v} \cdot n \, dS = \int_C \mathbf{v} \cdot d\mathbf{s}$$

Theorem 4.3 (Divergence Theorem)**.** *Let* $\mathbf{v}$ *be a vector field in* $\mathbb{R}^3$ *and* Ω *be a volume in* $\mathbb{R}^3$ *with boundary* Σ *which is a closed simple surface. Then*

$$\int_{\Omega} \nabla \cdot \mathbf{v} \, d\mathbf{x} = \int_{\Sigma} \mathbf{v} \cdot \mathbf{n} \, dS.$$

I shall give an intuitive proof of the divergence theorem. Let us consider $\Omega = (a,b) \times (c,d) \times (e,f)$, a box. The six faces of the box and their outer normals are

$$\begin{aligned}
\Sigma_1 &= \{a\} \times (c,d) \times (e,f), \text{outer normal: } (-1,0,0) \\
\Sigma_2 &= \{b\} \times (c,d) \times (e,f), \text{outer normal: } (1,0,0) \\
\Sigma_3 &= (a,b) \times \{c\} \times (e,f), \text{outer normal: } (0,-1,0) \\
\Sigma_4 &= (a,b) \times \{d\} \times (e,f), \text{outer normal: } (0,1,0) \\
\Sigma_5 &= (a,b) \times (c,d) \times \{e\}, \text{outer normal: } (0,0,-1) \\
\Sigma_6 &= (a,b) \times (c,d) \times \{f\}, \text{outer normal: } (0,0,1)
\end{aligned}$$

Suppose the vector field $\mathbf{v} = (P,Q,R)$. The flux passes through Σ_1 is

$$\int_{\Sigma_1} \mathbf{v} \cdot \mathbf{n} \, dS = \int_c^d \int_e^f -P(a,y,z) \, dy \, dz.$$

The flux passes through Σ_2 is

$$\int_{\Sigma_2} \mathbf{v} \cdot \mathbf{n} \, dS = \int_c^d \int_e^f P(b,y,z) \, dy \, dz.$$

Hence,

$$\int_{\Sigma_1 \cup \Sigma_2} \mathbf{v} \cdot \mathbf{n} \, dS = \int_c^d \int_e^f P(b,y,z) - P(a,y,z) \, dy \, dz = \int_c^d \int_e^f \int_a^b P_x(x,y,z) \, dx \, dy \, dz$$

Similarly, the flux passes through Σ_3 and Σ_4 are

$$\int_{\Sigma_3} \mathbf{v} \cdot \mathbf{n} \, dS = \int_a^b \int_e^f -Q(x,c,z) \, dx \, dz.$$

$$\int_{\Sigma_4} \mathbf{v} \cdot \mathbf{n} \, dS = \int_a^b \int_e^f Q(x,d,z) \, dx \, dz.$$

and we get

$$\int_{\Sigma_3\cup\Sigma_4} \mathbf{v}\cdot\mathbf{n}\,dS = \int_a^b\int_e^f Q(x,d,z)-Q(x,c,z)\,dx\,dz = \int_a^b\int_e^f\int_c^d Q_y(x,y,z)\,dy\,dx\,dz$$

Similarly, we get

$$\int_{\Sigma_5\cup\Sigma_6} \mathbf{v}\cdot\mathbf{n}\,dS = \int_a^b\int_c^d R(x,y,f)-R(x,y,e)\,dx\,dy = \int_a^b\int_c^d\int_e^f R_z(x,y,z)\,dz\,dy\,dx$$

We sum the integrals over these six faces, then we get the divergence theorem for the box case. For general domain, we cut the domain into small boxes, apply the divergence theorem on each small boxes. Notice that the surface integrals on two adjacent boxes are cancelled (the outer normals have opposite signs). The remaining surface integral is the surface of the domain Ω. This gives the divergence for general domain.

Remark. From the Stokes and divergence theorems, we can get the physical meaning of the curl and div as the follows.

- $\nabla\cdot\mathbf{v}(\mathbf{x}) = \lim_{|\Omega|\to 0}\frac{1}{|\Omega|}\int_\Sigma \mathbf{v}\cdot\mathbf{n}\,dS$, where Ω is a small volume containing $\mathbf{x}$ and $|\Omega|$ is its volume. The surface integral measure the **flux** flows outward from V through Σ.

- $[\nabla\times\mathbf{v}(x)]\cdot\mathbf{n} = \lim_{|\Sigma|\to 0}\frac{1}{\Sigma}\int_C \mathbf{v}\cdot d\mathbf{s}$, where Σ is a small piece of surface containing $\mathbf{x}$ and is perpendicular to $\mathbf{n}$. $|\Sigma|$ is the area of Σ and C is its boundary. The integral measure the **circulation** of $\mathbf{v}$ along a curve C. Thus, the direction of $\nabla\times\mathbf{v}$ is the direction with the strongest circulation, and its magnitude is this circulation.

We have seen that $\nabla\times\nabla u = 0$ for any scalar field u in $\mathbb{R}^3$. The converse is also true.

Theorem 4.4. *Let* $\mathbf{v}$ *be a vector field in* $\mathbb{R}^3$*. Then* $\mathbf{v} = \nabla u$ *for some scalar field* $\mathbf{u}$ *if and only if* $\nabla\times\mathbf{v} = 0$.

Sketch Proof. The function u is defined by

$$u(\mathbf{x}) = \int_0^{\mathbf{x}} \mathbf{v}\cdot d\mathbf{s}$$

where the above line integral is allowed to follow any path connecting 0 to $\mathbf{x}$. This line integral is independent of path. The reason is that any two such paths forms a surface in

$\mathbb{R}^3$. We can apply the Stokes theorem and using $\nabla \times \mathbf{v} = 0$ to see that it is independent of path. Once u is defined, it is not difficult to see that its gradient is $\mathbf{v}$.

Next, we have also seen that $\nabla \cdot (\nabla \times \mathbf{w}) = 0$ for any vector field $\mathbf{w}$ in $\mathbb{R}^3$. Its converse is also true in $\mathbb{R}^3$.

Theorem 4.5. *If a vector field $\mathbf{v}(\mathbf{x})$ in $\mathbb{R}^3$ is divergence-free, that is, $\nabla \cdot \mathbf{v} = 0$, then there exists a vector-valued function $\mathbf{w}(\mathbf{x})$ such that*

$$\mathbf{v}(\mathbf{x}) = \nabla \times \mathbf{w}(\mathbf{x})\,.$$

Sketch idea of proof. To find $\mathbf{w}$, we apply curl to $\mathbf{v} = \nabla \times \mathbf{w}$:

$$\nabla \times (\nabla \times \mathbf{w}) = \nabla \times \mathbf{v}.$$

We use the identity $\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \triangle \mathbf{v}$, and use $\nabla \cdot \mathbf{v} = 0$, then we get

$$-\triangle \mathbf{w} = \nabla \times \mathbf{v}.$$

We then solve this partial differential equation component by component. In the whole $\mathbb{R}^3$ space, this is not difficult to solve. Indeed, there is an exact solution formula for

$$-\triangle u = f.$$

It is

$$u(\mathbf{x}) = \int \frac{1}{4\pi|\mathbf{x} - \mathbf{y}|} f(\mathbf{y})\, d\mathbf{y}.$$

4.1.3 Green's Theorems

Finally, the theorem corresponding to the technique of integration-by-part is called the Green's theorem. which is frequently used in the variational formulations for various partial differential equations. Their proofs are based on the divergence theorem.

1. For two scalar functions u and v, we have

$$\int_\Omega u_{x_i}\, v\, d\mathbf{x} = -\int_\Omega u\, v_{x_i}\, d\mathbf{x} + \int_{\partial\Omega} u\, v\, n_i\, dS$$

where $\Omega \subset R^k$, $\mathbf{n} = (n_1, n_2, \cdots, n_k)^T$ is the unit outward normal direction to the boundary $\partial\Omega$ of Ω.

2. For a vector-valued function $\mathbf{u}$ and a scalar function v, we have

$$\int_\Omega (\nabla \cdot \mathbf{u})\, v \, d\mathbf{x} = -\int_\Omega (\mathbf{u} \cdot \nabla v) \, d\mathbf{x} + \int_{\partial\Omega} (\mathbf{u} \cdot \mathbf{n})\, v \, dS.$$

3. For two vector-valued functions $\mathbf{u}$ and $\mathbf{v}$, we have

$$\int_\Omega (\nabla \times \mathbf{u}) \cdot \ \mathbf{v} \, d\mathbf{x} = \int_\Omega \mathbf{u} \ \cdot \nabla \times \mathbf{v} \, d\mathbf{x} + \int_{\partial\Omega} (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{n} \, dS \, .$$

Proofs. We first prove 2. We use the vector field identity

$$\nabla \cdot (\mathbf{u}v) = \nabla \cdot \mathbf{u}v + \mathbf{u} \cdot \nabla v.$$

We integrate this identity over a domain Ω, then apply the divergence theorem.

Case 1 is a special case of case 2 with $\mathbf{u} = (0, .., u, 0..)$, only the ith component is nontrial.

For the third statement, we use the vector identity:

$$\nabla \cdot (\mathbf{u} \times \mathbf{v}) = (\nabla \times \mathbf{u}) \cdot \mathbf{v} - \mathbf{u} \cdot (\nabla \times \mathbf{v}).$$

We then integrate it over a domain and then apply the divergence theorem.

Homeworks.

- pp. 217, 3.4.8
- pp. 218, 3.4.15
- pp. 218, 3.4.16
- pp. 218, 3.4.17
- pp. 219, 3.4.33
- pp. 217, 3.4.13
- pp. 219, 3.4.34

4.1.4 Vector Calculus in special coordinate systems

In this and next subsections, we introduce three most frequently used coordinate systems.

In the standard rectangular coordinate system, each point is uniquely determined by three real numbers x, y and z, called the coordinates of the point, often denoted as the triple (x, y, z). Many partial differential equations can be conveniently written in this standard rectangular coordinate system.

For example, the Poisson equation in the orthogonal coordinate system takes the form

$$u_{xx} + u_{yy} + u_{zz} = f(x, y, z)$$

or equivalently

$$\Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f(x, y, z). \tag{4.1}$$

This Poisson equation will have a different form in a different coordinate system.

Cylindrical coordinate system (r, θ, z) Consider the following transformation

$$\begin{aligned} x &= r\cos\theta, \\ y &= r\sin\theta, \\ z &= z \end{aligned}$$

where r and θ change in the following ranges

$$0 \le \theta \le 2\pi, \quad 0 \le r < \infty.$$

Note that we always have

$$x^2 + y^2 = r^2\,.$$

Using this transformation, any point in the standard rectangular coordinate (x, y, z) can also be uniquely determined by the triple (r, θ, z), and the coordinate system (r, θ, z) is called *the cylindrical coordinate system.* In this system, the Poisson equation takes the following form

$$\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial w}{\partial r}) + \frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2} + \frac{\partial^2 w}{\partial z^2} = F(r, \theta, z) \tag{4.2}$$

where w and F are related to u and f in the equation (4.1) by the relations:

$$u(x, y, z) = u(r\,\cos\theta, r\,\sin\theta, z) \equiv w(r, \theta, z), \tag{4.3}$$

$$f(x, y, z) = f(r\,\cos\theta, r\,\sin\theta, z) \equiv F(r, \theta, z). \tag{4.4}$$

• To get some feelings about the above transformations, one may see what are the functions $w(r,\theta,z)$ when

$$u(x,y,z) = x + y + z, \quad u(x,y,z) = x^2 + y^2 + z^2 .$$

To derive the equation (4.2) from the equation (4.1), we need to use the following relations:

$$\begin{aligned} u_x &= w_r \frac{\partial r}{\partial x} + w_\theta \frac{\partial \theta}{\partial x}, \\ u_y &= w_r \frac{\partial r}{\partial y} + w_\theta \frac{\partial \theta}{\partial y}, \end{aligned}$$

then from these we can compute the second order derivatives. But for the above computings, we need to first find the derivatives $\frac{\partial r}{\partial x}$, $\frac{\partial \theta}{\partial x}$, $\frac{\partial r}{\partial y}$ and $\frac{\partial \theta}{\partial y}$. How to compute these partial derivatives ?

Note that the following relation is not always true:

$$\frac{\partial r}{\partial x} = \left(\frac{\partial x}{\partial r}\right)^{-1}.$$

This holds only for the case with one variable: $x = x(r)$. In fact, we have

$$1 = \frac{dx}{dx} = \frac{dx}{dr}\frac{dr}{dx}.$$

But for our current case, we always have the following Jacobi relation

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \Big[\frac{\partial(r,\theta)}{\partial(x,y)}\Big]^{-1}.$$

This comes by writting

$$x = \phi(r,\theta) = \phi(r(x,y),\theta(x,y)), \quad y = \psi(r,\theta) = \psi(r(x,y),\theta(x,y))$$

and directly checking

$$\frac{\partial(x,y)}{\partial(r,\theta)}\frac{\partial(r,\theta)}{\partial(x,y)} = \mathrm{I}.$$

Thus we have

$$\frac{\partial(r,\theta)}{\partial(x,y)} = \left(\frac{\partial(x,y)}{\partial(r,\theta)}\right)^{-1} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}^{-1} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\frac{\sin\theta}{r} & \frac{\cos\theta}{r} \end{pmatrix}.$$

One can also derive this formula directly. To do this, using

$$x = r\cos\theta, \quad y = r\sin\theta,$$

we obtain

$$\begin{aligned} 1 &= \frac{\partial x}{\partial x} = \frac{\partial r}{\partial x}\cos\theta + r(-\sin\theta)\frac{\partial\theta}{\partial x}\ , \\ 0 &= \frac{\partial y}{\partial x} = \frac{\partial r}{\partial x}\sin\theta + r\cos\theta\frac{\partial\theta}{\partial x}. \end{aligned}$$

From this we have

$$\frac{\partial r}{\partial x} = \frac{\begin{vmatrix} 1 & -r\sin\theta \\ 0 & r\cos\theta \end{vmatrix}}{\begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}} = \frac{r\cos\theta}{r} = \cos\theta\ ,$$

$$\frac{\partial\theta}{\partial x} = \frac{\begin{vmatrix} \cos\theta & 1 \\ \sin\theta & 0 \end{vmatrix}}{\begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}} = \frac{-\sin\theta}{r}\ .$$

Similarly, using the transformation

$$x = r\cos\theta, \quad y = r\sin\theta$$

again, we obtain

$$\begin{aligned} 0 &= \frac{\partial x}{\partial y} = \frac{\partial r}{\partial y}\cos\theta + r(-\sin\theta)\frac{\partial\theta}{\partial y}\ , \\ 1 &= \frac{\partial y}{\partial y} = \frac{\partial r}{\partial y}\sin\theta + r\cos\theta\frac{\partial\theta}{\partial y}. \end{aligned}$$

From this we have

$$\frac{\partial r}{\partial y} = \frac{\begin{vmatrix} 0 & -r\sin\theta \\ 1 & r\cos\theta \end{vmatrix}}{\begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}} = \frac{r\sin\theta}{r} = \sin\theta\ ,$$

$$\frac{\partial\theta}{\partial y} = \frac{\begin{vmatrix} \cos\theta & 0 \\ \sin\theta & 1 \end{vmatrix}}{\begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}} = \frac{\cos\theta}{r}\ .$$

Now, we can compute as follows:

$$\begin{array}{rcl} u_x & = & w_r \dfrac{\partial r}{\partial x} + w_\theta \dfrac{\partial \theta}{\partial x} = w_r \cos\theta - w_\theta \dfrac{\sin\theta}{r}, \\ u_y & = & w_r \dfrac{\partial r}{\partial y} + w_\theta \dfrac{\partial \theta}{\partial y} = w_r \sin\theta + w_\theta \dfrac{\cos\theta}{r}, \\ u_{xx} & = & (w_r)_x \cos\theta + (w_r)(\cos\theta)_x - (w_\theta)_x \dfrac{\sin\theta}{r} - w_\theta \left(\dfrac{\sin\theta}{r}\right)_x \\ \cdots & & \cdots \end{array}$$

Please complete the remaining computations to derive (4.2).

Spherical coordinate system (r, θ, φ) Consider the following transformation

$$\begin{array}{rcl} x & = & r\cos\theta\sin\varphi \ , \\ y & = & r\sin\theta\sin\varphi \ , \\ z & = & r\cos\varphi \end{array}$$

where r, θ and φ change in their ranges

$$0 \le r < \infty, \quad 0 \le \theta \le 2\pi \ , \quad 0 \le \varphi \le \pi \ .$$

Note that we always have

$$x^2 + y^2 + z^2 = r^2 \ .$$

Using this transformation, any point in the standard rectangular coordinate (x, y, z) can be also uniquely determined by the triple (r, θ, φ), and the coordinate system (r, θ, φ) is called *the spherical coordinate system.* In this system, the Poisson equation takes the following form

$$\frac{1}{r^2}\frac{\partial}{\partial r}\big(r^2\frac{\partial w}{\partial r}\big) + \frac{1}{r^2\sin^2\varphi}\frac{\partial^2 w}{\partial \theta^2} + \frac{1}{r^2\sin\varphi}\frac{\partial}{\partial\varphi}\big(\sin\varphi\frac{\partial w}{\partial\varphi}\big) = F(r, \theta, \varphi) \qquad (4.5)$$

where w and F are related to u and f in the equation (4.1) by the relations:

$$\begin{array}{rcl} u(x, y, z) & = & u(r\ \cos\theta\sin\varphi, r\ \sin\theta\sin\varphi, r\ \cos\varphi) \equiv w(r, \theta, \varphi), \\ f(x, y, z) & = & f(r\ \cos\theta\sin\varphi, r\ \sin\theta\sin\varphi, r\ \cos\varphi) \equiv F(r, \theta, \varphi). \end{array}$$

• To get some feelings about this transfrmation, one may find out the functions $w(r, \theta, \varphi)$ when $u(x, y, z)$ is of the following two simple functions:

$$u(x, y, z) = x + y + x, \quad u(x, y, z) = x^2 + y^2 + z^2 \ .$$

Similarly as in the cylindrical coordinate system, we may use the following relation to derive the equation (4.5):

$$\frac{\partial(x,y,z)}{\partial(r,\theta,\varphi)} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \varphi} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \varphi} \end{pmatrix} = \left(\frac{\partial(r,\theta,\varphi)}{\partial(x,y,z)}\right)^{-1}.$$

The Jacobi matrix on the left hand is easy to compute, so we can get the Jacobi matrix on the right.

Remark 4.1. *It would be an excellent* **test** *for students to derive the equation (4.5) from the equation (4.1).*

4.2 Modeling heat conduction

4.2.1 The heat equation

We want to model heat conduction process in a domain $D \in \mathbb{R}^3$. The conduction process is described by a temperature function $u(\mathbf{x}, t)$. The variation of temperature in space generate heat flux $\mathbf{q}$. It flows from high temperature to low temperature. Fourier gave an emperical law of the heat flux:

$$\mathbf{q} = -\kappa \nabla u, \tag{4.6}$$

where $\kappa > 0$ is called the heat conductivity. It is a material dependent parameter. This emperical law is called the Fourier law.

It is known the heat is indeed a form of energy and the energy density is linearly proportitional to temperature:

$$h = c_v u,$$

where c_v is called the specific heat. The heat conduction model is based on *conservation of energy* described below. Let us consider an arbitrary domain $\Omega \in \mathbb{R}^3$. The energy in Ω is $\int_\Omega c_v u(\mathbf{x}, t)\, d\mathbf{x}$. Its rate of change in t is

$$\frac{d}{dt}\int_\Omega c_v u(\mathbf{x}, t)\, d\mathbf{x}.$$

The increase of energy in Ω must be the same as the heat transported into Ω through the boundary $\partial\Omega$ per unit time. Its is given by the following surface integral:

$$\int_{\partial\Omega} \mathbf{q} \cdot (-\mathbf{n})\, dS.$$

Here, $\mathbf{n}$ is the unit outer normal of $\partial\Omega$, and $\mathbf{q}\cdot(-\mathbf{n})dS$ is the net heat transported into Ω through dS per unit time. Thus, the conservation of energy states that

$$\frac{d}{dt}\int_\Omega c_v u(\mathbf{x},t)\,d\mathbf{x} = \int_{\partial\Omega} \mathbf{q}\cdot(-\mathbf{n})\,dS.$$

The left-hand side equals

$$\frac{d}{dt}\int_\Omega c_v u(\mathbf{x},t)\,d\mathbf{x} = \int_\Omega c_v u_t(\mathbf{x},t)\,d\mathbf{x},$$

whereas the right-hand side, by apply the divergence theorem, is

$$\int_{\partial\Omega} \mathbf{q}\cdot(-\mathbf{n})\,dS = -\int_\Omega \nabla\cdot\mathbf{q}\,d\mathbf{x}$$

This is valid for any arbitrary domain Ω. Hence, we get

$$c_v u_t + \nabla\cdot\mathbf{q} = 0. \tag{4.7}$$

According to the Fourier law, $\mathbf{q} = -\kappa\nabla u$, we get

$$u_t = k\nabla^2 u = k\triangle u. \tag{4.8}$$

Here, $k = \kappa/c_v > 0$ is the temperature conductivity.

4.2.2 Boundary condition

There are three kinds of boundary conditions for heat conduction model:

- Dirichlet: $u(\mathbf{x},t) = g(\mathbf{x}),\ \mathbf{x}\in\partial D$. This means that the domain D is attached to an environment (a heat bed) whose temperature is given by $g(\mathbf{x})$ on $\partial\Omega$.

- Neumann: $\nabla u(\mathbf{x})\cdot\mathbf{n} = 0$, $\mathbf{x}\in\partial D$. This means that there is no heat flux on the boundary. In other word, the domain is insulated. We may also impose $\nabla u(\mathbf{x})\cdot\mathbf{n} = h(\mathbf{x})$, $\mathbf{x}\in\partial D$, for some function h. This means that we inject energy into Ω with rate h.

- Robin: $\nabla u(\mathbf{x})\cdot\mathbf{n} = \lambda(g(\mathbf{x}) - u(\mathbf{x},t))$. Here, $\lambda > 0$ is a constant. This means that the difference of u and its surrouding temperature g generates a heat flux with rate linearly proportitional to this difference. This is called the Newton's conduction law. The sign $\lambda > 0$ means that if $u(\mathbf{x},t)$ on the boundary is greater than its surroundary temperature $g(\mathbf{x})$, then the heat flux $\quad = -\kappa\nabla u$ projected onto the outer normal direction $\mathbf{n}$, which is $-\kappa\nabla u\cdot\mathbf{n}$ is positive. This means the heat flows outward. This is consistent with the second law of thermodynamics, the heat flows from high temperature to low temperature.

Initial Condition. At time $t = 0$, the distribution of temperature $u_0(\mathbf{x})$ is given. That is

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \mathbf{x} \in D. \tag{4.9}$$

This is called the initial condition.

Initial boundary value problem. The mathematical problem for heat conduction is to find solution $u(\mathbf{x}, t)$ which satisfis (4.8), one of the above boundary condition and the initial condition (4.9).

4.3 Modeling electrostatics

4.3.1 The integral and differential forms of electrostatics

Electrostatics study the electric field in an environment containing conductors and charges. There are two basic physical laws:

- Gauss's law: the total electric flux over a closed surface is proportional to the total enclosed chages. In mathematical form:

$$\int_{\partial\Omega} \mathbf{E} \cdot \mathbf{n}\, dS = \frac{1}{\epsilon_0} \int_{\Omega} \rho(\mathbf{x})\, d\mathbf{x}. \tag{4.10}$$

  where ϵ_0 is a constant.

- The line integral of $\mathbf{E}$ over any closed loop is zero.

$$\int_C \mathbf{E} \cdot d\mathbf{s} = 0. \tag{4.11}$$

  This is a version of Farady's law.

These two equations are called the integral form of the electrostatics.

Using the divergence theorem and the Stokes theorem, we can get the differential form of the eletrostatics.

$$\epsilon_0 \nabla \cdot \mathbf{E} = \rho, \tag{4.12}$$

and

$$\nabla \times \mathbf{E} = 0. \tag{4.13}$$

The second equation implies that there exists a potential u such that $\mathbf{E} = \nabla u$. Plug this into the first equation. We get an equation for u:

$$\triangle u(\mathbf{x}) = f(\mathbf{x}). \tag{4.14}$$

where $f = \rho/\epsilon_0$. This is called the Poisson equation.

In the region where there is no charge, $f(\mathbf{x}) = 0$ in that region. In this case, the equation becomes

$$\triangle u(\mathbf{x}) = 0.$$

This is called the Laplace equation. This problem appears when the space contains conductors, and there is no charges outside conductor surfaces.

4.3.2 Boundary conditions

If the boundary is a perfect conductor, the tangent component of $\mathbf{E}$ must be zero. Otherwise the nonzero tangential electric field will drive charge motion on conductor, which violates the assumption of statics (no charge motion). In mathematical formulation, it reads

$$\mathbf{E} \cdot \mathbf{t} = 0 \text{ on surfaces of conductors.} \tag{4.15}$$

In terms of potential u, we get

$$\nabla u \cdot \mathbf{t} = 0 \text{ on surfaces of conductors.}$$

This leads to

$$u(\mathbf{x}) = const. \text{ on surfaces of conductors.}$$

If we have two conductors, we need to impose two constants for potential on the two surfaces. This kind of boundary condition is the Dirichlet boundary condition.

More general Dirichlet boundary condition reads

$$u(\mathbf{x}) = g(\mathbf{x}) \text{ for } \mathbf{x} \in \partial D. \tag{4.16}$$

4.4 *Modeling Fluid Flows

4.4.1 Flux and Continuity equation

Let consider a gas flow in three dimensions. Let $\rho(\mathbf{x})$ and $\mathbf{v}(\mathbf{x})$ be density and velocity of this flow at $\mathbf{x}$. Here, I mean "at $\mathbf{x}$" means the averages of the quantity in a small vicinity of $\mathbf{x}$. Let us consider a region $\Omega \in \mathbb{R}^3$. We want to study the variation of mass in Ω per unit time. That is

$$\frac{d}{dt}\int_\Omega \rho(\mathbf{x},t)\,d\mathbf{x}.$$

The mass increases in Ω must come from those flow into Ω throught its boundary $\partial\Omega$ in a unit small time Δt. To see how much amount of mass flows in, we consider a small piece of surface element dS with unit outer normal $\mathbf{n}$. Consider the parallel volume with base dS and $\mathbf{v}\cdot\Delta t$ as its one side. The gas in this parallel volume will flow into Ω in Δt through dS. And only this gas flows into Ω through dS in this period Δt. The volume of this parallel volume is $\mathbf{v}\Delta t\cdot\mathbf{n}dS$. Thhe mass in this vlume is $\rho\mathbf{v}\Delta t\cdot\mathbf{n}dS$. Thus, the total amount of masses flow into Ω in Δt is

$$-\int_{\partial\Omega}\rho\mathbf{v}\Delta t\cdot\mathbf{n}\,dS$$

The minus sign indicates the amount of gas *flows in* and notice that $\mathbf{n}$ is the outer normal. The quantity $\rho\mathbf{v}$ is called the *mass flux*. Thus, the amount of gas flows into Ω per unit time is

$$-\int_{\partial\Omega}\rho\mathbf{v}\cdot\mathbf{n}\,dS$$

Thus, we have

$$\frac{d}{dt}\int_\Omega \rho(\mathbf{x},t)\,d\mathbf{x} = -\int_{\partial\Omega}\rho\mathbf{v}\cdot\mathbf{n}\,dS$$

The left-hand side equals

$$\frac{d}{dt}\int_\Omega \rho(\mathbf{x},t)\,d\mathbf{x} = \int_\Omega \frac{\partial}{\partial t}\rho(\mathbf{x},t)\,d\mathbf{x}$$

By the divergence theorem, the right-hand side is

$$-\int_{\partial\Omega}\rho\mathbf{v}\cdot\mathbf{n}\,dS = -\int_\Omega \nabla\cdot(\rho\mathbf{v})\,d\mathbf{x}.$$

Thus, we get

$$\int_\Omega \rho_t + \nabla\cdot(\rho\mathbf{v})\,d\mathbf{x} = 0.$$

This is valid for any domain Ω. Thus, we conclude

$$\rho_t + \nabla\cdot(\rho\mathbf{v}) = 0. \tag{4.17}$$

This is called the continuity equation, or the conservation law of masses. It is the most fundamental equation in continuum mechanics.

4.4.2 Material derivative

Given a velocity field $\mathbf{v}(\mathbf{x}, t)$, we can define the particle path $\mathbf{x}(\cdot, \mathbf{X})$ to be the solution of the ordinary differential equation with initial position $\mathbf{X}$:

$$\dot{\mathbf{x}} = \mathbf{v}(\mathbf{x}, t), \mathbf{x}(0, \mathbf{X}) = \mathbf{X}.$$

The initial position $\mathbf{X}$ is usually called the material coordinate or the *Lagrange coordinate* of the flow. Given a flow qantity, say $f(\mathbf{x}, t)$, we can study how it changes along a particle path. This means we should fix the material coordinate and differentiate in t. That is

$$\begin{aligned} \left(\frac{d}{dt}\right)_{\mathbf{X}} f(\mathbf{x}(t, \mathbf{X}), t) &= \frac{\partial}{\partial t} f + \nabla_{\mathbf{x}} f \cdot \frac{d}{dt}\mathbf{x}(t, \mathbf{X}) \\ &= \left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}}\right) f \\ &:= \frac{D}{Dt} f. \end{aligned}$$

The derivative D/Dt is called the material derivative.

If $D\rho/Dt = 0$, the fluid is called *incompressible.* By the continuity equation

$$0 = \rho_t + \nabla \cdot (\rho \mathbf{v}) = \rho_t + \mathbf{v} \cdot \nabla \rho + \rho \nabla \cdot \mathbf{v}$$

We get that the incompressibility is equivalent to

$$\nabla \cdot \mathbf{v} = 0. \tag{4.18}$$

4.4.3 Momentum equation

Now, let us study the momentum change in an arbitrary domain Ω. According to Newton's force law, the change of momentum is due to the force it is exterted. There are two kinds of forces, (i) surface force T, and a body force f. The surface force is called the *stress.* It is a 3×3 matrix $T = (T_{ij})_{3\times 3}$. We shall explain more later. The change of momentum in Ω is

$$\frac{d}{dt} \int_{\Omega} \rho \mathbf{v} \, d\mathbf{x}.$$

There are three terms cause this change:

- the momentum flows into Ω through $\partial\Omega$: $-\int_{\partial\Omega} (\rho \mathbf{v})(\mathbf{v} \cdot \mathbf{n}) \, dS$
- the surface force on $\partial\Omega$: $\int_{\partial\Omega} T \cdot n \, dS$

- the body force in Ω: $\int_\Omega \rho f \, d\mathbf{x}$.

Thus, we get

$$\frac{d}{dt}\int_\Omega \rho\mathbf{v}\, d\mathbf{x} = -\int_{\partial\Omega} (\rho\mathbf{v})(\mathbf{v}\cdot\mathbf{n})\, dS + \int_{\partial\Omega} T\cdot n\, dS + \int_\Omega \rho f\, d\mathbf{x}.$$

By applying the divergence theorem, we get

$$\Big(\rho\mathbf{v}\Big)_t + \nabla\cdot\Big(\rho\mathbf{v}\mathbf{v}\Big) = \nabla\cdot T + \rho f.$$

We should be careful about the term $\nabla\cdot\rho\mathbf{v}\mathbf{v}$. It means that $\mathbf{v}\mathbf{v}$ is a matrix with entry v_iv_j. The ith component of $\nabla\cdot(\rho\mathbf{v}\mathbf{v})$ is $\sum_j \partial_j(\rho v_i v_j)$.

One can show by using the continuity equation to get that

$$\Big(\rho\mathbf{v}\Big)_t + \nabla\cdot\Big(\rho\mathbf{v}\mathbf{v}\Big) = \rho\frac{D\mathbf{v}}{Dt}$$

So sometimes, the momentum equation is expressed as

$$\rho\frac{D\mathbf{v}}{Dt} = \nabla\cdot T + \rho f.$$

Stress There are two kinds of fluid we consider.

- perfact fluid: $T = -pI$, no viscosity,
- viscous fluid: $T = -pI + \sigma$,

where p is the *pressure*, which is a function of ρ from theory of thermodynamics. And

$$\sigma = \mu(\nabla\mathbf{v} + (\nabla\mathbf{v})^T) + \lambda\nabla\cdot\mathbf{v}$$

is called the shear stress. The quantity $\frac{1}{2}(\nabla\mathbf{v} + (\nabla\mathbf{v})^T))$ is called the strain e. That is

$$e_{ij} = \frac{1}{2}\Big(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\Big)$$

The above relation between shear stress and strain is a generalized Hook's law.

Euler equation. For perfect fluid, we get the Euler equation

$$\left\{\begin{array}{l} \rho_t + \nabla\cdot(\rho\mathbf{v}) = 0 \\ \rho\frac{D\mathbf{v}}{Dt} + \nabla p = \rho f. \end{array}\right. \tag{4.19}$$

There are 4 equations with 4 unknowns $\rho, \mathbf{v}$. The pressure is a given function of ρ. In the second equation, we can rewrite it as

$$\mathbf{v}_t + \mathbf{v}\cdot\nabla\mathbf{v} + \frac{\nabla p}{\rho} = f$$

We can define a function $P(\rho)$ (called *entherpy* or *potential*) such that

$$P'(\rho) = \frac{p'(\rho)}{\rho}$$

Then the momentum equation becomes

$$\mathbf{v}_t + \mathbf{v}\cdot\nabla\mathbf{v} + \nabla P = f$$

In the case of incompressible flow, the continuity equation becomes

$$\nabla\cdot\mathbf{v} = 0.$$

The unknowns become P and $\mathbf{v}$.

Navier-Stokes equation. For *incompressible and viscous* fluid, we get the Navier-Stokes equation

$$\left\{\begin{array}{l} \nabla\cdot\mathbf{v} = 0. \\ \rho\frac{D\mathbf{v}}{Dt} + \nabla p = \rho f + \mu\nabla^2\mathbf{v}. \end{array}\right. \tag{4.20}$$

Remarks.

- For fluid mechanics, you may read pp. 227-231.
- If you want to have a better understanding on the strain and stress, read pp. 220-225, or you can read first few pages of the andau-Lipschitz's book on Fluid Mechanics.

4.4.4 Boundary Conditions

For viscous flows, the usual boundary condition is so called non-slip boundary condition

$$\mathbf{v} = 0 \text{ on } \partial\Omega$$

This means that the particles is attached to the boundary due to viscosity. However, if the flow is inviscid, we usually impose the following boundary condition

$$\mathbf{v} \cdot \mathbf{n} = 0.$$

This means that the fluid cannot flow into or out of the boundary.

4.4.5 Potential flows

An important quantity to study fluid dynamics is the *vorticity*. Physically, the vorticity is mainly generated from the friction of fluid with the bundary, or with surounding fluid. For perfect fluid under a conservative force $f = \nabla\Phi$, we claim if there is no vorticity initially, then there is no vorticity in all later time. To see this, we derive the vorticity for the perfect fluid. Define

$$\omega := \nabla \times \mathbf{v},$$

Take curl on the Euler equation for incompressible flow:

$$\nabla \times \Big(\mathbf{v}_t + \mathbf{v} \cdot \nabla \mathbf{v} + \nabla P - \nabla \Phi\Big) = 0.$$

The conservative force terms $\nabla \times (\nabla P - \nabla\Phi)$ disappear. We get

$$\frac{D\omega}{Dt} = (\omega \cdot \nabla)\mathbf{v}$$

If $\mathbf{v}$ is given, this is a linear ODE for ω. If $\omega = 0$ initially, then $\omega \equiv 0$ in all later time. Such kind of flows are called irrotational flows. An incompressible and irrotational flow satisfies

$$\left\{ \begin{array}{l} \nabla \cdot \mathbf{v} = 0 \\ \nabla \times \mathbf{v} = 0. \end{array} \right. \tag{4.21}$$

Since $\mathbf{v}$ is curl free, there exist a potential ϕ such that

$$\mathbf{v} = \nabla\phi.$$

Thus, an irrotational flow is called a *potential flow*. The incompressibility $\nabla \cdot \mathbf{v} = 0$ yields

$$\triangle \phi = 0. \tag{4.22}$$

Thus, the potential ϕ of a potential flow satisfies Laplace equation. On the boundary, we impose non-slip boundary condition

$$\nabla \phi \cdot \mathbf{n} = 0.$$

Alternatively, we can use $\mathbf{v}$ being divergence free in two dimensions, there exists a function ψ such that

$$\mathbf{v} = \psi^{\perp} := (\psi_y, -\psi_x).$$

The level set of ψ is tangent to $\psi^{\perp}$, hence to $\mathbf{v}$. This means that $\mathbf{v}$ is tangent to the level set of ψ. So, $\psi = const.$ are the stream lines of the flow. From $\nabla \times \mathbf{v} = 0$, we get

$$\nabla \times \mathbf{v} = \nabla \times (\psi_y, -\psi_x, 0) = (0, 0, -\triangle \psi) = 0.$$

we get

$$\triangle \psi = 0.$$

The non-slip boundary condition now reads

$$0 = \mathbf{v} \cdot \mathbf{n} = (\psi_y, -\psi_x) \cdot \mathbf{n} = (\psi_x, \psi_y) \cdot \mathbf{t}.$$

where $\mathbf{t}$ is the tangent of the boundary. That is, the tangential derivative of ψ along the boundary is zero. Hence, ψ is a constant along the boundary.

Homeworks.

- Below, we can view two dimensional potential flows are defined on complex plane. Verify the following function are the potentials of a potential flow. Plot their stream lines.

    1. $\phi = Re(z^2) = x^2 - y^2$
    2. $\phi = Re(\log z) = \log(\sqrt{x^2 + y^2})$
    3. $\phi = Re(-i \log z) = \theta(x, y) = \tan^{-1}(y/x)$

4.4.6 Stokes flows

In the incompressible viscous flows, if the flow moves slowly, then the acceleration term $D\mathbf{v}/Dt$ is usually relatively small. In this case, we can ignor it. The resulting equation is

$$\begin{aligned} \nabla \cdot \mathbf{v} &= 0 \\ -\mu \nabla^2 \mathbf{v} + \nabla p &= f. \end{aligned}$$

with boundary condition:

$$\mathbf{v} = 0 \text{ on boundary.}$$

We may write them in matrix form:

$$\begin{bmatrix} C^{-1} & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

where $C^{-1} = -\mu\nabla^2$, $A = \nabla$, the gradient operator, and $A^T = -\nabla\cdot$ is the minus divergence operator.

Homework.

- Show that
$$\int_\Omega \nabla p \cdot \mathbf{v}\, dx = \int_\Omega p\, \nabla \cdot \mathbf{v}\, dx$$
for any scalar field p and any vector field $\mathbf{v}$ with $\mathbf{v} = 0$ on $\partial\Omega$.

4.5 *Equilibruim of Elastic Material

The deformation of an elastic material is characteristized by the disciplacement $\mathbf{u}(\mathbf{x})$. If there is no deformation, $\mathbf{u}(\mathbf{x}) = 0$. If the orginal position of a small piece of material is $\mathbf{x}$ and it is deformed to ξ, then the discplacement is defined to be $\mathbf{u}(\mathbf{x}) := \xi - \mathbf{x}$. The Jacobian $J = \partial\mathbf{u}/\partial\mathbf{x}$ characterizes the local change of the material. If the material is expanding uniforming, then $\xi = \lambda\mathbf{x}$ with $\lambda > 1$. The corresponding $\mathbf{u} = (\lambda - 1)\mathbf{x}$, and the Jacobian $J = (\lambda - 1)I$. According to Hook's law, we should expect a restoration force σ which should point to the center at every point. Thus, expansion or shrinking of material causes a restoration force formed inside the material. In three dimensions, the material can stretch in x and y directions and dilate in the z-direction. Or in general, it can stretch (or dilate) in one direction and dilate (or stretch) on its orthogonal plane. This will introduces restoration force inside the material. It is a surface force. However, rotation of a material will introduce no restoration force inside the material. The infinitesimal deformation of is described by the Jacobian J. The anti-symmetric part of J corresponds to the rotation of the material. The symmetric part corresponds to the expansion or shrinking (non-isometric) of the material. This symmetric part is called the strain of the material. That is,

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

The divergence of J is responsible for the isometric expansion or shrinking of the material.

The generalized Hook's law is

$$\sigma_{i,j} = 2\mu e_{ij} + \lambda(e_{11} + e_{22} + e_{33})\delta_{ij}$$

σ is called the stress tensor. The parameters λ and μ are called Lamé constants. On a small piece of surface with normal $\mathbf{n}$, the surface force due to this stress is $\sigma\mathbf{n}$, which is a vector. It is the restoration force on this small piece of surface.

Consider any domain Ω in the material. The force balance law in Ω reads

$$\int_{\partial\Omega} \sigma \cdot \mathbf{n}\, dS + \int_\Omega f\, d\mathbf{x}$$

Apply the divergence theorem, we get

$$\int_\Omega (\nabla \cdot \sigma + f)\, d\mathbf{x}$$

for any domain Ω. We then get the force balance equilibrium equation

$$\nabla \cdot \sigma(\mathbf{x}) + f(\mathbf{x}) = 0,\ \mathbf{x} \in D \tag{4.23}$$

The boundary condition can be classified into two, one is Dirichlet, the other is Neumann. Namely, $\partial D = \Gamma_D \cup \Gamma_N$. On which, the following boundary conditions are imposed:

$$\begin{cases} \mathbf{u}(\mathbf{x}) = 0 & \mathbf{x} \in \Gamma_D \\ (\sigma \cdot \mathbf{n})(\mathbf{x}) = g(\mathbf{x}) & \mathbf{x} \in \Gamma_N \end{cases} \tag{4.24}$$

4.6 *Applications to Maxwell equations

We now apply the knowledge on vector calculus we learnt in the previous few subsections to simplify the important Maxwell systems. The Maxwell system is a system of first-order partial differential equations which can describe most electromagnetic phenomena and are given by the system in three dimensions:

$$\begin{aligned}
\varepsilon\,\frac{\partial \mathbf{E}}{\partial t} - \nabla \times \mathbf{H} &= -\mathbf{J} \quad (\mathbf{Ampere's\ law}) && (4.25)\\
\mu\,\frac{\partial \mathbf{H}}{\partial t} + \nabla \times \mathbf{E} &= 0 \quad (\mathbf{Faraday's\ law}) && (4.26)\\
\nabla \cdot (\varepsilon \mathbf{E}) &= \rho \quad (\mathbf{Gaussian\ laws}) && (4.27)\\
\nabla \cdot (\mu \mathbf{H}) &= 0 \quad (\mathbf{Gaussian\ laws}) && (4.28)
\end{aligned}$$

where **E** and **H** are the electric and magnetic field of the involved physical medium, $\varepsilon(\mathbf{x})$ and $\mu(\mathbf{x})$ are the electric permittivity and magnetic permeability of the medium.

Steady-state Maxwell system. In the steady-state case, the Maxwell system can be decoupled as follows:

$$\begin{aligned}
-\nabla \times \mathbf{H} &= -\mathbf{J}\,, \quad \nabla \cdot (\mu \mathbf{H}) = 0 && (4.29)\\
\nabla \times \mathbf{E} &= 0\,, \quad \nabla \cdot (\varepsilon \mathbf{E}) = \rho && (4.30)
\end{aligned}$$

The first order system may not be so easy to solve. So we would like to find a better system for both **E** and **H**.

First consider **E**. Using the first equation in (4.30), we can write $\mathbf{E} = \nabla u$ for some scalar u. Substituting it into the second equation in (4.30), we get a nice equation for u:

$$\nabla \cdot (\varepsilon \nabla u) = \rho\,.$$

After u is available, we can calculate **E** using the relation $\mathbf{E} = \nabla u$.

Then we consider **H**. Using the second equation in (4.29), we can write $\mu\,\mathbf{H} = \nabla \times \mathbf{A}$ for some vector-valued function **A**. Substituting it into the first equation in (4.29), we get a nice equation for **A**:

$$\nabla \times (\mu^{-1} \nabla \times \mathbf{A}) = \mathbf{J}\,, \quad \nabla \cdot \mathbf{A} = 0\,.$$

where the second equation is added to ensure the uniqueness of the solutions. After **A** is available, we can calculate **H** using the relation $\mu\,\mathbf{H} = \nabla \times \mathbf{A}$.

Time-dependent Maxwell system. It is easy to see that the two unknown functions $\mathbf{E}$ and $\mathbf{H}$ are coupled together in the time-dependent Maxwell system (4.25)-(4.28). Usually it is very difficult and expensive to solve this coupled system. A simple approach is to decouple the two unknown functions. To do so, we first take the time derivative of the Ampere's equation to get

$$\varepsilon(\mathbf{x})\,\frac{\partial^2\mathbf{E}}{\partial t^2} - \nabla\times\frac{\partial\mathbf{H}}{\partial t} = -\frac{\partial\mathbf{J}}{\partial t}\,,$$

then using the Faraday's equation we derive an equation for $\mathbf{E}$:

$$\varepsilon(\mathbf{x})\,\frac{\partial^2\mathbf{E}}{\partial t^2} + \nabla\times\Big(\mu(\mathbf{x})^{-1}\nabla\times\mathbf{E}\Big) = -\frac{\partial\mathbf{J}}{\partial t}\,.$$

Similarly, by taking the time derivative of the Faraday's equation, we get

$$\mu\,\frac{\partial^2\mathbf{H}}{\partial t^2} + \nabla\times\frac{\partial\mathbf{E}}{\partial t} = 0\,,$$

then using the Ampere's equation we derive an equation for $\mathbf{H}$:

$$\mu\,\frac{\partial^2\mathbf{H}}{\partial t^2} + \nabla\times\Big(\varepsilon(\mathbf{x})^{-1}\nabla\times\mathbf{H}\Big) = \nabla\times\left(\varepsilon(\mathbf{x})^{-1}\mathbf{J}\right).$$

4.7 Variation Formulation

We shall reformulate the Poisson equation

$$-\triangle u(\mathbf{x}) = f(\mathbf{x}),\ \ \mathbf{x}\in D \tag{4.31}$$

with Dirichlet boundary condition

$$u(\mathbf{x}) = 0, \mathbf{x}\in\partial D \tag{4.32}$$

as a minimization problem. Consider the functional

$$P(u) = \frac{1}{2}\int_D |\nabla u|^2\,d\mathbf{x} - \int_D f(\mathbf{x})u(\mathbf{x})\,d\mathbf{x} \tag{4.33}$$

This functional is called the Dirichlet integral. It is defined for all function u in C^1 So let us consider the following minimization problem:

$$\min_u P(u) \text{ with } u(\mathbf{x}) = 0 \text{ on } \partial D. \tag{4.34}$$

The minimum occurs at $P'(u^*) = 0$. To compute $P'(u)$, we perform the following directional derivative. Let v be a function in C^1 which is a variation of u. Since u is fixed on ∂D, we should consider $v(\mathbf{x}) = 0$ for $\mathbf{x} \in \partial D$. Consider $dP(u+tv)/dt$ at $t = 0$. This is the change of P at u in the direction v, or the directional derivative of P at u in the direction v.

$$\begin{aligned} P(u+tv) &= \int_D \left(\frac{1}{2}|\nabla(u+tv)|^2 - f(u+tv)\right) d\mathbf{x} \\ &= P(u) + t\int_D (\nabla u \cdot \nabla v - fv)\, d\mathbf{x} + \frac{t^2}{2}\int_D |\nabla v|^2\, d\mathbf{x} \end{aligned}$$

Hence,

$$\left.\frac{d}{dt}\right|_{t=0} P(u+tv) = \int_D (\nabla u \cdot \nabla v - fv)\, d\mathbf{x}$$

By applying the Green's theorem, we get

$$\int_D \nabla u \cdot \nabla v\, d\mathbf{x} = -\int_D \nabla \cdot (\nabla u) v\, d\mathbf{x} + \int_{\partial D} v \nabla u \cdot \mathbf{n}\, dS$$

From the condition $v(\mathbf{x}) = 0$ for $\mathbf{x} \in \partial D$, we get the boundary term is 0. Hence, we get

$$(P'(u), v) := \left.\frac{d}{dt}\right|_{t=0} P(u+tv) = \int_D (-\nabla^2 u - f) v\, d\mathbf{x} \tag{4.35}$$

$P'(u)$ is the gradient of P at u. It is

$$P'(u) = -\nabla^2 u - f \tag{4.36}$$

We find that $P'(u) = 0$ is exactly the equation (4.31). The equation $P'(u) = 0$ is called the Euler-Lagrange equation of the minimization problem (4.34).

Conversely, if u^* satisfies (4.31) with boundary condition (4.32), we claim it is the minimum of P among all function u satisfying the same boundary condition. Let u be any function in C^1 satisfies the same boundary condition. Let $v = u - u^*$. Then $v(\mathbf{x}) = 0$

for $\mathbf{x} \in \partial D$. We compare $P(u)$ and $P(u^*)$:

$$\begin{aligned}
P(u) - P(u^*) &= P(u^* + v) - P(u^*) \\
&= \int_D \left(\frac{1}{2} |\nabla(u^* + v)|^2 - \frac{1}{2} |\nabla u^*|^2 - fv \right) dx \\
&= \int_D (\nabla u \cdot \nabla v) - fv \, dx + \frac{1}{2} \int_D |\nabla v|^2 \, dx \\
&= \int_D ((-\nabla \cdot \nabla u^*) v - fv) \, dx + \int_{\partial D} v \nabla u \cdot n \, dS + \frac{1}{2} \int_D |\nabla v|^2 \, dx \\
&= \int_D ((-\nabla \cdot \nabla u^*) v - fv) \, dx + \frac{1}{2} \int_D |\nabla v|^2 \, dx \\
&= \frac{1}{2} \int_D |\nabla v|^2 \, dx \geq 0.
\end{aligned}$$

This shows u^* is the minimum.

In general, we consider the following Dirichlet integral

$$P(u) := \int_D \left(\frac{c(x)}{2} |\nabla u(x)|^2 - f(x) u(x) \right) dx \tag{4.37}$$

where $c(x)$ is a strictly positive function, i.e. $c(x) \geq c_0 > 0$ for all $x \in D$. The Dirichlet principle states

The function u^* which minimizes $P(u)$ with $u(x) = g(x)$ for $x \in \partial D$ must satisfies the Euler-Lagrange equation

$$-\nabla \cdot (c(x) \nabla u) = f \tag{4.38}$$

with boundary condition

$$u(x) = g(x), x \in \partial D. \tag{4.39}$$

The converse is also true.

Next, we express the Euler-Lagrange equation in variation form. Let us define

$$a(u, v) := \int_D c(x) \nabla u(x) \cdot \nabla v(x) \, dx$$

and

$$(f, v) = \int_D f(x) v(x) \, dx$$

We have the following weak form:

The function u^* with boundary condition (4.39) satisfies the Euler-Lagrange equation (4.38) if and only if it satisfies the following variation equation:

$$a(u^*, v) = (f, v) \textbf{ for all } v \textbf{ with } v(x) = 0 \textbf{ on } \partial D. \tag{4.40}$$

This variation formulation will be used in the finite element method for (4.38) (4.39).

Homework

- Prove the above general Dirichlet principle.

5 Analytical methods

5.1 Motivation of Fourier series: Solving heat equation on a circle

In this section, we are going to present one of the most powerful analytical methods – Fourier series. We start from Fourier's original approach, solving heat equation on a circle as our motivation. We then explain the theory of Fourier series and with an application to solve Laplace equation on a disk. Finally, we extend the Fourier series to Fourier transform and its application to solving some differential equation on line.

We consider the heat equation on a circle:

$$u_t = u_{xx},\ x \in [0, 2\pi]$$

with periodic boundary condition

$$u(0,t) = u(2\pi, t),$$

and initial condition

$$u(x,0) = f(x).$$

The reason why we consider periodic boundary condition is for simlicity. The method can also be extended to arbitrary finite interval with Dirichlet boundary condition or Neumann boundary condition.

Fourier had two important observations for this equation:

- If we differentiate $\cos kx$ twice, we get the same $\cos kx$:

  $$\frac{d^2}{dx^2} \cos kx = -k^2 \cos kx.$$

  Here, k is an integer. In modern language, $\cos kx$ is an eigenvector of the differential operator with eigenvalue $-k^2$. Thus, we can guess a solution looks like

  $$u(x,t) = a(t) \cos kx.$$

  Using this *ansatz*, we plug it into equation, we get

  $$\dot{a}(t) \cos kx = -k^2 a(t) \cos kx.$$

We eliminate $\cos kx$ and find that $a(t)$ satisfies

$$\dot{a}(t) = -k^2 a(t).$$

This can be solved immediately. Namely

$$a(t) = a(0)e^{-k^2 t}.$$

Thus,

$$u(x,t) = a(0)e^{-k^2 t}\cos kx$$

is a solution of the heat equation. Same property also happens on $\sin kx$. We also have solution looks like

$$u(x,t) = b(0)e^{-k^2 t}\sin kx$$

for any positive integer k.

- If we know two solutions u_1 and u_2, then their linear combination is also a solution. This is because the equation is linear.

Combining these two observations, we immediately get that all functions

$$u(x,t) = \sum_{k=0}^{N}\Big(a_k \cos kx + b_k \sin kx\Big)e^{-k^2 t}.$$

are solutions of the heat equation on circle. The corresponding initial data is

$$u(x,0) = \sum_{k=0}^{N} a_k \cos kx + b_k \sin kx$$

The function $\sum_{k=0}^{N} a_k \cos kx + b_k \sin kx$ is called a trigonometric polynomial. Just like the case of Taylor series: a nice general function can be approximated by polynomial with infinite terms, we would also like to approximate any 2π-periodic function by Fourier series. So

Question 1. *Can any 2π-periodic function $f(x)$ be represented as a trigonometric series?*
That is

$$f(x) = \sum_{k=0}^{\infty}\Big(a_k \cos kx + b_k \sin kx\Big)$$

for some coefficients a_k and b_k. How to find these coefficients?

Question 2. *Does the function*

$$u(x,t) = \sum_{k=0}^{\infty} \Big(a_k \cos kx + b_k \sin kx\Big) e^{-k^2 t} \tag{5.1}$$

solve the heat equation with the initial condition $u(x,0) = f(x)$?

This is a motivation to develop the theory of Fourier series.

Notice that we have the following *orthogonality* of $\cos kx$ and $\sin kx$:

$$\int_0^{2\pi} \cos kx \cos mx \, dx = \begin{cases} 2\pi & \text{if } k = m = 0 \\ \pi & \text{if } k = m \neq 0 \\ 0 & \text{if } k \neq m \end{cases}$$

Same property for $\sin kx$. Further, $\sin kx$ is always orthogonal to $\cos mx$ for any integers k, m:

$$\int_0^{2\pi} \sin kx \cos mx \, dx = 0.$$

To find the coefficient a_m, we multiply (5.1) by $\cos mx$) and integrate from 0 to 2π. Using the orthogonality of $\cos kx$ and $\sin kx$, we get all terms are disappeared except the term a_m:

$$\int_0^{2\pi} f(x) \cos mx \, dx = \pi a_m, m > 0.$$

This is for the case $m \neq 0$. When $m = 0$, we get

$$\int_0^{2\pi} f(x) \, dx = 2\pi a_0$$

For b_m, we have

$$\int_0^{2\pi} f(x) \sin mx \, dx = \pi b_m, m > 0.$$

The above procedure is at least formally true. In fact, the exploration of Fourier series opened the door of *modern analysis.*

Before we formally introduce Fourier theory, let us make the Fourier remarks on theessential keys of this theory. According to Euler, the two key roles $\cos kx$ and $\sin kx$ can be put together into one as

$$e^{ikx} = \cos kx + i \sin kx.$$

The very essential keys of Fourier series are

1. The function e^{ikx} is an eigenvector of the differential operator d/dx. In other words, e^{ikx} can be used to *diagonalize* differential operators.

2. The functions e^{ikx} are *oscillatory* and wiggle more and more as k increase. They can be ued to approximate any function with oscillation at any scale.

3. The functions e^{ikx} are orthogonal! That is
$$\int_0^{2\pi} e^{ikx}e^{imx}\,dx = 2\pi\delta_{km},$$
where
$$\delta_{km} = \begin{cases} 1 & \text{if } k = m, \\ 0 & \text{if } k \neq m. \end{cases}$$

5.2 Fourier Series

5.2.1 Inner product function spaces

Let us learn some basic teminology about periodic functions.

Definition 5.1 (Periodic functions). *A function $f(x)$ is called a periodic function with period d if*
$$f(x+d) = f(x) \quad \forall\, x\ .$$

For example, e^{ikx}, $\cos kx$ and $\sin kx$ are all periodic functions with period 2π. But
$$\cos 2kx, \quad \sin 2kx$$
are periodic functions with period 2π and also π. Because of the periodicity, we can consider any interval of length 2π for the Fourier expansions (5.9) and (5.2). We often take $[-\pi, \pi]$ or $[0, 2\pi]$. In our subsequent discussions, we will always use the interval $[-\pi, \pi]$.

Now suppose $f(x)$ is a function with period 2π, i.e., $f(x+2\pi) = f(x)$ $\forall x$. In this case, the graph of $f(x)$ in any interval of length 2π will be repeated in its neighboring interval of length 2π.

Let us denote all real-valued 2π-periodic functions by V. It is a vector space over $\mathbb{R}$. In V, we define the inner product
$$(f, g) = \int_{-\pi}^{\pi} f(x)g(x)\,dx.$$
One can show that it satisfies

1. $(f, f) \geq 0$ and $(f, f) = 0$ if and only if $f = 0$
2. $(f, g) = (g, f)$
3. $(f_1 + f_2, g) = (f_1, g) + (f_2, g)$
4. $(\alpha f, g) = \alpha(f, g)$

We can define the norm

$$\|f\| = \sqrt{(f, f)}.$$

We can see that the inner product satisfies the following Cauchy's inequality:

$$|(f, g)| \leq \|f\| \, \|g\|$$

This is because

$$0 \leq (f + tg, f + tg) = \|f\|^2 + 2t(f, g) + t^2\|g\|^2$$

for any $t \in \mathbb{R}$. Therefore, we have

$$(f, g)^2 - \|f\|^2\|g\|^2 \leq 0.$$

From this Cauchy's inequality, it is natural to define the angle θ between two vectors f and g by

$$\cos\theta = \frac{(f, g)}{\|f\| \, \|g\|}$$

Definition 5.2 (Orthogonal functions)**.** *Two real-valued functions $f(x)$ and $g(x)$ are said to be orthogonal on the interval $[a, b]$ if the following holds*

$$(f, g) := \int_a^b f(x)g(x)dx = 0.$$

Similarly one can verify that the following three sequences

$$\{\cos kx\}_{k=0}^{\infty}, \quad \{\sin kx\}_{k=0}^{\infty}, \quad \{\cos kx, \sin kx\}_{k=0}^{\infty}$$

are all orthogonal sequences of functions on $[-\pi, \pi]$ or $[0, 2\pi]$.

5.2.2 Real Fourier series

We first discuss how to find the Fourier series

$$f(x) = \sum_{k=0}^{\infty}(a_k \cos kx + b_k \sin kx). \tag{5.2}$$

We need to find all the coefficients $\{a_k\}$ and $\{b_k\}$. Notice that for $k = 0$, $\sin kx \equiv 0$, so we don't need the coefficent b_0. Recall that $\{\cos kx, \sin kx\}$ are orthogonal on $[-\pi, \pi]$, namely for any $k \neq l$,

$$\int_{-\pi}^{\pi} \cos kx \cos lx dx = 0, \tag{5.3}$$

$$\int_{-\pi}^{\pi} \cos kx \sin lx dx = 0, \tag{5.4}$$

$$\int_{-\pi}^{\pi} \sin kx \sin lx dx = 0 \ . \tag{5.5}$$

Find the coefficient a_k in (5.2). Multiply both sides of (5.2) by $\cos kx$, integrate then over $[-\pi, \pi]$ and use the orthogonality (5.3)-(5.5). We have

$$\int_{-\pi}^{\pi} f(x) \cos kx dx = \int_{-\pi}^{\pi} a_k \cos kx \cos kx dx.$$

From this we obtain

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos kx dx \tag{5.6}$$

since

$$\int_{-\pi}^{\pi} \cos^2 kx dx = \int_{-\pi}^{\pi} \frac{1 + \cos 2kx}{2} dx = \pi \ .$$

Find the coefficient b_k in (5.2). Multiply both sides of (5.2) by $\sin kx$, integrate then over $[-\pi, \pi]$ and use the orthogonality (5.3)-(5.5). We have

$$\int_{-\pi}^{\pi} f(x) \sin kx dx = \int_{-\pi}^{\pi} b_k \sin kx \sin kx dx.$$

From this we obtain

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \sin kx dx \, . \tag{5.7}$$

Find the coefficient a_0 in (5.2). Multiply both sides of (5.2) by the constant 1, then integrate over $[-\pi, \pi]$ to obtain

$$\int_{-\pi}^{\pi} f(x) dx = \int_{-\pi}^{\pi} a_0 dx,$$

therefore,

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx\,, \tag{5.8}$$

that is, the first coefficient a_0 is the average of $f(x)$ on $[-\pi, \pi]$.

In summary, we can expand $f(x)$ as follows:

$$f(x) = a_0 + a_1\cos x + b_1 \sin x + a_2\cos 2x + b_2\sin 2x + \cdots$$

where all the coefficients are given by

$$\begin{aligned} a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx\,, \\ a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx\,, \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx\,. \end{aligned}$$

5.2.3 Complex Fourier series

First let us consider class of all complex-valued 2π-periodic functions W. It is a vector spaces over $\mathbb{C}$. We define the following inner product:

$$(f, g) := \int_{-\pi}^{\pi} f(x)\,\overline{g(x)}\,dx.$$

One can show that it satisfies

1. $(f, f) \geq 0$ and $(f, f) = 0$ if and only if $f = 0$
2. $(f, g) = \overline{(g, f)}$
3. $(f_1 + f_2, g) = (f_1, g) + (f_2, g)$
4. $(\alpha f, g) = \alpha(f, g)$

We can define the norm

$$\|f\| = \sqrt{(f, f)}.$$

Through

$$0 \leq (f + tg, f + tg) = \|f\|^2 + 2Re(f, g)t + t^2\|g\|^2$$

for any real t, we get that

$$Re(f, g) \leq \|f\|\,\|g\|$$

If we replace g by $g_1 := g(f,g)/|(f,g)|$, then we get

$$|(f,g)| = Re(f,g_1) \le \|f\|\,\|g_1\| = \|f\|\,\|g\|.$$

This leads to the Cauchy's inequality

$$|(f,g)| \le \|f\|\,\|g\|.$$

Definition 5.3 (Orthogonal functions)**.** *Two complex functions $f(x)$ and $g(x)$ are said to be orthogonal on the interval $[a,b]$ if the following holds*

$$(f,g) := \int_a^b f(x)\overline{g(x)}dx = 0$$

where $\overline{g(x)}$ is the conjugate of $g(x)$. For example, $\{e^{ikx}\}_{k=-\infty}^{\infty}$ is an orthogonal sequence of functions on $[-\pi,\pi]$ or $[0,2\pi]$, since

$$(e^{ikx}, e^{ilx}) = \int_0^{2\pi} e^{ikx}\overline{e^{ilx}}dx = 0 \quad \forall\, l \neq k$$

In fact,

$$\int_0^{2\pi} e^{ikx}\overline{e^{ilx}}dx = \int_0^{2\pi} e^{ikx}e^{-ilx}dx = \frac{1}{i(k-l)}e^{i(k-l)x}\Big|_0^{2\pi} = 0\ .$$

Now we shall discuss how to expand a complex-valued function f interms of $\{e^{ikx}\}_{k\in\mathbb{Z}}$:

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \tag{5.9}$$

We need to find all the coefficients $\{c_k\}$. We know that $\{e^{ikx}\}$ are orthogonal on $[-\pi,\pi]$, namely for any $k \neq l$,

$$(e^{ikx}, e^{ilx}) = \int_{-\pi}^{\pi} e^{ikx}e^{-ilx}dx = 0 \quad \forall\, k \neq l.$$

Thus multiply both sides of (**??**) by e^{-ikx} and use the orthogonality of $\{e^{ikx}\}$, we obtain

$$\int_{-\pi}^{\pi} f(x)e^{-ikx}dx = \int_{-\pi}^{\pi} c_k e^{ikx}e^{-ikx}dx\ ,$$

or

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx\ . \tag{5.10}$$

That is, the Fourier series is

$$f(x) = c_0 + c_1e^{ix} + c_{-1}e^{-ix} + c_2e^{2ix} + c_{-2}e^{-2ix} + \cdots \tag{5.11}$$

with coefficients c_k defined by (5.10).

Remark 5.1. *Note that in the Fourier series (5.11), the function $f(x)$ can be a real function. For a real function, one can choose the real Fourier expansion (5.2) or the complex form (5.11).*

Think about why we can choose the complex form (5.11) for a real function. Any contradiction ?

5.2.4 Relation between the real and complex Fourier series

There are close relations between the real and complex Fourier series.

(a) The coefficients c_k in the complex form (5.11) can be derived from the coefficients a_k and b_k in the real form (5.2). In fact, we know

$$e^{ikx} = \cos kx + i \sin kx, \quad e^{-ikx} = \cos kx - i \sin kx. \tag{5.12}$$

Multiply both sides of the second equation by $f(x)$ and integrate over $[-\pi, \pi]$. We obtain

$$\int_{-\pi}^{\pi} f(x) e^{-ikx} dx = \int_{-\pi}^{\pi} f(x) \cos kx dx - i \int_{-\pi}^{\pi} f(x) \sin kx dx \ ,$$

That implies

$$2c_k = a_k - i\, b_k\,. \tag{5.13}$$

This can be written as

$$c_k = \frac{1}{2} a_k - \frac{i}{2} b_k \ .$$

Similarly, we can derive from the first equation of (5.12):

$$c_{-k} = \frac{1}{2} a_k + \frac{i}{2} b_k \ .$$

(b) The coefficients a_k and b_k in (5.2) can be recovered from the complex coefficients c_k in (5.11). Using the formula

$$\cos kx = \frac{1}{2}\big(e^{ikx} + e^{-ikx}\big), \quad \sin kx = \frac{1}{2i}\big(e^{ikx} - e^{-ikx}\big) \ .$$

Therefore

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx = c_k + c_{-k} \ , \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx = \frac{1}{i}(c_{-k} - c_k). \end{aligned}$$

c If f is *real-valued*, it is easy to check that the follows are equivalent:

- f is real-valued
- a_k and b_k are real for all k
- $c_k = \overline{c_{-k}}$

5.3 Examples of Fourier series

We now give some examples to illustrate the calculations of the Fourier series.

Example 5.1. *Find the Fourier series of $f(x) = \cos^2 x$.*

<u>**Solution**</u>. By definition, we have

$$\begin{aligned}
a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1+\cos 2x}{2}dx = \frac{1}{2}\ , \\
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kxdx = \frac{1}{2\pi}\int_{-\pi}^{\pi}(1+\cos 2x)\cos kxdx = \begin{cases} 0\ , & k \neq 2 \\ \frac{1}{2}\ , & k = 2\ , \end{cases} \\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kxdx = \frac{1}{2\pi}\int_{-\pi}^{\pi}(1+\cos 2x)\sin kxdx = 0\ .
\end{aligned}$$

Therefore the Fourier series of $f(x)$ is

$$f(x) = \frac{1}{2} + \frac{1}{2}\cos 2x\ .$$

This is a well-known formula. ♯

- Try the Fourier expansions of the functions $\sin^2 x$, $\cos 2x$, $\sin x + \cos x$.
- Try the Fourier expansions of the functions

1. $f(x) = \begin{cases} -1 & -\pi < x < 0 \\ 1 & 0 < x < \pi, \end{cases}$

2. $f(x) = \begin{cases} x+\pi & -\pi < x < 0 \\ \pi - x & 0 < x < \pi, \end{cases}$

3. $f(x) = (\pi - x)(x + \pi)$

Example 5.2. *Find the Fourier series of $f(x) = \delta(x)$ on $[-\pi, \pi]$. This function is called a delta function and it is one of the most important functions used in physics and engineering. The delta function has the following properties*

$$\int_{-\pi}^{\pi} g(x)\delta(x)dx = g(0) \quad \forall\ g \in C\ [-\pi, \pi]$$

and

$$\delta(x) = 0 \quad \text{for any} \quad x \neq 0.$$

Solution. By definition, the Fourier coefficients are

$$\begin{aligned} a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \delta(x)dx = \frac{1}{2\pi}\ , \\ a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} \delta(x)\cos kx dx = \frac{1}{\pi}\cos 0 = \frac{1}{\pi}\ , \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} \delta(x)\sin kx dx = 0\ , \end{aligned}$$

therefore

$$\delta(x) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{k=1}^{\infty} \cos kx\ ,\ x \in\ [-\pi, \pi] \tag{5.14}$$

In the complex case,

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx = \frac{1}{2\pi}\ ,$$

so we have

$$\delta(x) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} e^{ikx}\ ,\ x \in\ [-\pi, \pi]\,. \tag{5.15}$$

We have from (5.15) that

$$\begin{aligned} \delta(x) &= \frac{1}{2\pi} + \frac{1}{2\pi}\sum_{k=1}^{\infty} \left(e^{ikx} + e^{-ikx}\right) \\ &= \frac{1}{2\pi} + \frac{1}{\pi}\sum_{k=1}^{\infty} \cos kx\ , \end{aligned}$$

this is the same as (5.14).

5.3.1 Sine series and cosine series

Odd and even functions. A function $f(x)$ is called an *even function* if it satisfies

$$f(-x) = f(x), \quad \forall\, x\,.$$

And it is called an *odd function* if it satisfies

$$f(-x) = -f(x), \quad \forall\, x\,.$$

It is easy to check the following properties:

For any odd function $f(x)$ on $[-\pi, \pi]$, we have

$$\int_{-\pi}^{\pi} f(x)dx = 0\,.$$

For any even function $f(x)$ on $[-\pi, \pi]$, we have

$$\int_{-\pi}^{\pi} f(x)dx = 2\int_{0}^{\pi} f(x)dx\ .$$

Example 5.3. *Find the Fourier series of the odd function*

$$f(x) = x\ ,\ x \in\ [-\pi, \pi]\ .$$

Solution. The Fourier coefficients are

$$\begin{aligned} a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = 0 \quad \text{(why ?)} \\ a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx = 0 \quad \text{(why ?)} \end{aligned}$$

Finally for the coefficients b_k, we have

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx = \frac{2}{\pi}\int_{0}^{\pi} x\sin kx dx$$

By integration by parts, we obtain

$$b_k = \frac{2}{k\pi}\int_{0}^{\pi}\cos kx dx - \frac{2}{\pi k} x\cos kx\Big|_0^{\pi} = -\frac{2\cos k\pi}{k}\ ,$$

that is,

$$b_1 = 2, \quad b_2 = -\frac{2}{2}, \quad b_3 = \frac{2}{3}, \cdots, \quad b_k = (-1)^{k+1}\frac{2}{k} \ ,$$

so the required Fourier series is

$$x = b_1 \sin x + b_2 \sin 2x + \cdots = 2(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \cdots), \quad -\pi < x < \pi \ .$$

Remark 5.2. *Note that the Fourier series above does not converge at $x = -\pi, \pi$, as the series is* 0 *at $x = -\pi$ and π.*

A very important observation:

The Fourier series of an even function has only cosine terms, since

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \sin kx dx = 0 \ .$$

The Fourier series of an odd function has only sine terms, since

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos kx dx = 0 \ .$$

Every function $f(x)$ can be written as a sum of an even and an odd function, i.e.,

$$f(x) = f_e(x) + f_o(x),$$

with

$$f_e(x) = \frac{f(x) + f(-x)}{2} \ , \quad f_o(x) = \frac{f(x) - f(-x)}{2} \ .$$

On the other hand, every function f on $(0, \pi)$ can have two ways of extension to $(-\pi, \pi)$:

- even extension: define $f(x) = f(-x)$ for $x \in (-\pi, 0)$,
- odd extension: define $f(x) = -f(-x)$ for $x \in (-\pi, 0)$.

Example 5.4. *The function $f(x) = 1$ is known on the half-period* $0 < x < \pi$. *Find its Fourier series when*

(a) $f(x)$ is extended to $(-\pi, \pi)$ as an even function;

(b) $f(x)$ is extended to $(-\pi, \pi)$ as an odd function.

Solution. By definition of even and odd functions, we have

(a) Extend $f(x)$ as an even function,

$$\begin{aligned}
a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = \frac{1}{\pi}\int_0^{\pi} 1dx = 1 \ , \\
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx = \frac{2}{\pi}\int_0^{\pi} \cos kx dx = 0 \ , \\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx = 0 \ ,
\end{aligned}$$

therefore the Fourier series of $f(x)$ is

$$f(x) = 1, \quad -\pi < x < \pi.$$

This recovers the original constant function.

(b) Extend $f(x)$ as an odd function,

$$\begin{aligned}
a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = 0 \ , \\
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx = 0 \ , \\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx = \frac{2}{\pi}\int_0^{\pi} \sin kx dx \\
&= -\frac{2}{k\pi}\big((-1)^k - 1\big) = \begin{cases} 0, & k \text{ is even} \\ \frac{4}{k\pi}, & k \text{ is odd} \end{cases} ,
\end{aligned}$$

so the Fourier series of $f(x)$ is

$$f(x) = \frac{4}{\pi}\Big\{\frac{\sin x}{1} + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots\Big\} \ , \quad -\pi < x < \pi.$$

This is very different from the original constant function.

5.4 Some properties of Fourier series

Best approximation by trigonometric polynomials Let consider the space

$$V_N := \{F(x) = \sum_{k=0}^{N}(A_k \cos kx + B_k \sin kx), A_k, B_k \in \mathbb{R}\}$$

It is the spaces of all trigonometric polynomials of order less or equal to N. For any 2π periodic function f, we measure the distance from f to V_N by

$$\min_{F\in V_N} \|f-F\|^2$$

This is equivalent to

$$\|f-F\|^2 = E(A_0, A_1, \cdots, A_n, B_0, B_1, \cdots, B_n) = \int_{-\pi}^{\pi} \big\{f(x)-\sum_{k=0}^{N}(A_k\cos kx+B_k\sin kx)\big\}^2 dx\,.$$

Then we claim that

The best trigonometric approximation of $f(x)$ on $[-\pi,\pi]$ in the mean-square sense is its Fouries series, i.e.,

$$E(a_0, a_1, \cdots, a_N, b_0, b_1, \cdots, b_N) = \min_{\forall\, A_k, B_k\in\mathbb{R}^1} E(A_0, A_1, \cdots, A_N, B_0, B_1, \cdots, B_N)$$

where $\{a_k\}$ and $\{b_k\}$ are the Fourier coefficients of $f(x)$.

To see this, let us assume $\{A_k, B_k\}_{K=0}^{n}$ is a minimizer of E, then

$$\begin{aligned}\frac{\partial E}{\partial A_k} =&2\int_{-\pi}^{\pi}\big\{f(x)-\sum_{K=0}^{n}(A_k\cos kx+B_k\sin kx)\big\}\cos kx dx\\ =&2\int_{-\pi}^{\pi}\big\{f(x)-A_k\cos kx\big\}\cos kx dx\\ =&2\int_{-\pi}^{\pi} f(x)\cos kx dx - 2\pi A_k = 0\ ,\end{aligned}$$

therefore

$$A_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx = a_k,$$

for $k\neq 0$. Similarly we have

$$A_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = a_0\ ,$$

and

$$B_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx = b_k\ ,\ \forall k\ .$$

This indicates that the minimizer $\{A_k, B_k\}_{k=0}^{n}$ is the Fourier coefficients of $f(x)$. $\sharp$

• Think about why we can claim what we get is the minimizer, not the maximizer. Think about the difference between $E(\{A_i\},\{B_i\})$ and $E(\{a_i\},\{b_i\})$.

Bessel's inequality Our second claim is:

Let $f_N(x)$ be the truncated Fourier series

$$f_N(x) = \sum_{k=0}^{N} (a_k \cos kx + b_k \sin kx)$$

then we have

$$\int_{-\pi}^{\pi} f_N^2(x)dx \le \int_{-\pi}^{\pi} f^2(x)dx\,.$$

First we have by using the orthogonality that for any $0 \le k \le N$,

$$\begin{aligned}\int_{-\pi}^{\pi} \big(f(x) - f_N(x)\big) \cos kx dx = 0\\ \int_{-\pi}^{\pi} \big(f(x) - f_N(x)\big) \sin kx dx = 0\end{aligned}$$

This implies that

$$\int_{-\pi}^{\pi} \big(f(x) - f_N(x)\big) f_N(x) dx = 0.$$

Thus

$$\begin{aligned}\int_{-\pi}^{\pi} f^2(x)dx &= \int_{-\pi}^{\pi} \big(f(x) - f_N(x) + f_N(x)\big)^2 dx\\ &= \int_{-\pi}^{\pi} \big(f(x) - f_N(x)\big)^2 dx + 2\int_{-\pi}^{\pi} \big(f(x) - f_N(x)\big) f_N(x) dx + \int_{-\pi}^{\pi} f_N^2(x)dx\\ &= \int_{-\pi}^{\pi} \big(f(x) - f_N(x)\big)^2 dx + \int_{-\pi}^{\pi} f_N^2(x)dx\\ &\ge \int_{-\pi}^{\pi} f_N^2(x)dx\,.\end{aligned}$$

♯

- Think about the interesting question. If we define a sequence $\{\alpha_N\}$ by

$$\alpha_N = \int_{-\pi}^{\pi} F_N^2(x)dx,$$

then the sequence $\{\alpha_N\}_{n=0}^{\infty}$ must be monotonely increasing.

Convergence of the Fourier series Our third claim is

Let f be a 2π-periodic smooth function. Let $f_N(x)$ be its truncated Fourier series

$$f_N(x) = \sum_{k=0}^{N}(a_k \cos kx + b_k \sin kx)$$

then we have

$$f(x) = \lim_N f_N(x) = \sum_{k=0}^{\infty}(a_k \cos kx + b_k \sin kx)$$

Notice that from the relation between real Fourier series and complex Fourier series, the partial sum has two expressions:

$$f_N(x) = \sum_{k=0}^{N}(a_k \cos kx + b_k \sin kx) = \sum_{-N}^{N} c_k e^{-ikx}$$

We plug

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y)e^{-iky}\,dy$$

into the above formula to get

$$\begin{aligned}
f_N(x) &= \sum_{-N}^{N} c_k e^{-ikx} \\
&= \sum_{-N}^{N} \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y)e^{-iky}\,dy e^{-ikx} \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y) \sum_{k=-N}^{N} e^{ik(x-y)}\,dy \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y)P_N(x-y)\,dy \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-y')P_N(y')\,dy'
\end{aligned}$$

where $P_N(x)$ is the partial sum of (5.15):

$$\begin{aligned} P_N(x) &= \sum_{k=-N}^{N} e^{ikx} = \frac{1}{2\pi} e^{-iNx} \sum_{k=-N}^{N} e^{i(N+k)x} \\ &= e^{-iNx} \sum_{k=0}^{2N} e^{ikx} \\ &= e^{-iNx} \frac{1 - e^{i(2N+1)x}}{1 - e^{ix}} \\ &= \frac{e^{i(N+\frac{1}{2})x} - e^{-i(N+\frac{1}{2})x}}{e^{ix/2} - e^{-ix/2}} \\ &= \frac{\sin(N+\frac{1}{2})x}{\sin\frac{1}{2}x} \ . \end{aligned}$$

We claim that

$$\frac{1}{2\pi} P_N(y') \to \delta(y'), \text{ as } N \to \infty.$$

If so, then we get

$$f_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-y') P_N(y')\, dy' \to \int_{-\pi}^{\pi} f(x-y') \delta(y')\, dy' = f(x).$$

We shall not prove this claim. Instead, we raise the following questions for students to think.

Can the series (5.15) really reflect the behavior of $\delta(x)$?

Study the following questions

1. For each given N, show that

$$\lim_{x\to 0} P_N(x) = 2(N + \frac{1}{2}).$$

   So $P_N(x)$ will tend to infinity at $x = 0$ when N goes larger and larger.

2. $\frac{1}{2\pi} \int_{-\pi}^{\pi} P_N(x)\, dx = 1$

3. Plot the figure for $P_N(x)$ using Matlab; and calculate the integral

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} P_N(x) f(x) dx$$

   approximately for $N = 10, 20, 30, 40, 50, 100$. Observe if $P_N(x)$ satisfies that

$$\lim_{N\to\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} P_N(x) f(x) dx = f(0).$$

   if so, $\frac{1}{2\pi} P_N(x)$ approximates $\delta(x)$.

Homeworks.

- pp. 285, 4.1.1
- pp. 286, 4.1.2
- pp. 286, 4.1.3
- pp. 286, 4.1.5
- pp. 286, 4.1.8
- pp. 286, 4.1.9
- pp. 287, 4.1.10
- pp. 287, 4.1.11

5.5 Application of Fourier expansion: Laplace's equation on a disk

In this section, we are going to apply the Fourier series to solve an important differential equation, i.e., the Laplace's equation:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x, y) \in \Omega \tag{5.16}$$

with the boundary condition

$$u(x, y) = u_0(x, y), \quad (x, y) \in \partial\Omega \tag{5.17}$$

where Ω is the unit circle, i.e.,

$$\Omega = \{(x, y);\ x^2 + y^2 < 1\}.$$

Since Ω is a circle, it is easier to use the polar coordinates:

$$x = r\cos\theta\ ,\ \ y = r\sin\theta\ .$$

Under the transformation, the domain Ω and the equation (5.16) are transformed into

$$\omega = \{(r, \theta);\ 0 \le r < 1,\ \ -\pi \le \theta < \pi\}$$

and

$$\frac{1}{r}\frac{\partial}{\partial r}\big(r\frac{\partial u}{\partial r}\big) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0\ . \tag{5.18}$$

The boundary condition (5.17) changes into

$$u(1, \theta) \ = \ u_0(\theta)\,. \tag{5.19}$$

Notice that we hav used the same notations u and u_0 for both the Cartesian coordinate and the polar coordinate. We are now going to find the solutions of (5.18). First, we can easily check that the following functions

$$1,\ r\cos\theta,\ r\sin\theta,\ r^2\cos 2\theta,\ r^2\sin 2\theta, \cdots \tag{5.20}$$

are all solutions of (5.18). For example, we take $u(r, \theta) = r^k \cos k\theta$ for $k \ge 2$, then

$$u_r = kr^{k-1}\cos k\theta\ ,$$
$$\frac{1}{r}\frac{\partial}{\partial r}(ru_r) = k^2 r^{k-2}\cos k\theta\ ,$$

while

$$\frac{1}{r^2}u_{\theta\theta} = -k^2 r^{k-2}\cos k\theta \ ,$$

therefore $u(r,\theta) = r^k \cos k\theta$ is a solution to the equation (5.18). Note that (5.18) is a linear equation, so any combination of two solutions $u_1(r,\theta)$ and $u_2(r,\theta)$ is still a solution (**why ?**). Thus the following combination of the above special solutions is a also solution:

$$u(r,\theta) = a_0 + a_1 r\cos\theta + b_1 r\sin\theta + \cdots + a_k r^k \cos k\theta + b_k r^k \sin k\theta + \cdots \ , \tag{5.21}$$

where a_k and b_k are arbitrary constants.

But we have to determine the coefficients a_k and b_k. This can be done by using the boundary condition (5.19). For this, we let $r = 1$ in (5.21) and obtain

$$u(1,\theta) = a_0 + a_1 \cos\theta + b_1 \sin\theta + \cdots + a_k \cos k\theta + b_k \sin k\theta + \cdots \ .$$

We know that $u(1,\theta) = u_0(\theta)$, so the coefficients a_k and b_k are nothing else but the Fourier coefficients of u_0, i.e.,

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)d\phi \ , \tag{5.22}$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} u_0(\phi)\cos k\phi d\phi \ , \ k = 1, 2, \cdots \tag{5.23}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} u_0(\phi)\sin k\phi d\phi \ , \ k = 1, 2, \cdots \ . \tag{5.24}$$

This indicates that $u(r,\theta)$ in (5.21) is the desired solution of the boundary value problem (5.18) with the coefficients a_k and b_k given by (5.22)-(5.24).

Example 5.5. *Find the solution of the following Laplace equation:*

$$\begin{cases} \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0, \quad 0 \le r < 1, \quad -\pi \le \theta < \pi \\ u(1,\theta) = \theta, \quad -\pi \le \theta < \pi \, . \end{cases}$$

and

$$\begin{cases} \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0, \quad 0 \le r < 1, \quad -\pi \le \theta < \pi \\ u(1,\theta) = \delta(\theta), \quad -\pi \le \theta < \pi \, . \end{cases}$$

We can plug the integral expression for a_k and b_k back to (5.21), we get

$$\begin{aligned}
u(r,\theta) &= \sum_{k=0}^{\infty} r^k \left(a_k \cos k\theta + b_k \sin k\theta\right) \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)\,d\phi + \sum_{k=1}^{\infty} \frac{1}{\pi}\int_{-\pi}^{\pi} u_0(\phi) r^k \left(\cos k\phi\, \cos k\theta + \sin k\phi\, \sin k\theta\right)\,d\phi \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)\,d\phi + \sum_{k=1}^{\infty} \frac{1}{\pi}\int_{-\pi}^{\pi} u_0(\phi) r^k \left(\cos k(\theta-\phi)\right)\,d\phi \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)\,d\phi + \sum_{k=1}^{\infty} \frac{1}{\pi}\int_{-\pi}^{\pi} u_0(\phi) Re\left(r^k e^{ik(\theta-\phi)}\right)\,d\phi \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)\,d\phi + \frac{1}{\pi}\int_{-\pi}^{\pi} u_0(\phi) Re\left(\sum_{k=1}^{\infty}(r^k e^{ik(\theta-\phi)})\right)\,d\phi \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)\left(1 + 2Re\left(\frac{re^{i(\theta-\phi)}}{1-re^{i(\theta-\phi)}}\right)\right)\,d\phi \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} u_0(\phi)\frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)}\,d\phi.
\end{aligned}$$

This last expression to represent the solution is called the Poisson formula.

Homeworks.

- pp. 287, 4.1.16
- pp. 287, 4.1.18

5.6 Orthogonal functions

In this section, we introduce some further knowledge on orthogonal functions.

For a given positive function $w(x)$ on $[a, b]$, we define an inner product

$$(f, g)_\omega = \int_a^b \omega(x)\, f(x)\, g(x) dx$$

for any two real functions $f(x)$ and $g(x)$ on $[a, b]$. And $\omega(x)$ will be called a *weight function.* We will often use the following norm:

$$\|f\|_\omega = \Big\{ \int_a^b \omega(x)\, f^2(x)\, dx \Big\}^{\frac{1}{2}} .$$

Definition 5.4 (Weighted orthogonal functions)**.** *Let $f(x)$ and $g(x)$ be two real functions on $[a, b]$. $f(x)$ is said to be orthogonal to $g(x)$ with respect to the inner product $(\cdot, \cdot)_\omega$ if $(f, g)_\omega = 0$.*

A sequence of functions $\{f_k\}_{k=0}^\infty$ is said to be orthonormal with respect to the inner product $(\cdot, \cdot)_\omega$ if the following holds:

$$(f_m, f_n)_\omega = 0 \quad \forall\, m \neq n$$

and each f_k is unitary, i.e.,

$$\|f_k\|_\omega = 1 .$$

• Check if function $\cos x$ is orthogonal to $g(x) = \sin x$ with respect to the inner product $(\cdot, \cdot)_\omega$ for $\omega(x) = 1, x, x^2$.

• Verify that any sequence of orthogonal functions $\{g_k\}_{k=1}^\infty$ on the interval $[a, b]$ are linearly independent.

Now we are going to demonstrate that

Any sequence of linearly independent functions $\{\phi_k\}_{k=0}^\infty$ defined on $[a, b]$ can generate a sequence of functions $\{q_k\}_{k=0}^\infty$ which are orthonormal with respect to the inner product $(\cdot, \cdot)_\omega$.

Gram-Schmidt orthogonalization is one of such orthogonalizing techniques. Below we introduce the Gram-Schmidt orthogonalization.

Gram-Schmidt orthogonalization process.

Given a sequence $\{\phi_k\}_{k=0}^{\infty}$ of linearly independent functions defined on $[a,b]$, we are going to construct a sequence of orthonormal functions $\{q_k\}_{k=0}^{\infty}$ as follows:

0) Set

$$\widetilde{q}_0(x) = \phi_0(x)\,.$$

Normalize $\widetilde{q}_0(x)$:

$$q_0(x) = \frac{\widetilde{q}_0(x)}{\|\widetilde{q}_0\|_\omega}\ ;$$

1) Set

$$\widetilde{q}_1(x) = \phi_1(x) - \alpha_{10}\, q_0(x)\,,$$

choose α_{10} such that

$$(\widetilde{q}_1, q_0)_\omega = \int_a^b \omega(x)\,\tilde{q}_1(x) q_0(x) dx = 0,$$

that gives,

$$\alpha_{10} = (\phi_1, q_0)_\omega = \int_a^b \omega(x)\,\phi_1(x) q_0(x) dx.$$

Normalize $\widetilde{q}_1(x)$:

$$q_1(x) = \frac{\widetilde{q}_1(x)}{\|\widetilde{q}_1\|_\omega}.$$

k) Suppose $q_0, q_1, \cdots, q_k$ are constructed such that

$$(q_i, q_j)_\omega = 0 \quad \forall i \neq j \quad \text{and} \quad \|q_i\|_\omega = 1\ .$$

We then construct q_{k+1} by

$$\widetilde{q}_{k+1}(x) = \phi_{k+1}(x) - \Big\{\alpha_{k+1,0} q_0(x) + \cdots + \alpha_{k+1,k} q_k(x)\Big\}$$

with

$$\alpha_{k+1,i} = (\phi_{k+1}, q_i)_\omega, \quad i = 0, 1, \cdots, k.$$

Normalize q_{k+1}:

$$q_{k+1}(x) = \frac{\widetilde{q}_{k+1}(x)}{\|\widetilde{q}_{k+1}\|_\omega}.$$

Then the sequence $\{q_k\}_{k=0}^{\infty}$ constructed above is an orthonormal sequence, i.e.,

$$(q_i, q_j)_\omega = 0 \quad \forall\ i \neq j\ ; \quad \|q_i\|_\omega = 1\ .$$

Example 5.6. *Given the sequence of polynomials*

$$1, x, x^2, \cdots, x^k, \cdots,$$

on the interval $[-1, 1]$*, use the Gram-Schmidt orthogonalization process to construct an orthonormal sequence of polynomials, and write down the first three constructed polynomials explicitly.*

<u>**Solution**</u> (exercise). The three polynomials are

$$P_0(x) = \frac{1}{\sqrt{2}}\ ,\ P_1(x) = \frac{\sqrt{3}}{\sqrt{2}}\, x\ ,\ P_2(x) = \frac{\sqrt{45}}{\sqrt{8}}\Big(x^2 - \frac{1}{3}\Big)\ .$$

Example 5.7. *Check if the Chebyshev polynomials*

$$T_0(x) = 1,\ T_1(x) = x,\ T_2(x) = 2x^2 - 1,\ T_3(x) = 4x^3 - 3x$$

are orthogonal on $[-1, 1]$ *with respect to the weight function* $w(x) = 1/\sqrt{1-x^2}$.

<u>**Solution**</u>. Use the transformation $x = cos\theta$.

Example 5.8. *Based on a given sequence of functions* $\{\phi_i(x)\}_{k=0}^{\infty}$*, which is orthogonal with respect to the inner product* $(\cdot, \cdot)_\omega$*, use the Gram-Schmidt orthogonalization to construct an orthonormal sequence of functions with respect to* $(\cdot, \cdot)_\omega$*. (*exercise*)*

Example 5.9. *Expand a given function* $f(x)$ *on* $[a, b]$ *in terms of a given orthogonal sequence of functions* $\{\phi_k(x)\}_{k=1}^{\infty}$ *with respect to the inner product* $(\cdot, \cdot)_\omega$.

<u>**Solution**</u>. Let

$$f(x) = \alpha_1\phi_1(x) + \alpha_2\phi_2(x) + \alpha_3\phi_3(x) + \cdots\ .$$

Think about how to find the coefficients $\{\alpha_k\}$. $\sharp$

5.7 Fourier transform

Fourier transforms play a very important role in mathematics, physics and engineering.

5.7.1 Definition and examples

Recall that

$$e^{ikx} = \cos(kx) + i\,\sin(kx).$$

From this expression, we can easily see that the magnitude of k determines the intensity of the oscillation of function $\exp(ikx)$, and k measures the frequencies of the oscillation.

To better understand the relation between the magnitude of k and the oscillation of $\exp(ikx)$, , one may plot and compare the figures of $\sin \pi x$, $\sin 4\pi x$ and $\sin 8\pi x$.

The Fourier series expansion for 2π-periodic functions can be extended to any $2T$-periodic functions by the following scaling method. Given a $2T$-periodic function f, we can rescale it as a 2π-periodic function f' by the following transformation:

$$x' = \frac{\pi x}{T}, \text{ define } f'(x') := f(x).$$

(Notice that f' here does not denote for the derivative of f. It simply means a rescaling of f.) Let us first write down the Fourier series expansion for the 2π-periodic function f':

$$f'(x') = \sum_{k'=-\infty}^{\infty} c_{k'} e^{ik'x'}$$

where

$$c_{k'} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f'(y') e^{-ik'y'}\, dy'$$

We now perform the following transform:

$$y = \frac{Ty'}{\pi},$$

we get

$$\begin{aligned} c_{k'} &= \frac{1}{2\pi} \int_{-T}^{T} f'\left(\frac{\pi y}{T}\right) e^{-ik'\pi y/T}\, d\frac{\pi y}{T} \\ &= \frac{1}{2T} \int_{-T}^{T} f(y) e^{-ik'\pi y/T}\, dy. \end{aligned}$$

We express x' in terms of x again:

$$f(x) = f'\left(\frac{\pi x}{T}\right) = \sum_{k'=-\infty}^{\infty} c_{k'} e^{ik'\pi x/T}$$

Combine the above two, f can be expanded as

$$f(x) = \sum_{k'=-\infty}^{\infty} c_{k'} e^{ik'\pi x/T} = \sum_{k'=-\infty}^{\infty} \frac{1}{2T} e^{ik'\pi x/T} \left[\int_{-T}^{T} f(y) e^{-ik'\pi y/T}\, dy\right]$$

If we take $T \to \infty$, this means that we are considering a general function defined on the whole line. The square bracket becomes

$$\left[\int_{-\infty}^{\infty} f(y) e^{-iky}\, dy\right] := g(k).$$

Here, $k = \frac{k'\pi}{T}$ is a real number. The summation part is an approximation to an integral: when $\Delta k' = 1$, the corresponding $\Delta k = \frac{\pi}{T}$.

$$f(x) = \sum_{k'=-\infty}^{\infty} \frac{1}{2T} e^{ikx} g(k) = \frac{1}{2\pi} \sum_{k'=-\infty}^{\infty} \Delta k e^{ikx} g(k) \to \frac{1}{2\pi} \int_{-\infty}^{\infty} g(k) e^{ikx}\, dk.$$

Thus, we have the following definition.

Definition 5.5. *For a given function $f(x)$ defined on $(-\infty, \infty)$, the Fourier transform of f is a function $\widehat{f}$ depending on* **frequency***:*

$$\widehat{f}(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} dx\,, \quad -\infty < k < \infty\,. \tag{5.25}$$

The inverse Fourier transform of $\widehat{f}(k)$ recovers the original function $f(x)$:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx} dk\,, \quad -\infty < x < \infty\,. \tag{5.26}$$

Example 5.10. *Find Fourier transform of the delta function $f(x) = \delta(x)$.*

<u>**Solution**</u>.

$$\widehat{f}(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} dx = \int_{-\infty}^{\infty} \delta(x) e^{-ikx} dx = 1, \quad \text{for all frequencies } k\ .$$

So the Fourier transform of the delta function is a constant function.

Example 5.11. *Find the Fourier transform of the function:*

$$f(x) = \text{ \textit{square pulse} } = \begin{cases} 1\,, & |x| \le a \\ 0\,, & |x| > a \end{cases} \quad .$$

<u>**Solution**</u>.

$$\widehat{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-ikx}dx = \int_{-a}^{a} e^{-ikx}dx = \frac{2\sin ka}{k} \ .$$

• Think about whether this function $\widehat{f}(k)$ makes sense at $k = 0$.

Example 5.12. *For $a > 0$, find the Fourier transform of the function:*

$$f(x) = \begin{cases} e^{-ax}\,, & x \ge 0 \\ -e^{ax}\,, & x < 0\ . \end{cases}$$

<u>**Solution**</u>. By definition, we have

$$\begin{aligned}\widehat{f}(k) &= \int_{-\infty}^{\infty} f(x)e^{-ikx}dx = \int_{0}^{\infty} e^{-ax-ikx}dx + \int_{-\infty}^{0} -e^{ax-ikx}dx \\ &= \frac{1}{a+ik} - \frac{1}{a-ik} = \frac{-2ik}{a^2+k^2} \ .\end{aligned}$$

• Justify the above process yourself.

Example 5.13. *Find the Fourier transform of*

$$f(x) = \textit{sign function} = \begin{cases} \ 1\,, & x > 0 \\ -1\,, & x < 0\ . \end{cases}$$

<u>**Solution**</u>. We have

$$\widehat{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-ikx}dx = \int_{0}^{\infty} e^{-ikx}dx + \int_{-\infty}^{0} -e^{-ikx}dx.$$

But what is $e^{-ikx}\big|_0^\infty$? It is difficult to know.

To solve this problem, we consider the function

$$f_a(x) = \begin{cases} e^{-ax}\,, & x > 0 \\ -\,e^{ax}\,, & x < 0 \end{cases},$$

it is easy to see that

$$\lim_{a\to 0^+} f_a(x) = f(x)\,,$$

then we can compute as follows:

$$\begin{aligned}\widehat{f}(k) &= \int_{-\infty}^{\infty} f(x)e^{-ikx}dx = \int_{-\infty}^{\infty} \lim_{a\to 0^+} f_a(x)e^{-ikx}dx \\ &= \lim_{a\to 0^+}\int_{-\infty}^{\infty} f_a(x)e^{-ikx}dx = \lim_{a\to 0^+}\frac{-2ik}{a^2+k^2} \\ &= \frac{-2i}{k} = \frac{2}{ik}\ .\end{aligned}$$

Example 5.14. *Find the Fourier transformation of the constant function*

$$f(x) = 1, \quad \forall\, x \in (-\infty,\infty)\ .$$

<u>**Solution**</u>. We have

$$\begin{aligned}\widehat{f}(k) &= \int_{-\infty}^{\infty} e^{-ikx}dx = \int_{0}^{\infty} e^{-ikx}dx + \int_{-\infty}^{0} e^{-ikx}dx \\ &= \lim_{a\to 0^+}\{\int_{0}^{\infty} e^{-ax}e^{-ikx}dx + \int_{-\infty}^{0} e^{ax}e^{-ikx}dx\} \\ &= \lim_{a\to 0^+}\{\frac{1}{a+ik}+\frac{1}{a-ik}\} = \begin{cases} 0 & k\neq 0 \\ ? & k=0\end{cases}\end{aligned}$$

What is $\widehat{f}(0)$? Note that $\widehat{f}(k)$ looks like a delta function. Let $\widehat{f}(k) = \alpha\delta(k)$, then by the inverse Fourier transform we have

$$1 = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(k)e^{ikx}dk = \frac{\alpha}{2\pi}\ ,$$

so

$$\alpha = 2\pi$$

or

$$\widehat{f}(k) = 2\pi\delta(k), \quad -\infty < k < \infty\ .$$

5.7.2 Two identities for Fourier transforms

(1) For a function $f(x)$ on $(-\infty,\infty)$ and its Fourier transform $\widehat{f}(k)$, we have

$$2\pi\int_{-\infty}^{\infty}|f(x)|^2dx = \int_{-\infty}^{\infty}|\widehat{f}(k)|^2dk\,. \tag{5.27}$$

(2) The inner product of any functions f and g satisfies

$$2\pi \int_{-\infty}^{\infty} f(x)\bar{g}(x)dx = \int_{-\infty}^{\infty} \widehat{f}(k)\bar{\widehat{g}}(k)dk$$

where $\bar{g}(x)$ is the conjugate of $g(x)$.

Example 5.15. *Check the relation (5.27) for the following function*

$$f(x) = \begin{cases} e^{-ax}\,, & x > 0 \\ 0\,, & x < 0\ . \end{cases}$$

Solution. We have

$$2\pi \int_{-\infty}^{\infty} |f(x)|^2 dx = 2\pi \int_{0}^{\infty} e^{-2ax}dx = \frac{\pi}{a},$$

while

$$\begin{aligned} \widehat{f}(k) &= \int_{-\infty}^{\infty} f(x)e^{-ikx}dx = \int_{0}^{\infty} e^{-ax}e^{-ikx}dx \\ &= -\frac{1}{a+ik}e^{-ax-ikx}\Big|_0^{\infty} = \frac{1}{a+ik}\ , \end{aligned}$$

therefore

$$\int_{-\infty}^{\infty} \left|\widehat{f}(k)\right|^2 dx = \int_{-\infty}^{\infty} \frac{dk}{|a+ik|^2} = \int_{-\infty}^{\infty} \frac{dk}{a^2+k^2} = \frac{\pi}{a}\ ,$$

that verifies (5.27). Here we have used the transformation $k = a\cos\theta/\sin\theta$. $\sharp$

5.7.3 Important properties of Fourier transform

This subsection discusses some more properties of Fourier transforms.

(1) One can directly verify from definition that for any complex number α,

$$\widehat{\alpha f}(k) = \alpha\widehat{f}(k).$$

(2) One can directly verify from definition that

$$\widehat{f+g}(k) = \widehat{f}(k) + \widehat{g}(k).$$

(3) The Fourier transform of $\frac{df}{dx}$ is $ik\widehat{f}(k)$, i.e.,

$$\widehat{\frac{df}{dx}}(k) = ik\widehat{f}(k)\ .$$

To see this, we use

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(k)e^{ikx}dk$$

to obtain

$$\frac{df}{dx}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} ik\widehat{f}(k)e^{ikx}dk.$$

Comparing with definition

$$\frac{df}{dx}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{\frac{df}{dx}}(k)e^{ikx}dk$$

gives

$$\widehat{\frac{df}{dx}}(k) = ik\widehat{f}(k).$$

(4) The transform of $F(x) = \int_a^x f(x)dx$ is $\frac{\widehat{f}(k)}{ik} + C\,\delta(k)$, i.e.,

$$\widehat{F}(k) = \frac{\widehat{f}(k)}{ik} + C\,\delta(k)$$

To see this, we use

$$\frac{dF(x)}{dx} = f(x),$$

or

$$\frac{d}{dx}\big(F(x) + C\big) = f(x) \quad \forall C \in \mathbb{R}^1.$$

Taking the transform on both sides,

$$ik\big(\widehat{F}(k) + 2C\,\pi\delta(k)\big) = \widehat{f}(k)\ ,$$

this is ,

$$\widehat{F}(k) = \frac{\widehat{f}(k)}{ik} + C\delta(k) \quad \forall\, C \in \mathbb{R}^1\,.$$

(5) The Fourier transform of $F(x) = f(x-d)$ is $e^{-ikd}\widehat{f}(k)$.

In fact, we have

$$\begin{aligned}\widehat{F}(k) =& \int_{-\infty}^{\infty} F(x)e^{-ikx}dx \\ =& \int_{-\infty}^{\infty} f(x-d)e^{-ikx}dx \\ =& \int_{-\infty}^{\infty} f(y)e^{-ik(y+d)}dy \\ =& e^{-ikd}\widehat{f}(k) \ .\end{aligned}$$

(6) The transform of $g(x) = e^{ixd}f(x)$ is $\widehat{f}(k-d)$.

By definition, we have

$$\begin{aligned}\widehat{g}(k) &= \int_{-\infty}^{\infty} g(x)e^{-ikx}dx = \int_{-\infty}^{\infty} f(x)e^{(id-ik)x}dx \\ &= \int_{-\infty}^{\infty} f(x)e^{-ik'x}dx = \widehat{f}(k') = \widehat{f}(k-d) \ .\end{aligned}$$

(7) The convolution of G and h is the function

$$u(x) = \int_{-\infty}^{\infty} G(x-y)h(y)dy \ ,$$

we often write

$$u(x) = (G * h)(x)$$

or

$$u(x) = \Big(G * h\Big)(x) = \int_{-\infty}^{\infty} G(x-y)h(y)dy \, .$$

We now show that

$$\widehat{u}(k) = \widehat{G}(k)\,\widehat{h}(k).$$

In fact, we have

$$\begin{aligned}\widehat{u}(k) =& \int_{-\infty}^{\infty} u(x)e^{-ikx}dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-y)h(y)e^{-ikx}dydx \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-y)h(y)e^{-ikx}dxdy = \int_{-\infty}^{\infty} h(y)\int_{-\infty}^{\infty} G(x-y)e^{-ikx}dxdy \\ =& \int_{-\infty}^{\infty} h(y)e^{-iky}dy\int_{-\infty}^{\infty} G(x')e^{-ikx'}dx' \quad (\text{let } x-y=x') \\ =& \widehat{G}(k)\,\widehat{h}(k) \, , \quad -\infty < k < \infty \, .\end{aligned}$$

5.7.4 Application of Fourier transform for differential equations

Fourier transforms can be applied to solve different types of differential equations. Here we consider one example.

Consider the differential equation

$$-\frac{d^2u}{dx^2} + a^2u = h(x)\ ,\ -\infty < x < \infty\ . \tag{5.28}$$

In order to solve the equation, we apply the Fourier transform

$$\widehat{u}(k) = \int_{-\infty}^{\infty} u(x)e^{-ikx}\, dx$$

to each term of the equation to obtain

$$-(ik)^2\widehat{u}(k) + a^2\widehat{u}(k) = \widehat{h}(k)\ .$$

This gives

$$\widehat{u}(k) = \frac{\widehat{h}(k)}{a^2+k^2}\ . \tag{5.29}$$

Let $G(x)$ be a function such that

$$\widehat{G}(k) = \frac{1}{a^2+k^2},$$

then we know

$$\widehat{u}(k) = \frac{\widehat{h}(k)}{a^2+k^2} = \widehat{G}(k)\widehat{h}(k).$$

By the convolution property, the solution $u(x)$ can be given by

$$u(x) = (G * h)(x) = \int_{-\infty}^{\infty} G(x-y)h(y)dy\,. \tag{5.30}$$

To find $G(x)$, we consider function $f(x) = e^{-a|x|}$. By definition, we have

$$\begin{aligned}
\widehat{f}(k) &= \int_{-\infty}^{\infty} f(x)e^{-ikx}dx \\
&= \int_0^{\infty} e^{-(a+ik)x}dx + \int_{-\infty}^{0} e^{(a-ik)x}dx \\
&= -\frac{1}{a+ik}e^{-(a+ik)x}\Big|_{x=0}^{x=\infty} + \frac{1}{a-ik}e^{(a-ik)x}\Big|_{x=-\infty}^{x=0} \\
&= \frac{1}{a+ik} + \frac{1}{a-ik} = \frac{2a}{a^2+k^2}\ .
\end{aligned}$$

this shows

$$\widehat{\frac{1}{2a}f}(k) = \frac{1}{a^2+k^2}\,,$$

so we have

$$G(x) = \frac{1}{2a}e^{-a|x|}\,.$$

Now we get from (5.30) that

$$u(x) = \frac{1}{2a}\int_{-\infty}^{\infty} e^{-a|x-y|}h(y)dy, \quad -\infty < x < \infty\,. \tag{5.31}$$

♯

• Check if the function $u(x)$ in (5.31) is indeed a solution to the differental equation (5.28).

Homeworks.

- pp. 326, 4.3.2 (a)
- pp. 326, 4.3.3 (a), (b)
- pp. 326, 4.3.4
- pp. 327, 4.3.6
- pp. 327, 4.3.7
- pp. 327, 4.3.10

5.8 *Complex variables and conformal mapping

5.8.1 Basics on complex variables

In this subsection, we shall discuss how to solve some special differential equations by complex methods. We start with an introduction of some basic knowledge on complex variables.

It is best to illustrate a complex number $z = x + i\,y$ in a plane. E.g., the number $z = -\sqrt{2} + \sqrt{2}\,i$ has real part $x = -\sqrt{2}$ and imaginary part $y = \sqrt{2}$, and its distance to the origin is $r = \sqrt{x^2+y^2} = 2$, the same as the absolute value $r = |z| = \sqrt{z\,\bar{z}}$. The angle is given by $\tan\theta = y/x = -1$, so it is 135^o or $3\pi/4$.

Let us look at the reciprocal of $w = 1/z$. We have

$$w = \frac{1}{x + i\,y} = \frac{x - i\,y}{x^2 + y^2}\,. \tag{5.32}$$

So its real part is x/r^2, and its imaginary part is $-y/r^2$. To see the relation between z and $w = 1/z$, one may write z as $z = r\,e^{i\theta}$ in polar coordinates which are much more convenient when multiplications or divisions are studied, then

$$w = \frac{1}{z} = \frac{1}{r\,e^{i\theta}} = \frac{1}{r}\,e^{-i\theta}\,.$$

From this we clearly see that a circle of radius r in the z-plane is mapped to a circle of radius $1/r$ in the w-plane, with the angle θ changed into $-\theta$. That is, moving counterclockwise around one circle takes us clockwise around the other. Using this, we directly find that

$$w = \frac{1}{5}\cos\theta - \frac{1}{5}\,i\,\sin\theta \quad \text{for} \quad z = 3\,\cos\theta + 4\,i\,\sin\theta\,.$$

It is easy to observe the important fact from (5.32) that every circle, whether its center is origin or not, is mapped into another circle by $w = 1/z$.

The unit circle ($r = 1$) in z-plane is special, it will be transformed to the unit circle in w-plane.

Now we look at the special division $z/\bar{z}$. We have

$$\frac{z}{\bar{z}} = \frac{re^{i\theta}}{re^{-i\theta}} = e^{2i\theta} \quad \text{or} \quad \frac{x + iy}{x - iy} = \frac{(x^2 - y^2) + 2ixy}{x^2 + y^2}\,.$$

Again we see the advantage of the polar coordinates. On the right of the above equation, one can not see anything directly. But we see immediately from the left that the magnitude of $z/\bar{z}$ is always 1, and $e^{2i\theta}$ is always on the unit circle.

5.8.2 Analytic functions and Laplace equation

Laplace equation $u_{xx}+u_{yy} = 0$ is a basic partial differential equation in the mathematical modelling. It is surprising that complex numbers may provide great help in solving the equation.

$u(x, y)$ is a real function depending on real numbers x and y. Now, we consider the complex combination $z = x + i\,y$. Then it is amazing to notice that any reasonable function $f(z) = f(x + i\,y)$ is automatically a solution to the Laplace equation. To see this, we first observe that

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} = \frac{\partial f}{\partial z}\,, \quad \frac{\partial f}{\partial y} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial y} = i\,\frac{\partial f}{\partial z}\,. \tag{5.33}$$

This leads to the following fundamental equation for $f(x + i\,y)$:

$$i\,\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\,. \tag{5.34}$$

Further differentiating the two relations in (5.33) gives

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial z^2}\,, \quad \frac{\partial^2 f}{\partial y^2} = -\,\frac{\partial^2 f}{\partial z^2}\,. \tag{5.35}$$

This indicates that $f(x + i\,y)$ satisfies the Laplace equation:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0\,. \tag{5.36}$$

This relation is very simple but very useful. Immediately we know that $f = (x + i\,y)^n$ and $f = e^{x+i\,y}$ are the solutions to the Laplace equation.

Now let us try to get some real solutions to the laplace equation. Let us write

$$f(x + i\,y) = u(x, y) + i\,s(x, y) \tag{5.37}$$

where $u(x, y)$ and $s(x, y)$ are both real functions. (Students should try some simple examples to find out what $u(x, y)$ and $s(x, y)$ are for a given function).

Now we substitute (5.37) in (5.36) to find out that both u and s satisfy laplace equation.

Next, we try to find out some relation between the real part $u(x, y)$ and the imaginary part $s(x, y)$. To do so, we substitute (5.37) into (5.34) to obtain

$$i\Big(\frac{\partial u}{\partial x} + i\,\frac{\partial s}{\partial x}\Big) = \Big(\frac{\partial u}{\partial y} + i\,\frac{\partial s}{\partial y}\Big)\,.$$

Comparing the real and imaginary parts give the following Cauchy-Riemann equation:

$$\frac{\partial u}{\partial x} = \frac{\partial s}{\partial y}\,, \quad \frac{\partial u}{\partial y} = -\frac{\partial s}{\partial x}\,. \tag{5.38}$$

Based on the above results, we introduce an important class of functions – analytic functions:

A function $f(z)$ is analytic at $z = a$ if in a neighborhood of point $z = a$

1. it depends on the combination $z = x + i\,y$ and satisfies $i\partial f/\partial x = \partial f/\partial y$;

2. its real and imaginary parts are connected by the Cauchy-Riemann equation $u_x = s_y$ and $u_y = -s_x$;

3. it is the sum of a convergent power series $c_0 + c_1(z-a) + c_2(z-a)^2 + \cdots$.

5.8.3 Conforming mapping

Consider solving the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{in the region } \Omega\,, \tag{5.39}$$

$$u = u_0 \quad \text{on the boundary } \partial\Omega\,. \tag{5.40}$$

We first discuss how to transform the boundary to simpler form. Consider a transformation

$$X = X(x,y), \quad Y = Y(x,y)\,.$$

Usually after the transformation, the Laplace equation will become more complicated. But if X and Y are the real and imaginary parts of a function $F(z) = F(x+iy)$, situation is completely different. The equation for $U(X,Y)$, in the new variables, is still Laplace equation.

This transformation is called a conforming mapping, when it is given by an analytic function $F(z)$. We have the following result:

Theorem 5.1. *Suppose the combination $X + i\,Y$ is a function F of the combination $x + iy$. Then if $U(X,Y)$ satisfies Laplace equation in the X, Y variables, the corresponding $u(x,y) = U(X(x,y), Y(x,y))$ satisfies Laplace equation in the x, y variables.*

Proof. We can look at U as the real part of some analytic function $f(X + i\,Y)$. But $X + i\,Y$ is $F(x+iy)$. Clearly in the x and y variables, u is the real part of $f(F(z))$. So it must satisfy the Laplace equation.

Example 5.16. *Consider the functions $f(X+iY) = e^{X+iY}$ and $X + iY = (x+iy)^2 = x^2 - y^2 + 2ixy$.*

Discussion. Clearly $f(X+iY)$ is an analytic function of $Z = X + iY$, so its real part $U(X,Y) = e^X \cos Y$ must satisfy the Laplace equation, so does $u = e^{x^2-y^2}\cos 2xy$.

The conforming mapping

$$X = X(x, y), \quad Y = Y(x, y)$$

is invertible if $F' = dF/dz \neq 0$, then the inverse mapping is also conforming.

Consider the region between the x-axis and the line $y = x$ with vertex $(0, 0)$. This region can be transformed to a better domain by a conforming mapping, e.g., to the first quardrant. This can be achieved by the conforming mapping $f(z) = z^2$. If we write $z = re^{i\theta}$, then the conforming mapping $f(z) = z^2$ doubles both the magnitude and the angle.

Using this mapping, we see that rays from the origin are rotated onto other rays, with the angle doubled. Circles around the origin are mapped onto other circles, with radius squared. Only the unit circle stays invariant.

Angle preservation. Conforming mapping has an amazing property: it preserves angles. It doubles any angle with vertex being the origin, but it preserves the angle between any two lines. Consider a triangle with vertices z, $z + \Delta$ and $z + \delta$, with Δ and δ small. With the conforming mapping $f(z) = z^2$, the three vertices become z^2, $z^2 + 2z\Delta$ and $z^2 + 2z\delta$, after ignoring the high order terms. So the original triangle is amplified but its angle at z is not changed. In fact, the straight edges of the original triangle become curved in the w-plane because of the neglected terms Δ^2 and δ^2, but they are nearly straight near z^2.

At the origin, the angle is not preserved, in fact it is doubled. But the mapping fails to be conforming at $z = 0$ since $F'(z) = 2z = 0$.

The mapping $f(z) = z^2$ simplifies the Laplace equation in the wedge formed by the x-axis and the line $y = x$. Let us assume that the mixed boundary conditions are enforced: $u = 0$ on the x-axis, and its derivative $\partial u/\partial n = 0$ on the line $y = x$. Clearly with the mapping, the first condition does not change: $U = 0$ on the line $Y = 0$.

To see how the other boundary condition changes, let N the normal direction on the Y-axis in the w-plane, then with the conforming mapping $X = X(x, y)$ and $Y = Y(x, y)$, the normal direction n on the line $y = x$ will be changed to N since angles are preserved. So the condition $\partial u/\partial n = 0$ on the line $y = x$ is changed to $\partial U/\partial N = 0$ on the Y-axis.

Then one can see that in the new coordinate system (X, Y), the solution is much easier to find. As u solves the Laplace equation, so U also solves the Laplace equation, i.e.,

$$U_{XX} + U_{YY} = 0\,.$$

But noting that the normal N is now in the X-direction, so $\frac{\partial U}{\partial N} = 0$ gives $\frac{\partial U}{\partial X} = 0$, that is, $U(X,Y) == U(Y)$. Then we have from the Laplace equation that $U_{YY} = 0$, or $U = cY + d$. Using the condition $U = 0$ on $Y = 0$, we know $U = cY$. The constant c is not determined by the problem, or it can be determined by the condition at infinity.

The solutions are the straight lines in the w-plane,$U = cY$. Going back to the z-plane, we know $Y = 2xy$, so the solutions $u(x,y)$ are hyperbolas in the z-plane, $u = 2cxy$. One can easily check that this function is indeed the solution to Laplace equation, and it vanishes on the x-axis, and its normal derivative on the line $y = x$ is zero. To see this, we have on the line $y = x$,

$$\begin{aligned}\frac{\partial u}{\partial n} &= \nabla u \cdot n = n_1 \frac{\partial u}{\partial x} + n_2 \frac{\partial u}{\partial y} \\ &= -\frac{1}{\sqrt{2}} 2cy + \frac{1}{\sqrt{2}} 2cx = 0\,.\end{aligned}$$

By Riemann's mapping theorem, every region without holes can be mapped conformingly onto every other such region. But it may be difficult to implement the corresponding conforming mapping.

In summary, here is the way to solve the Laplace equations by conforming mappings:

Conforming mapping tries to separate difficulties with the geometry of the domain from difficulties with the boundary conditions. First, we look for a mapping $F(z)$ such that it can simplify the geometry. Then we try to find a function $U(X,Y)$ that satisfies the new boundary conditions and the Laplace equation in the new coordinate system. Finally get back to the original solution $u(x,y)$ using the conforming mapping.

Due to the specialty of the method, it works only for two dimensions.

5.8.4 Some important conforming mappings

In this subsection we list some important conforming mappings.

(1). $w = f(z) = e^{x+iy}$.

This conforming mapping compresses the whole real axis into positive half the real axis since e^x is always positive. Similarly it maps the entire line $y = \pi$ into negative half the real axis. So the mapping turns the two parallel boundaries ($y = 0$ and $y = \pi$) of an infinite horizontal strip into a single line (the x-axis). Points inside the strip go above this line in the w-plane, by noting that $w = e^x(\cos y + i \sin y)$.

(2). $w = (az + b)/(cz + d)$.

This conforming mapping takes every circle into another circle; see the textbook for details.

(3). $w = (z + 1/z)/2$.

To see the effect of this conforming mapping, we consider $z = e^{i\theta}$ lying on the unit circle. Then we have

$$w = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) = \cos\theta\,.$$

So the whole unit circle goes to one part of a line: $-1 \leq w \leq 1$. The points outside the unit circle fill the rest of the w-plane, the same for the points inside the circle.

6 Numerical methods

6.1 Symmetric positive definite matrices

Definition 6.1 (Symmetric matrix)**.** *A matrix A is called symmetric if $A^T = A$.*

Think about what conclusions you can draw for a symmetric matrix !

Definition 6.2 (Symmetric positive definite matrix)**.** *A matrix A is called symmetric and positive definite if A is symmetric and*

$$x^T A\, x > 0 \quad \forall\, x \neq 0.$$

Think about what conclusions you can draw for a symmetric positive matrix A.

A direct observation is that the diagonal entries a_{ii} of A must be positive, **why ?**

Definition 6.3 (Eigenvalue and eigenvector)**.** *For a given $n \times n$ matrix A, if there exists a non-zero vector x such that*

$$Ax = \lambda\, x$$

for some number λ (maybe real or complex), then λ is called an eigenvalue of the matrix A and x is called a corresponding eigenvector of A.

• Think about if the eigenvectors associated with one eigenvalue are unique.

• If there are two eigenvectors of the matrix A associated with one eigenvalue λ, check if any combination of the two eigenvectors is still an eigenvector.

Lemma 6.1. *If A is a $n \times n$ real symmetric matrix, then all its eigenvalues are real numbers. And A always has real eigenvectors.*

Proof. Try to check this property yourself by definition !

Lemma 6.2. *A matrix A is symmetric positive definite if and only if A is symmetric and all its eigenvalues are positive.*

Proof. If A is symmetric positive definite, and λ_i is the i-th eigenvalue of A, i.e., we have for some $x \neq 0$ that

$$Ax = \lambda_i x\,.$$

But by deinition, $x^T Ax > 0$, this leads to

$$\lambda_i x^T x > 0.$$

So we must have $\lambda_i > 0$.

Think about the proof of the other half of the proposition !

6.2 Least-squares solutions to general linear algebraic systems

We now consider how to solve the following general system of linear algebraic equations:

$$Ax = b, \tag{6.41}$$

where A is a $m \times n$ matrix with $m > n$. Clearly this system is not likely to have a solution as the number of equations is larger than the number of unknowns.

When *will this system have a solution x ?*

To answer this question, let us write A as

$$A = (a_1, a_2, \cdots, a_n),$$

then we can write (6.41) into

$$b = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n \,.$$

This indicates

The system (6.41) has a solution only when b lies in the subspace spanned by the column vectors of A.

And in this case, think about

Under **what conditions** *will the solutions be unique ?*

Below, we consider the most general case:

b does not lie in the subspace spanned by the column vectors of A.

We know that the system (6.41) has no solutions in this case.

What will then be some reasonable solutions we should look for ?

In practical applications, it is meaningful to find some vector x which minimizes the error $(Ax - b)$ in some sense. One of the most popular solutions of this kind is called the **least-squares solution**.

The *least-squares solution* of the system $Ax = b$ is the vector x that minimizes the error $(Ax - b)$, namely,

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 \tag{6.42}$$

• To find the least-squares solution, we first discuss some property of the matrix A^TA. Let A be a $m \times n$ matrix, and all its columns are linearly independent. Then

$$\boxed{A^TA \textbf{ must be symmetric positive definite.}}$$

Clearly, A^TA is symmetric. And for any $x \in \mathbb{R}^n$ and $x \neq 0$, we have

$$x^T(A^TA)x = (Ax)^T(Ax) > 0. \tag{6.43}$$

Therefore A^TA is positive definite. If (6.43) is not true, then $(Ax)^T(Ax) = 0$, or $Ax = 0$. That means the columns of A are linearly dependent, a contradiction. ♯

• Further, we look at a simpler minimization problem:

$$\min_{x \in \mathbb{R}^n} \left\{ x^TBx - 2x^Tb \right\}$$

where B is a symmetric possitive definite $n \times n$ matrix. Let

$$J(x) = x^TBx - 2x^Tb, .$$

We next verify that $J(x)$ takes its minimum at the solution x of the equation $Bx = b$. In fact, we easily see that for any $y \in \mathbb{R}^n$,

$$\begin{aligned} J(y) - J(x) &= (y^TBy - 2y^Tb) - (x^TBx - 2x^Tb) \\ &= (y^TBy - 2y^Tb) - (x^TBx - 2x^TBx) \\ &= (y^TBy - 2y^TBx) + x^TBx \\ &= (y - x)^TB(y - x) \\ &\geq 0\,, \end{aligned}$$

so $J(x)$ takes its minimum at the solution to $Bx = b$.

In summary, we have

For any given vector $b \in \mathbb{R}^n$ and any symmetric possitive definite $n \times n$ matrix B, the minimizer of the problem

$$\min_{x \in \mathbb{R}^n} \left\{ x^T B x - 2x^T b \right\}$$

is the unique solution of the system of linear algebraic equations

$$Bx = b\,.$$

The converse is also true.

- Finally, let us find the least-squares solution (6.42). We can write

$$\begin{aligned} \|Ax - b\|^2 &= (Ax-b)^T(Ax-b) \\ &= x^T(A^TA)x - 2x^TA^Tb + b^Tb\,. \end{aligned}$$

For the minimization, the constant term b^Tb plays no role. So using the previously discussed results, we know the minimizer for $\|Ax-b\|^2$ is the solution $x \in \mathbb{R}^n$ such that

$$A^TAx = A^Tb\,,$$

that is,

$$x = (A^TA)^{-1}A^Tb$$

is the least-squares solution to the system $Ax = b$. $\sharp$

Cholesky factorization for least-squares solutions. As we have known earlier, to find the least-squares solution of the system $Ax = b$, where A is a $m \times n$ matrix with linearly indepedent columns, one needs to solve the normal equation

$$A^TAx = A^Tb\,.$$

Since A^TA is a symmetric and positive definite matrix, we may first find its Cholesky factorization

$$A^TA = LL^T,$$

where L is a lower triangular matrix. Then solving the normal equation $A^TAx = A^Tb$ is equivalent to solving the following two triangular systems

$$Lc = A^Tb, \quad L^Tx = c\,.$$

- Construct or find a few examples from the textbooks to practice how to find the least-squares solutions using the Cholesky factorization.

Exercise. For the least-squares solution $x = (A^TA)^{-1}A^Tb$, find out the relations between the error vector $(Ax - b)$ and the column vectors of A.

6.3 *Nonlinear equations and Newton's method

In this and next subsections, we will introduce some iterative methods for solving linear and nonlinear algebraic system of equations, some simple boundary value and initial-value problems. As this course is an introductory course, we will have no time to discuss each numerical method very deeply.

We will start with the powerful and effective Newton's methods for a single nonlinear equation of one variable.

6.3.1 Newton's method for a scalar nonlinear equation

Consider solving the nonlinear equation

$$g(x) = 0, \quad x \in R^1 \tag{6.44}$$

where $g(x)$ is a nonlinear function in $\mathbb{R}^1$. The point x^* is called a solution of the equation (6.44) if the condition $g(x^*) = 0$ holds, i.e., x^* is the point where the graph of $g(x)$ crosses the horizontal axis. It is easy to see that the equation $g(x) = 0$ may have more than one solution. But the equation often has a unique solution when it is restricted on a small interval.

Let x^* be a solution of the equation $g(x) = 0$, i.e.,

$$g(x^*) = 0$$

Our subsequent task is to study how to find the solution x^*.

Newton's method. Newton's method is an iterative method. It starts with a given approximate solution x^0 of x^*. Then it tries to find a better approximation x^1 of the form:

$$x^1 = x^0 + d^0.$$

To find d^0, we expand $g(x^1)$ at x^0 by Taylor expansion:

$$0 = g(x^1) \approx g(x^0) + g'(x^0)d^0 .$$

Let $g(x^0) + g'(x^0)d^0 = 0$, we obtain

$$d^0 = -\frac{g(x^0)}{g'(x^0)} \ ,$$

so the next approximation x^1 is given by

$$x^1 = x^0 - \frac{g(x^0)}{g'(x^0)} \ .$$

Repeating this procedure, we have

$$x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)} \ , \quad k = 0, 1, 2, \cdots$$

This iterative method is called the *Newton's method.*

• Note that the Newton's method is derived using the Taylor expansion, so the initial guess can not be too much away from the exact solution. Such iterative methods are called *iterative methods with local convergence.*

• Geometrically, the Newton's method is equivalent to finding the intersection of the tangent line of $g(x)$ with x-axis at $x = x^0, x^1, x^2, \cdots$. Try to understand this process and derive the Newton's method using this approach.

Example 6.1. *Solve the equation*

$$g(x) = x^2 - 10 = 0$$

by the Newton's method with an initial guess $x^0 = 3$.

Solution. Clearly this example is too simple. But we use this simple example just to demonstrate how fast the Newton's method may converge.

By using the computer, we know the exact solution is

$$x^* = \sqrt{10} = 3.167227766016838.$$

But by Newton's method, we have

$$\begin{aligned}
x^0 =& 3.0, \\
x^1 =& x^0 - \frac{g(x^0)}{g'(x^0)} = x^0 - \frac{(x^0)^2 - 10}{2x^0} \approx 3.16 \\
x^2 =& x^1 - \frac{g(x^1)}{g'(x^1)} \approx 3.1622 \\
x^3 =& 3.16227766016 \\
x^4 =& 3.16227766016838 \ .
\end{aligned}$$

We see that Newton's method converges **very fast**.

6.3.2 Newton's method for system of nonlinear equations

Consider solving the nonlinear system of equations

$$\begin{cases} g_1(x_1, x_2, \cdots, x_n) = 0 \ , \\ g_2(x_1, x_2, \cdots, x_n) = 0 \ , \\ \cdots\cdots \\ g_n(x_1, x_2, \cdots, x_n) = 0 \end{cases} \tag{6.45}$$

where $g_i(x_1, x_2, \cdots, x_n)$ are all nonlinear functions. For convenience, we often write this system as the following vector-form:

$$g(x) = 0 \tag{6.46}$$

where x and $g(x)$ are the vectors given by

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad g(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \cdots \\ g_n(x) \end{pmatrix}.$$

If $g_i(x)$ are linear functions, e.g.,

$$g_i(x_1, x_2, \cdots, x_n) = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i, \quad i = 1, 2, \cdots, n,$$

then we can write equation (6.46) into

$$Ax = b \tag{6.47}$$

with

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

So the linear system of equations (6.47) is a special case of the nonlinear system (6.46).

To solve the nonlinear system (6.46), we use the iterative Newton's method again. Suppose we have an initial guess x^0, then we want to find a new approximation x^1 of the form

$$x^1 = x^0 + d^0 \ .$$

To find d^0, expand $g(x^1)$ at x^0 by the Taylor series:

$$g(x^1) \approx g(x^0) + ? \tag{6.48}$$

To see the ? term in (6.48), we define a function of one variable t:

$$f(t) = g(x^0 + t(x^1 - x^0)) \ ,$$

then by the Taylor expansion,

$$f(t) \approx f(0) + f'(0)\, t \, . \tag{6.49}$$

But

$$f'(t) = \frac{d}{dt} \begin{pmatrix} g_1(x^0 + t(x^1 - x^0)) \\ \cdots \\ g_n(x^0 + t(x^1 - x^0)) \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x^1 - x^0)_1 + \cdots + \frac{\partial g_1}{\partial x_n}(x^1 - x^0)_n \\ \cdots \\ \frac{\partial g_n}{\partial x_1}(x^1 - x^0)_1 + \cdots + \frac{\partial g_n}{\partial x_n}(x^1 - x^0)_n \end{pmatrix}$$

So we have

$$f'(0) = J(x^0)\,(x^1 - x^0)$$

where J is called the *Jacobian matrix* at x_0, and it is given by

$$J(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \frac{\partial g_1}{\partial x_2}(x) & \cdots & \frac{\partial g_1}{\partial x_n}(x) \\ \frac{\partial g_2}{\partial x_1}(x) & \frac{\partial g_2}{\partial x_2}(x) & \cdots & \frac{\partial g_2}{\partial x_n}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_n}{\partial x_1}(x) & \frac{\partial g_n}{\partial x_2}(x) & \cdots & \frac{\partial g_n}{\partial x_n}(x) \end{pmatrix} \tag{6.50}$$

Now letting $t = 1$ in (6.49), we obtain

$$0 = g(x^1) \approx g(x^0) + J(x^0)(x^1 - x^0) \ ,$$

setting $g(x^0) + J(x^0)(x^1 - x^0) = 0$, we derive

$$x^1 = x^0 - J(x^0)^{-1} g(x^0) \ . \tag{6.51}$$

This gives the Newton's method as follows:

$$x^{k+1} = x^k - J(x^k)^{-1} g(x^k), \quad k = 0, 1, 2, \cdots \tag{6.52}$$

Example 6.2. *Use the Newton's method to find a solution of the system*

$$\begin{cases} x_1x_2 = x_3^2 + 1 \ , \\ x_1x_2x_3 + x_2^2 = x_1^2 + 2 \ , \\ e^{x_1} + x_3 = e^{x_2} + 3 \end{cases} \tag{6.53}$$

with an initial guess $(x_1^0, x_2^0, x_3^0) = (1, 1, 1)$.

Solution. Let

$$\begin{aligned} g_1(x_1, x_2, x_3) &= x_1x_2 - x_3^2 - 1 \\ g_2(x_1, x_2, x_3) &= x_1x_2x_3 + x_2^2 - x_1^2 - 2 \\ g_3(x_1, x_2, x_3) &= e^{x_1} + x_3 - e^{x_2} - 3 \ , \end{aligned}$$

then the system (6.53) can be written as

$$g(x) = 0$$

where x and $g(x)$ are the vectors given by

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} , \quad g(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \cdots \\ g_n(x) \end{pmatrix} .$$

It is easy to compute the Jacobian matrix of $g(x)$:

$$J(x) = \begin{pmatrix} x_2 & x_1 & -2x_3 \\ x_2x_3 - 2x_1 & x_1x_3 + 2x_2 & x_1x_2 \\ e^{x_1} & -e^{x_2} & 1 \end{pmatrix} ,$$

then the Newton's method gives

$$\begin{aligned} &x^1 = x^0 - J(x^0)^{-1}g(x^0) \ , \\ &x^2 = x^1 - J(x^1)^{-1}g(x^1) \ , \\ &\cdots \end{aligned}$$

Using the computer, we obtain

k	x_1^k	x_2^k	x_3^k
0	1.0000000	1.0000000	1.0000000
1	2.1893261	1.5984752	1.3939006
2	1.8462202	1.4380561	1.2866010
3	1.7780014	1.4222229	1.2421980
4	1.7774387	1.4237095	1.2378039
5	1.7776553	1.4239426	1.2374951
6	1.7776707	1.4239593	1.2374728
7	1.7776718	1.4239605	1.2374712

6.4 Classical iterative methods for general linear systems

In this section, we consider solving the linear system of equations

$$A\,x = b \tag{6.54}$$

by some classical iterative methods. Here A is a $n \times n$ matrix and $b \in R^n$.

6.4.1 Splitting methods

A very important class of iterative methods for solving the equation (6.54) is based on the following splitting of the matrix A:

$$A = M - N \tag{6.55}$$

Usually, M is the major part and N is the minor part. Using this splitting, the system $Ax = b$ can be equivalently written as

$$Mx = Nx + b.$$

This suggests the following iterative method:

Given an initial guess x^0, and find $x^1, x^2,\ x^3,\ \cdots$ by solving the following system of equations

$$Mx^{k+1} = Nx^k + b \tag{6.56}$$

where M is a $n \times n$ matrix to be chosen such that the system (6.56) is much easier to solve than the system (6.54). For example, the system (6.56) will be easily solvable if M is a diagonal matrix, a upper triangular or lower triangular matrix.

One can check directly that

If the sequence $\{x^k\}_{k=1}^{\infty}$ generated by (6.56) converges to some limit x^*, then x^* is the solution of $A\,x = b$.

6.4.2 Jacobi method

There are many different choices of the matrix M in the splitting (6.55). We will introduce three most frequently used iterative methods of the form (6.56).

<u>Jacobi method</u>. Given an initial approximate solution of $A\,x = b$:

$$x^0 = (x_1^0,\ x_2^0,\ \cdots,\ x_n^0)^T.$$

The Jacobi method will generate a sequence $\{x^k\}_{k=1}^{\infty}$ by solving

$$\begin{cases} a_{11}x_1^{k+1} + a_{12}x_2^k + \cdots + a_{1n}x_n^k = b_1 & \text{for} \quad x_1^{k+1}\ , \\ a_{21}x_1^k + a_{22}x_2^{k+1} + \cdots + a_{2n}x_n^k = b_2 & \text{for} \quad x_2^{k+1}\ , \\ \cdots & \\ a_{n1}x_1^k + a_{n2}x_2^k + \cdots + a_{nn}x_n^{k+1} = b_n & \text{for} \quad x_n^{k+1}\ . \end{cases} \tag{6.57}$$

Let D be the diagonal matrix

$$D = \text{ diag } (A) = \begin{pmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{pmatrix}$$

Then we can write (6.57) as

$$Dx^{k+1} + (A - D)x^k = b \tag{6.58}$$

or equivalently as

$$Dx^{k+1} = (D - A)x^k + b\,. \tag{6.59}$$

That is, the Jacobi's method is a special case of form (6.56) with $M = $ diag (A).

Example 6.3. *Solve the system*

$$A\,x = b$$

by the Jacobi method. Here

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\ , \quad b = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

<u>Solution</u>. We have

$$D = \text{diag } (A) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\ , \quad D - A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then the Jacobi's method is

$$Dx^{k+1} = (D - A)x^k + b\ ,$$

or

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1^{k} \\ x_2^{k} \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix} ,$$

or

$$x^{k+1} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} x^k + \begin{pmatrix} \frac{1}{2} \\ -1 \end{pmatrix} .$$

If we take the initial guess $x^0 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$, then we have

k	0	1	2	3	4	5	6	10
x^k	$\begin{pmatrix} 5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 3.0 \\ 1.5 \end{pmatrix}$	$\begin{pmatrix} 1.25 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.75 \\ -0.375 \end{pmatrix}$	$\begin{pmatrix} 0.3125 \\ -0.625 \end{pmatrix}$	$\begin{pmatrix} 0.1875 \\ -0.8438 \end{pmatrix}$	$\begin{pmatrix} 0.0781 \\ -0.9062 \end{pmatrix}$	$\begin{pmatrix} 0.0049 \\ -0.9941 \end{pmatrix}$

6.4.3 Gauss-Seidel method

Gauss-Seidel method. Given an initial approximate solution:

$$x^0 = (x_1^0,\ x_2^0,\ \cdots,\ x_n^0)^T$$

to the solution of $A\,x = b$. The Gauss-Seidel method generates the sequence $\{x^k\}_{k=1}^{\infty}$ by solving the following system

$$\left\{ \begin{array}{llllllllll} a_{11}x_1^{k+1} & + & a_{12}x_2^{k} & + & a_{13}x_3^{k} & + \cdots + & a_{1n}x_n^{k} & = b_1 & \text{for } x_1^{k+1} , \\ a_{21}x_1^{k+1} & + & a_{22}x_2^{k+1} & + & a_{23}x_3^{k} & + \cdots + & a_{2n}x_n^{k} & = b_2 & \text{for } x_2^{k+1} , \\ a_{31}x_1^{k+1} & + & a_{32}x_2^{k+1} & + & a_{33}x_3^{k+1} & + \cdots + & a_{3n}x_n^{k} & = b_3 & \text{for } x_3^{k+1} , \\ \cdots & \cdots & & & & & & & \\ a_{n1}x_1^{k+1} & + & a_{n2}x_2^{k+1} & + & a_{n3}x_3^{k+1} & + \cdots + & a_{nn}x_n^{k+1} & = b_n & \text{for } x_n^{k+1} . \end{array} \right. \tag{6.60}$$

Let us decompose

$$A = D + L + U,$$

where $D =$ diag (A), $L =$ the lower triangular of A and $U =$ the upper triangular of A. Then (6.60) can be written as

$$Lx^{k+1} + Dx^{k+1} + Ux^k = b \tag{6.61}$$

or

$$(D+L)x^{k+1} = (D+L-A)x^k + b.$$

That is, the Gauss-Seidel method is a special case of form (6.56) with $M = D + L$.

Example 6.4. *Solve the system*

$$A\,x = b$$

by the Gauss-Seidel method. Here we have

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

Solution. We have

$$D = \text{ diag } (A) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}$$

and

$$U = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \ .$$

Then the Gauss-Seidel method is

$$(D+L)x^{k+1} = -Ux^k + b$$

or

$$\begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix} x^{k+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x^k + \begin{pmatrix} 1 \\ -2 \end{pmatrix} \ .$$

If we take the initial guess $x^0 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$, then the Gauss-Seidel method gives the following result:

k	0	1	2	3	4	5	6	7
x^k	$\begin{pmatrix} 5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 3.0 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.75 \\ -0.625 \end{pmatrix}$	$\begin{pmatrix} 0.1875 \\ -0.9062 \end{pmatrix}$	$\begin{pmatrix} 0.0469 \\ -0.9766 \end{pmatrix}$	$\begin{pmatrix} 0.0117 \\ -0.9941 \end{pmatrix}$	$\begin{pmatrix} 0.0029 \\ -0.9985 \end{pmatrix}$	$\begin{pmatrix} 0.0007 \\ -0.9996 \end{pmatrix}$

Comparing this result with the previous result obtained by the Jacobi method, we can see that the Gauss-Seidel method converges about twice as fast as the Jacobi method.

6.4.4 *Successive overrelaxation method

Successive overrelaxation method (often called the SOR method).

Given an approximate solution $x^k = (x_1^k, x_2^k, \cdots, x_n^k)^T$ of $A\,x = b$. To find the next approximate solution

$$x^{k+1} = (x_1^{k+1}, x_2^{k+1}, \cdots, x_n^{k+1})^T,$$

we solve

$$\left\{\begin{array}{l} a_{11}x_1 + a_{12}x_2^k + \cdots + a_{1n}x_n^k = b_1 \quad \text{for} \quad x_1 = \bar{x}_1^{k+1} \ , \\ \qquad x_1^{k+1} = x_1^k + \omega(\bar{x}_1^{k+1} - x_1^k) \ ; \\ a_{21}x_1^{k+1} + a_{22}x_2 + a_{23}x_3^k + \cdots + a_{2n}x_n^k = b_2 \quad \text{for} \quad x_2 = \bar{x}_2^{k+1} \ , \\ \qquad x_2^{k+1} = x_2^k + \omega\,(\bar{x}_2^{k+1} - x_2^k) \ ; \\ \qquad \cdots\cdots \\ a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + \cdots + a_{n,n-1}x_{n-1}^{k+1} + a_{nn}x_n = b_n \quad \text{for} \quad x_n = \bar{x}_n^{k+1} \ , \\ \qquad x_n^{k+1} = x_n^k + \omega\,(\bar{x}_n^{k+1} - x_n^k) \ . \end{array}\right. \tag{6.62}$$

If we decompose

$$A = D + L + U,$$

where $D = \text{diag}\,(A)$, $L =$ the lower triangular part of A and $U =$ the upper triangular part of A, then the iteration (6.62) can be written as

$$Lx^{k+1} + D\bar{x}^{k+1} + Ux^k = b \ , \tag{6.63}$$

$$x^{k+1} = x^k + \omega\,(\bar{x}^{k+1} - x^k) \ . \tag{6.64}$$

We easily see from (6.64) that

$$\bar{x}^{k+1} = \frac{1}{\omega}\big(x^{k+1} + (\omega - 1)x^k\big) \ ,$$

substituting it into (6.63), we derive

$$(L + \frac{1}{\omega}D)x^{k+1} + \frac{1}{\omega}\big(\omega\,U + (\omega - 1)D\big)x^k = b$$

or

$$(L + \frac{1}{\omega}D)x^{k+1} = \big(\frac{1}{\omega}D - (D + U)\big)x^k + b\,. \tag{6.65}$$

Clearly, this is still a special case of (6.56) with

$$M = L + \frac{1}{\omega}D$$

and we know in this case that

$$M - A = (L + \frac{1}{\omega}D) - (D + L + U) = \frac{1}{\omega}D - (D + U) \ .$$

Remark. The parameter ω in the SOR method is often called a *relaxation parameter*, and it takes values in the interval $0 < \omega < 2$. The SOR method reduces to the Gauss-Seidel method when $\omega = 1$.

• Find some examples of linear algebraic systems and solve them using the Jacobi method, Gauss-Seidel method and the successive over-relaxation method. Then compare the effectiveness of the three iterative methods.

Homeworks.

- pp. 424, 5.3.1
- pp. 424, 5.3.2
- pp. 424, 5.3.3
- pp. 424, 5.3.4

6.5 *Orthogonalization and eigenvalue problems

We first recall some *basic concepts*. Two given two vectors $a, b \in \mathbb{R}^n$ are said to be *orthogonal* if

$$a^T b = \sum_{i=1}^{n} a_i b_i = 0 \ .$$

The *norm* of a vector $a \in \mathbb{R}^n$ is defined as

$$\|a\| = (a^T a)^{1/2} = \Big\{ \sum_{i=1}^{n} a_i^2 \Big\}^{1/2} \ .$$

A set of vectors $q_1, q_2, \cdots, q_n$ are called to be *orthonormal* if

$$q_i^T q_j = \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j \ . \end{cases}$$

And in this case, any vector $b \in \mathbb{R}^n$ can be expressed as

$$b = \sum_{i=1}^{n} \alpha_i q_i \, ,$$

and it is easy to see $\alpha_i = q_i^T b$, $i = 1, 2, \cdots, n$.

For a given $n \times n$ matrix Q, it is called *orthonormal* if

$$Q^T Q = I \ .$$

For an orthonormal matrix Q, we easily see

$$Q^{-1} = Q^T .$$

Furthermore, we can show the following properties:

1. A $n \times n$ orthonormal matrix Q preserves the norm of a vector, namely

$$\|Q\, x\| = \|x\| \quad \forall x \in R^n \ .$$

This is clear since

$$\|Q\, x\|^2 = (Q\, x)^T (Q\, x) = x^T (Q^T Q)\, x = x^T x = \|x\|^2 \ .$$

2. For any $n \times n$ orthonormal matrix Q, we have

$$Q^T Q = Q Q^T = I \ ,$$

from which we know the column vectors of Q are orthonormal, and so are the row vectors.

3. The solution of an orthonormal system is easy to compute. Consider

$$Q\,x = b$$

where Q is a $n \times n$ orthonormal matrix, $b \in R^n$. It is easy to find its solution

$$x = Q^T b.$$

6.5.1 *Revisit of least-squares solutions

Consider the linear system

$$A\,x = b \tag{6.66}$$

where A is a $m \times n$ matrix and $b \in \mathbb{R}^m$, m and n are positive integers and $m > n$.

Since the number of equations in (6.66) is larger than the number of unknowns, so the system (6.66) may not have a solution.

Assume that rank $(A) = n$, i.e., the columns of A are linearly independent. We have

$$\boxed{\textbf{If rank } (A) = n\textbf{, then } A^T A \textbf{ is nonsingular.}}$$

We check this by contradiction. If $A^T A$ is singular, then there exists a nonzero $x \in \mathbb{R}^n$ such that

$$A^T A\,x = 0.$$

Then we have

$$x^T A^T A\,x = (Ax)^T (Ax) = 0,$$

or

$$A\,x = 0,$$

i.e., the columns of A are linearly dependent. This is a contradiction. ♯

Now we discuss **how to find** the least-squares solution to the system (6.66). Recall that the least-squares solution is the solution of the following equation:

$$A^T A\, x = A^T b\,. \tag{6.67}$$

This equation is called the *normal equation* of $A\,x = b$. As $A^T A$ is nonsingular, so the normal equation (6.67) has a unique solution

$$x = (A^T A)^{-1} A^T b\ \ . \tag{6.68}$$

As we know that the least squares solution may not be a solution of the equation $A\,x = b$.

Orthonormal systems. If A is a $m \times n$ orthonormal matrix, i.e., $A^T A = I$, then the least squares solution of $A\,x = b$ reduces to

$$x = A^T b\ .$$

So in this case, the normal equation is extremely easy to solve.

In the general case where A is not an orthonormal matrix, it is difficult to solve the system since $(A^T A)^{-1}$ is not easy to calculate.

6.5.2 *Solve the normal equation by Gram-Schmidt orthogonalizing process

Consider the normal equation

$$A^T A\, x = A^T b\,. \tag{6.69}$$

We next discuss how to solve the system (6.69) by Gram-Schmidt orthogonalizing process. To do so, we first orthogonalize the column vectors of A. Write A as

$$A = (a_1, a_2, \cdots, a_n), \quad \text{with} \quad a_j \in \mathbb{R}^m\,.$$

Gram-Schmidt orthogonalization.

Step 1 Set $\tilde{q}_1 = a_1$; normalize $\tilde{q}_1$ as $q_1 = \frac{\tilde{q}_1}{\|\tilde{q}_1\|}$.

Step 2 Set

$$\tilde{q}_2 = a_2 - \alpha_{12}\, q_1$$

choose α_{12} such that

$$\tilde{q}_2^T q_1 = 0, \quad \text{that is,} \quad \alpha_{12} = q_1^T a_2\,.$$

normalize $\tilde{q}_2$ as $q_2 = \frac{\tilde{q}_2}{\|\tilde{q}_2\|}$.

Step k Suppose $q_1, q_2, \cdots, q_{k-1}$ are constructed such that they are orthonormal. Then set

$$\tilde{q}_k = a_k - \big\{\alpha_{1k} q_1 + \alpha_{2k} q_2 + \cdots + \alpha_{k-1,k}\, q_{k-1}\big\},$$

and choose $\alpha_{1k} \cdots, \alpha_{k-1,k}$ such that

$$\tilde{q}_k^T q_j = 0 \ , \ j = 1, 2, \cdots, k-1 \ ,$$

that gives

$$\alpha_{jk} = q_j^T a_k \ , \quad j = 1, 2, \cdots, k-1.$$

normalize $\tilde{q}_k$ as $q_k = \frac{\tilde{q}_k}{\|\tilde{q}_k\|}$.

In this process, the n vectors $q_1, q_2, \cdots, q_n$ constructed above will be a set of orthonormal vectors. And we can write this process as follows:

$$\begin{aligned} &a_1 = r_{11} q_1 \ , \\ &a_2 = r_{12} q_1 + r_{22} q_2 \ , \\ &\cdots \\ &a_k = r_{1k} q_1 + r_{2k} q_2 + \cdots + r_{k-1,k}\, q_{k-1} + r_{kk}\, q_k \ , \\ &\cdots \end{aligned}$$

Equivalently, we can write this as

$$A = \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \cdots & \vdots \\ & & & r_{nn} \end{pmatrix} \ \equiv \ QR \ .$$

Therefore, the Gram-Schmidt orthogonalizing process is also called *the QR factorization* of the matrix A. As $q_1, q_2, \cdots, q_n$ are orthonormal, so we have

$$Q^T Q = I_{n\times n} \quad \text{(Note } Q \text{ is not a square matrix, but } R \text{ is !)} .$$

Solve the normal equation. Substituting the QR decomposition of A into the normal equation (6.69), we obtain

$$(QR)^T (QR)\, x = (QR)^T b \ ,$$

that is,

$$R^T R\, x = R^T Q^T b,$$

but R is a square matrix and nonsingular (**why ?**), therefore we have

$$R\,x = Q^T b\,. \tag{6.70}$$

Note that R is a upper triangular matrix, so (6.70) is easy to solve, and the solution of (6.70) is the least squares solution of the normal equation (6.69), that is,

$$x = (A^T A)^{-1} A^T b = R^{-1} Q^T b.$$

Example 6.5. *Find the QR factorization of the matrix*

$$A = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

and use it to find the least squares solution of the equation

$$A\,x = b$$

where $b = (2, -1, 2)^T$.

Solution. Practice yourself !

6.5.3 *Eigenvalue problems

We shall introduce two effective methods for finding the eigenvalues of a given matrix: power method and QR iterative method.

Let A be a $n \times n$ real matrix[10], and has n eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ associated with n eigenvectors $x_1, x_2, \cdots, x_n$. Then we have $AX = X\Lambda$, where $X = (x_1, x_2, \cdots, x_n)$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$.

We assume that

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|\,,$$

then the *power method* can help us find the eigenvalue with largest magnitude, that is, find the eigenvalue λ_1.

[10] Power method is also valid for complex matrices.

Power Method. Given an initial vector $v^{(0)} \in \mathbb{R}^n$, generate a sequence of vectors $v^{(k)}$ as follows:

```
FOR  $k = 1, 2, 3, \cdots, $DO
     $z^{(k)} = Av^{(k-1)}$;
     $\lambda^{(k)} = z_i^{(k)}$   where $|z_i^{(k)}| = \|z^{(k)}\|_\infty$
     $v^{(k)} = z^{(k)}/\lambda^{(k)}$
END
```

This iteration is based on the following observation. If $v^{(0)} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ and $a_1 \neq 0$, then it follows that

$$A^k v^{(0)} = a_1\lambda_1^k \Big[x_1 + \sum_{j=2}^{n} \frac{a_j}{a_1}\left(\frac{\lambda_j}{\lambda_1}\right)^k x_j\Big] .$$

Clearly $A^k v^{(0)}$, thus $v^{(k)}$, converges to the first eigenvector. And we have

$$|\lambda_1 - \lambda^{(k)}| = O\Big[\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^k\Big] .$$

Example 6.6. *Using the power method to find the eigenvalue of largest magnitude for the following matrix*

$$A = \begin{pmatrix} -261 & 209 & -49 \\ -530 & 422 & -98 \\ -800 & 631 & -144 \end{pmatrix}$$

Solution. This is a simple example, and it has three eigenvalues $\lambda_1 = 10$, $\lambda_2 = 4$ and $\lambda_3 = 3$. We use this simple example to demonstrate how effective the power method is.

Applying the Power Method with the initial guess $v^{(0)} = (1, 0, 0)^T$, we obtain the following results:

k	$\lambda^{(k)}$	k	$\lambda^{(k)}$
1	994.49	6	10.0198
2	13.0606	7	10.0063
3	10.7191	8	10.0020
4	10.2073	9	10.0007
5	10.0633	10	10.0002

Indeed, we observe that the Power Method converges very fast.

Next, we introduce another powerful method for finding the eigenvalues of a matrix A: the QR factorization.

QR Algorithm. Given an $n \times n$ matrix A, let $A^{(0)} := A$.

```
FOR  $k = 0, 1, 2, \cdots$ , DO
     $A^{(k)} = Q_k R_k$;    ($QR$ factorization)
     $A^{(k+1)} := R_k Q_k$
END
```

We can easily see that

$$A^{(1)} = R_0 Q_0 = (Q_0)^{-1}(Q_0 R_0) Q_0 = (Q_0)^{-1} A^{(0)} Q_0,$$

So $A^{(1)}$ and A have the same eigenvalues.[11]

Similarly we can see that $A^{(k)}$ has the same eigenvalues as $A^{(k-1)}$. And the following result holds:

The sequence $\{A^{(k)}\}$ generated by the QR algorithm converges to a upper triangular matrix and the diagonals of $A^{(k)}$ converge to the eigenvalues of A

Example 6.7. *Use the QR factorization to find the eigenvalues of the following matrix*

$$A = \begin{pmatrix} -261 & 209 & -49 \\ -530 & 422 & -98 \\ -800 & 631 & -144 \end{pmatrix}$$

Example 6.8. *Use the QR factorization to find the eigenvalues of*

$$A = \begin{pmatrix} 2 & -1 \\ -3 & 0 \end{pmatrix} .$$

Solution. The two eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = -1$. Practice yourself !

[11] We see that

$$|A_1 - \lambda I| = |Q^{-1}(A - \lambda I)Q| = |Q^{-1}||A - \lambda I||Q|,$$

therefore

$$|A_1 - \lambda I| = 0 \Leftrightarrow |A - \lambda I| = 0.$$

This indicates that A_1 and A have the same eigenvalues.

6.6 Discrete Fourier trasform

Recall the complex form of the Fourier series

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \ , \quad x \in [0, 2\pi]$$

where $f(x)$ is a periodic function with period 2π and

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx = \frac{1}{2\pi}\int_{0}^{2\pi} f(x)e^{-ikx}dx \ .$$

We now discuss how to calculate the Fourier coefficients $\{c_k\}$ by the computer. As we know, if we want to use computers to calculate these coefficients $\{c_k\}$, then it is expensive to use the integral form for the calculation.

Let us first see the nature of the problem. Basically, this calculation is of the form:

input: Given $f(x)$ for all $x \in [0, 2\pi]$ (infinite data);

output: find $c_0, c_1, c_2, \cdots, c_k, \cdots$ (infinite data).

Note that the computer can only handle the finite data. So we will need to truncate the Fourier series and approximate it by its discrete Fourier transformation. The discrete Fourier transformation is of the form:

input: Given n input data $f_0, f_1, \cdots, f_{n-1}$;

output: find $c_0, c_1, c_2, \cdots, c_{n-1}$ (discrete Fourier coefficients).

Let T denote this transformation, then we can write

$$T \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} ,$$

that is, T can be represented by a matrix.

6.6.1 Examples of Discrete Fourier transform

As an example, we first study the behaviour of this transformation T for a simple case. Suppose we have 4 data:

$$f_0 = 2, \quad f_1 = 4, \quad f_2 = 6, \quad f_3 = 8,$$

we try to find a truncated 4-term Fourier series

$$F_4(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + c_3 e^{3ix}$$

such that

$$F_4(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + c_3 e^{3ix} = \begin{array}{l} \begin{cases} f_0 = 2 \quad \text{at} \quad x = 0 \\ f_1 = 4 \quad \text{at} \quad x = \frac{\pi}{2} \\ f_2 = 6 \quad \text{at} \quad x = \pi \\ f_3 = 8 \quad \text{at} \quad x = \frac{3\pi}{2} \end{cases} \\ \hline \text{note } {\scriptstyle F_4(x) = f_0 = 2} \text{ at } {\scriptstyle x = 2\pi} \\ \hline \end{array} \tag{6.71}$$

This process is equivalent to finding a transformation T such that

$$T \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} \tag{6.72}$$

What is the matrix T ? From (6.71) we have

$$\begin{array}{lclll} x = 0 & : & c_0 + c_1 + c_2 + c_3 & = f_0 & = 2 \\ x = \frac{\pi}{2} & : & c_0 + c_1 e^{\frac{\pi}{2} i} + c_2 e^{2\frac{\pi}{2} i} + c_3 e^{3\frac{\pi}{2} i} & = f_1 & = 4 \\ x = \pi & : & c_0 + c_1 e^{\pi i} + c_2 e^{2\pi i} + c_3 e^{3\pi i} & = f_2 & = 6 \\ x = \frac{3\pi}{2} & : & c_0 + c_1 e^{\frac{3\pi}{2} i} + c_2 e^{2\frac{3\pi}{2} i} + c_3 e^{3\frac{3\pi}{2} i} & = f_3 & = 8 \end{array} \tag{6.73}$$

which can be written as follows:

$$A \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix} \tag{6.74}$$

where the matrix A is given by

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{\frac{\pi}{2} i} & e^{2\frac{\pi}{2} i} & e^{3\frac{\pi}{2} i} \\ 1 & e^{\pi i} & e^{2\pi i} & e^{3\pi i} \\ 1 & e^{\frac{3\pi}{2} i} & e^{2\frac{3}{2}\pi i} & e^{3\frac{3\pi}{2} i} \end{pmatrix} \overset{e^{\pi i/2} = i}{=\!=\!=} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & i^2 & i^3 \\ 1 & i^2 & i^4 & i^6 \\ 1 & i^3 & i^6 & i^9 \end{pmatrix}$$

Comparing (6.72) and (6.74), we know

$$T \;=\; A^{-1}\;.$$

We next calculate $T = A^{-1}$. Take the complex conjugate of A (changing every i to $-i$) and obtain

$$\overline{A} \;=\; \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & (-i)^2 & (-i)^3 \\ 1 & (-i)^2 & (-i)^4 & (-i)^6 \\ 1 & (-i)^3 & (-i)^6 & (-i)^9 \end{pmatrix}$$

Now it is easy to verify

$$\bar{A}A \;=\; \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \;=\; 4I\,.$$

Thus

$$T = A^{-1} = \frac{1}{4}\bar{A},$$

which gives the following formula for computing the coefficients c_0, c_1, c_2 and c_3:

$$\begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \frac{1}{4}\bar{A}\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix}\;.$$

6.6.2 Discrete Fourier transform in n variables

In this section we consider the general discrete Fourier transform. Given a function $f(x)$ on the interval $[0, 2\pi]$, we want to find a finite Fourier series

$$F_n(x) = \sum_{k=0}^{n-1} c_k e^{ikx}\;, \quad x \in [0, 2\pi] \tag{6.75}$$

such that

$$F_n(x_j) = f(x_j), \quad x_j = \frac{j \cdot 2\pi}{n}, \quad j = 0, 1, 2, \cdots, n-1\,. \tag{6.76}$$

Using the n conditions above, we can determine the n coefficients in (6.75). In fact, let $x = x_j = \frac{j \cdot 2\pi}{n}$ in (6.75) and use (6.76), we have

$$\begin{array}{ll} \text{at } x_0: & c_0 + c_1 + c_2 + \cdots + c_{n-1} = f_0 \equiv f(x_0) \\ \text{at } x_1: & c_0 + c_1 e^{ix_1} + c_2 e^{2ix_1} + \cdots + c_{n-1} e^{(n-1)ix_1} = f_1 \equiv f(x_1) \\ \text{at } x_2: & c_0 + c_1 e^{ix_2} + c_2 e^{2ix_2} + \cdots + c_{n-1} e^{(n-1)ix_2} = f_2 \equiv f(x_2) \\ \cdots & \\ \text{at } x_{n-1}: & c_0 + c_1 e^{ix_{n-1}} + c_2 e^{2ix_{n-1}} + \cdots + c_{n-1} e^{(n-1)ix_{n-1}} = f_{n-1} \end{array} \tag{6.77}$$

Set

$$w = e^{2\pi i/n} = e^{ix_1},$$

then we have

$$e^{ix_2} = w^2, \quad e^{ix_3} = w^3, \cdots,$$

so the system (6.77) can be written as

$$\begin{array}{l} c_0 + c_1 + c_2 + \cdots + c_{n-1} = f_0 \\ c_0 + c_1 w + c_2 w^2 + \cdots + c_{n-1} w^{n-1} = f_1 \\ c_0 + c_1 w^2 + c_2 w^4 + \cdots + c_{n-1} w^{2(n-1)} = f_2 \\ \cdots \\ c_0 + c_1 w^{n-1} + c_2 w^{2(n-1)} + \cdots + c_{n-1} w^{(n-1)^2} = f_{n-1} \end{array} \tag{6.78}$$

and further written as

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w & w^2 & \cdots & w^{n-1} \\ 1 & w^2 & w^4 & \cdots & w^{2(n-1)} \\ \cdots & & & & \\ 1 & w^{n-1} & w^{2(n-1)} & \cdots & w^{(n-1)^2} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{n-1} \end{pmatrix} \tag{6.79}$$

For convenience, we write this system as

$$Ac = f.$$

Let us see how to find the solution vector $c = (c_1, c_2, \cdots, c_{n-1})^T$ in (6.79). Note that

$$1 + w + w^2 + \cdots + w^{n-1} = \frac{1 - w^n}{1 - w} = 0\ , \tag{6.80}$$

and

$$\begin{aligned}
&1\cdot 1 + w^j\bar{w}^k + w^{2j}\bar{w}^{2k} + \cdots + w^{(n-1)j}\bar{w}^{(n-1)k} \\
=&1 + e^{\frac{j(2\pi i)}{n}}e^{-\frac{k(2\pi i)}{n}} + e^{\frac{2j(2\pi i)}{n}}e^{-\frac{2k(2\pi i)}{n}} + \cdots + e^{\frac{(n-1)j(2\pi i)}{n}}e^{-\frac{(n-1)k(2\pi i)}{n}} \\
=&1 + e^{\frac{(j-k)(2\pi i)}{n}} + e^{\frac{2(j-k)(2\pi i)}{n}} + \cdots + e^{\frac{(n-1)(j-k)(2\pi i)}{n}} \qquad (6.81) \\
=&\begin{cases} n \quad \text{for} \ \ j = k; \\ \frac{1-e^{(j-k)(2\pi i)}}{1-e^{\frac{(j-k)(2\pi i)}{n}}} = 0 \quad \text{for} \ \ j \neq k\,. \end{cases}
\end{aligned}$$

Using this relation we obtain (**please check yourself !**):

$$A\bar{A} = \bar{A}A = nI \qquad (6.82)$$

or

$$A^{-1} = \frac{1}{n}\bar{A}\,.$$

Therefore the required discrete Fourier coefficients $\{c_i\}_{i=0}^{n-1}$ can be calculated by

$$c = A^{-1}f = \frac{1}{n}\bar{A}f \ . \qquad (6.83)$$

Or equivalently, we have for $k = 0, 1, \cdots, n-1$ that

$$c_k = \frac{1}{n}\big(f_0 + \bar{w}^k f_1 + \bar{w}^{2k} f_2 + \cdots + \bar{w}^{(n-1)k} f_{n-1}\big) \ .$$

But $\bar{w} = e^{-\frac{2\pi i}{n}}$, thus

$$c_k = \frac{1}{n}\sum_{j=0}^{n-1} f_j\,\bar{w}^{jk} = \frac{1}{n}\sum_{j=0}^{n-1} f_j\,e^{-i\frac{2jk\pi}{n}} \ , \quad k = 0, 1, \cdots, n-1\,. \qquad (6.84)$$

6.6.3 Relations between the discrete and continuous Fourier coefficients

The discrete Fourier coefficients in (6.84) have very close relations with the continuous Fourier coefficients

$$\hat{c}_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}dx\,. \qquad (6.85)$$

Let us see how these two sets of coefficients are related. To do so, we divide $[0, 2\pi]$ into n equally spaced subintervals:

$$0 = x_0 < x_1 < \cdots < x_{n-1} < x_n = 2\pi$$

with $x_j = jh, h = \frac{2\pi}{n}$, then we can write $\hat{c}_k$ in (6.85) as

$$\hat{c}_k = \frac{1}{2\pi} \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} \tilde{f}(x) dx \ , \quad \tilde{f}(x) = f(x) e^{-ikx} \, .$$

Now using the trapezoidal rule on each interval, we obtain

$$\begin{aligned} \hat{c}_k =& \frac{1}{2\pi} \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} \tilde{f}(x) dx \approx \frac{1}{2\pi} \sum_{j=0}^{n-1} \frac{h}{2} \big(\tilde{f}(x_j) + \tilde{f}(x_{j+1}) \big) \\ =& \frac{h}{4\pi} \sum_{j=0}^{n-1} \big(\tilde{f}(x_j) + \tilde{f}(x_{j+1}) \big) = \frac{h}{4\pi} \big\{ \tilde{f}(x_0) + \tilde{f}(x_n) + 2 \sum_{j=1}^{n-1} \tilde{f}(x_j) \big\} \, . \end{aligned}$$

Note that $f(x)$ is always assumed to be a periodic function with a period of 2π, so

$$\tilde{f}(x_n) = \tilde{f}(2\pi) = \tilde{f}(x_0).$$

Thus

$$\hat{c}_k \approx \frac{h}{2\pi} \sum_{j=0}^{n-1} \tilde{f}(x_j) = \frac{1}{n} \sum_{j=0}^{n-1} f(x_j) e^{-i \frac{2jk\pi}{n}},$$

this is exactly the previouly discussed discrete Fourier coefficients in (6.84).

In summary, we see from above that the discrete Fourier coefficients in (6.84) are the approximations of the continuous Fourier coefficients in (6.85) by using the trapezoidal rule for computing the integrals.

6.7 The fast Fourier transform

Recall the Fourier series of a function $f(x)$:

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \ , \ x \in [0, 2\pi] \ .$$

Now let us first look at the basis functions

$$\{e^{ikx}\}_{k=-\infty}^{\infty} \ , \quad x \in [0, 2\pi] \ .$$

If we divide $[0, 2\pi]$ into n equally spaced subintervals:

$$0 = x_0 < x_1 < \cdots < x_{n-1} < x_n = 2\pi \ ,$$

then the grid points are

$$x_j = j\,\frac{2\pi}{n}\ , \quad j = 0, 1, 2, \cdots, n\ .$$

Note that e^{ikx} has the same value at $x = x_0$ and $x = x_n$, so we consider only the n different values of e^{ikx} at the following points

$$x_0,\ x_1,\ x_2,\ \cdots,\ x_{n-1}\ ,$$

namely the following function values

$$e^{ikx_0},\ e^{ikx_1},\ \cdots,\ e^{ikx_{n-1}}\ .$$

Taking $k = 0, 1, 2, \cdots, n-1$, we obtain the following discrete Fourier coefficient matrix

$$F_n \ = \ \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ e^{ix_0} & e^{ix_1} & e^{ix_2} & \cdots & e^{ix_{n-1}} \\ e^{i(2x_0)} & e^{i(2x_1)} & e^{i(2x_2)} & \cdots & e^{i(2x_{n-1})} \\ \cdots & & & & \\ e^{i(n-1)x_0} & e^{i(n-1)x_1} & e^{i(n-1)x_2} & \cdots & e^{i(n-1)x_{n-1}} \end{pmatrix}$$

Let $w_n = e^{i\frac{2\pi}{n}} = e^{ix_1}$, then we can write

$$F_n \ = \ \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w_n & w_n^2 & \cdots & w_n^{n-1} \\ 1 & w_n^2 & w_n^4 & \cdots & w_n^{2(n-1)} \\ \cdots & & & & \\ 1 & w_n^{n-1} & w_n^{2(n-1)} & \cdots & w_n^{(n-1)^2} \end{pmatrix}$$

This matrix is exactly the same as matrix A in (6.79), and arises in many applications. The fast Fourier transform (FFT) is an efficient method to compute $F_n\,x$ (for any given vector x) faster. We now discuss this fast Fourier transform.

First, let us see how expensive the usual way to compute the multiplication $F_n\,x$. For, $n = 2m$, the usual calculation of $F_n\,x$ requires n^2 multiplication. But the FFT requires only

$$\frac{1}{2} n \log_2 n \quad \text{operations}\ .$$

If $n = 2^{12}$, then $n^2 = 2^{24} = 2^9 \times 2^{15}$ but $n \log_2 n = 6 \times 2^{12} < 2^{15}$. So if 2^{15} operations need one second, then n^2 operations needs 2^9 seconds =8.5 minutes. If 2^{15} operations need one minute: then n^2 operations needs 512 minutes = 8.5 hours. So FFT will save huge time if the multiplication $F_n\,x$ is needed many times. This is indeed often the case in real applications.

Next, we discuss how to compute $y = F_n\,x$, we needs to compute the n components

$$y_0, \quad y_1, \quad \cdots, \quad y_{n-1}\ .$$

By definition, we have

$$y_j = \sum_{k=0}^{n-1} w_n^{kj} x_k = \sum_{k=0}^{2m-1} w_n^{kj} x_k\ . \tag{6.86}$$

Divide $k = 0, 1, 2, \cdots, 2m-1$ into

$$\begin{aligned} k &= 0, 2, \cdots, 2(m-1)\ , \\ k &= 1, 3, \cdots, 2m-1\ , \end{aligned}$$

then we can write

$$y_j = \sum_{k=0}^{m-1} w_n^{2kj} x_{2k} + \sum_{k=0}^{m-1} w_n^{(2k+1)j} x_{2k+1}\ . \tag{6.87}$$

Set

$$x' = (x_0, x_2, \cdots, x_{2(m-1)})^T, \quad x'' = (x_1, x_3, \cdots, x_{2m-1})^T,$$

then for $j = 0, 1, 2, \cdots, m-1$, we have

$$y_j = \sum_{k=0}^{m-1} w_m^{kj} x'_k + w_n^j \sum_{k=0}^{m-1} w_m^{kj} x''_k$$

where we have used

$$w_n^2 = e^{i\frac{4\pi}{n}} = e^{i\frac{2\pi}{m}} = w_m,$$

thus comparing with (6.86), we see

$$y_i = (F_m x')_j + w_n^j (F_m x'')_j, \quad j = 0, 1, 2, \cdots, m-1\ .$$

Let $y' = F_m x'$, $y'' = F_m x''$, then

$$y_j = y'{}_j + w_n^j y''_j\ , \quad j = 0, 1, 2, \cdots, m-1\ . \tag{6.88}$$

Next replacing j in (6.87) by $j+m$ ($j = 0, 1, \cdots, m-1$), we obtain

$$\begin{aligned} y_{j+m} &= \sum_{k=0}^{m-1} w_n^{2k(j+m)} x_{2k} + \sum_{k=0}^{m-1} w_n^{(2k+1)(j+m)} x_{2k+1} \\ &= \sum_{k=0}^{m-1} w_m^{kj+km} x'_k + w_n^{j+m} \sum_{k=0}^{m-1} w_m^{k(j+m)} x''_k\ . \end{aligned}$$

Note that

$$w_m^{km} = e^{(i\frac{2\pi}{m})km} = 1 \ , \quad w_n^m = e^{(i\frac{2\pi}{n})m} = e^{i\pi} = -1,$$

hence

$$\begin{aligned} y_{j+m} &= \sum_{k=0}^{m-1} w_m^{kj} x_k' - w_n^j \sum_{k=0}^{m-1} w_m^{kj} x_k'' \\ &= (F_m x')_j - w_n^j (F_m x'')_j \\ &= y_j' - w_n^j y_j'' \ , \quad j = 0, 1, 2, \cdots, m-1 \ . \end{aligned} \tag{6.89}$$

From (6.89) and (6.88), we know that $y = F_n x$ can be calculated as follows:

1. Split x into

$$x' = (x_0, x_2, \cdots, x_{2(m-1)})^T \ , \quad x'' = (x_1, x_3, \cdots, x_{2m-1})^T \ .$$

2. Compute $y' = F_m x'$ and $y'' = F_m x''$.

3. Compute the components of $y = F_n x$ by

$$\begin{aligned} y_j &= y_j' + w_n^j y_j'' \ , \quad j = 0, 1, 2, \cdots, m-1 \ , \\ y_{j+m} &= y_j' - w_n^j y_j'' \ , \quad j = 0, 1, 2, \cdots, m-1 \ . \end{aligned}$$

Then for the calculation of $y' = F_m x'$ and $y'' = F_m x''$, we can again reduce to the multiplication of $F_{m/2}$, and finally to F_1. Let us denote the number of computations of F_m by C_m. We see that $C_1 = 1$ and

$$C_{2m} = 2C_m + 3m.$$

The number 3 here includes one multiplication and two additions. If we choose $n = 2^l$, then C_{2^l} is

$$C_{2^l} = 2C_{2^{l-1}} + 3 \cdot 2^{l-1}$$

This gives

$$2^{-l} C_{2^l} = 2^{-l+1} C_{2^{l-1}} + 3 \cdot 2^{-1}$$

This implies

$$2^{-l} C_{2^l} = \frac{3}{2} \cdot l,$$

or

$$C_{2^l} = \frac{3}{2} \cdot l \cdot 2^l = \frac{3}{2} n \log_2 n.$$

Homeworks

- pp. 468, 5.5.1
- pp. 468, 5.5.3
- pp. 468, 5.5.4
- pp. 468, 5.5.5
- pp. 468, 5.5.8
- pp. 469, 5.5.10