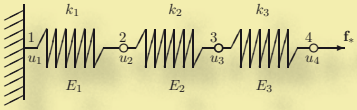
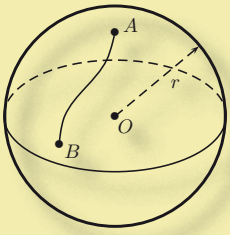
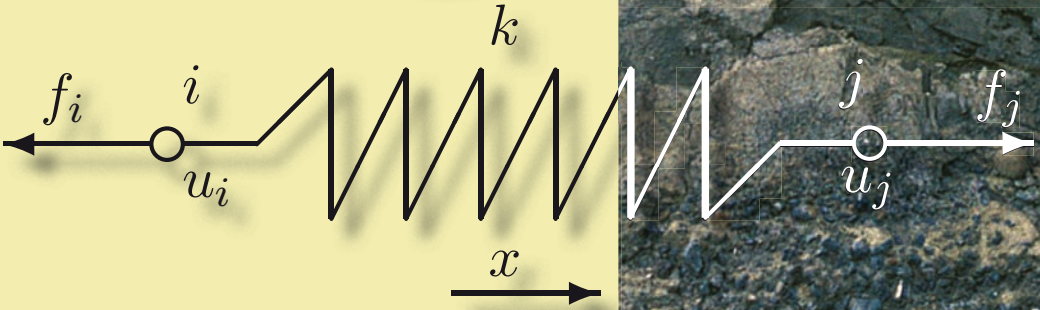
# Mathematical Modelling for Earth Sciences

Xin-She Yang




$$\sigma = -\frac{2RT}{\nu_m L^2}\int_0^L r\ln[1-\frac{\rho_s \dot{e}\bar{d}}{4c_0 D_{gb} w}(L^2-r^2)]dr$$




DUNEDIN

# Mathematical Modelling for Earth Sciences

**Xin-She Yang**

Department of Engineering, University of Cambridge

DUNEDIN

Published by
Dunedin Academic Press ltd
Hudson House
8 Albany Street
Edinburgh EH1 3QB
Scotland

www.dunedinacademicpress.co.uk

ISBN: 978-1-903765-92-0



British Library Cataloguing in Publication data
A catalogue record for this book is available from the British Library



Printed in the United Kingdom by Cpod.

# Contents

# Preface

Mathematical modelling and computer simulations are an essential part of the analytical skills for earth scientists. Nowadays, computer simulations based on mathematical models are routinely used to study various geophysical, environmental and geological processes, from geophysics to petroleum engineering, from hydrology to environmental fluid dynamics. The topics in earth sciences are very diverse and the syllabus itself is evolving. From a mathematical modelling point of view, therefore, this is a decision to select topics and limit the number of chapters so that the book remains concise and yet comprehensive enough to include important and interesting topics and popular algorithms. Furthermore, we use a 'theorem-free' approach in this book with a balance of formality and practicality. We will increase dozens of worked examples so as to tackle each problem in a step-by-step manner, thus the style will be especially suitable for non-mathematicians, though there are enough topics, such as the calculus of variation and pattern formation, that even mathematicians may find them interesting.

This book strives to introduce a wide range of mathematical modelling and numerical techniques, especially for undergraduates and graduates. Topics include vector and matrix analysis, ordinary differential equations, partial differential equations, calculus of variations, integral equations, probability, geostatistics, numerical integration, optimisation, finite difference methods, finite volume methods and finite element methods. Application topics in earth sciences include reaction-diffusion system, elasticity, fracture mechanics, poroelasticity, and flow in porous media. This book can serve as a textbook in mathematical modelling and numerical methods for earth sciences.

This book covers many areas of my own research and learning from experts in the field, and it represents my own personal odyssey through the diversity and multidisciplinary exploration. Over these years, I have received valuable help in various ways from my mentors, friends, colleagues, and students. First and foremost, I would like to thank my mentors, tutors and colleagues: A. C. Fowler, C. J. Mcdiarmid and S. Tsou at Oxford University for introducing me to the wonderful world of applied mathematics; J. M. Lees, C. T. Morley and G. T. Parks at Cambridge University for giving me the opportunity to work on the applications of mathematical methods and numerical simulations in various research projects; and A. C. McIntosh, J. Brindley, K. Seffan and T. Love who have all helped me in various ways.

I thank many of my students who have directly and/or indirectly tried some parts of this book and gave their valuable suggestions. Special thanks to Hugo Scott Whittle, Charles Pearson, Ryan Harper, J. H. Tan, Alexander Slinger and Adam Gordon at Cambridge University for their help in proofreading the book.

In addition, I am fortunate to have discussed many important topics with many international experts: D. Audet and H. Ockendon at Oxford, J. A. D. Connolly at ETHZ, A. Revil at Colorado, D. L. Turcotte at Cornell, B. Zhou at CSIRO, and E. Holzbecher at WIAS. I would like to thank them for their help.

I also would like to thank the staff at Dunedin Academic Press for their kind encouragement, help and professionalism. Special thanks to the publisher's referees, especially to Oyvind Hammer of the University of Oslo, Norway, for their insightful and detailed comments which have been incorporated in the book.

Last but not least, I thank my wife, Helen, and son, Young, for their help and support.

While every attempt is made to ensure that the contents of the book are right, it will inevitably contain some errors, which are the responsibility of the author. Feedback and suggestions are welcome.

Xin-She Yang
Cambridge, 2008

# Part I

# Mathematical Methods

# Chapter 1

# Mathematical Modelling

## 1.1 Introduction

### 1.1.1 Mathematical Modelling

Mathematical modelling is the process of formulating an abstract model in terms of mathematical language to describe the complex behaviour of a real system. Mathematical models are quantitative models and often expressed in terms of ordinary differential equations and partial differential equations. Mathematical models can also be statistical models, fuzzy logic models and empirical relationships. In fact, any model description using mathematical language can be called a mathematical model. Mathematical modelling is widely used in natural sciences, computing, engineering, meteorology, and of course earth sciences. For example, theoretical physics is essentially all about the modelling of real world processes using several basic principles (such as the conservation of energy, momentum) and a dozen important equations (such as the wave equation, the Schrodinger equation, the Einstein equation). Almost all these equations are partial differential equations.

An important feature of mathematical modelling and numerical algorithms concerning earth sciences is its interdisciplinary nature. It involves applied mathematics, computer sciences, earth sciences, and others. Mathematical modelling in combination with scientific computing is an emerging interdisciplinary technology. Many international companies use it to model physical processes, to design new products, to find solutions to challenging problems, and increase their competitiveness in international markets.

The basic steps of mathematical modelling can be summarised as meta-steps shown in Fig. 1.1. The process typically starts with the analysis of a real world problem so as to extract the fundamental phys-

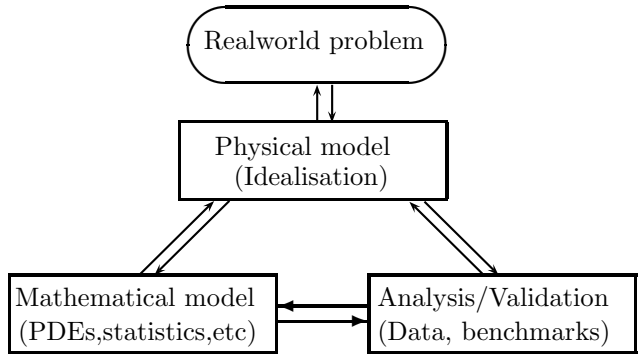



Figure 1.1: Mathematical modelling.

ical processes by idealisation and various assumptions. Once an idealised physical model is formulated, it can then be translated into the corresponding mathematical model in terms of partial differential equations (PDEs), integral equations, and statistical models. Then, the mathematical model should be investigated in great detail by mathematical analysis (if possible), numerical simulations and other tools so as to make predictions under appropriate conditions. Then, these simulation results and predictions will be validated against the existing models, well-established benchmarks, and experimental data. If the results are satisfactory (which they rarely are at first), then the mathematical model can be accepted. If not, both the physical model and mathematical model will be modified based on the feedback, then the new simulations and prediction will be validated again. After a certain number of iterations of the whole process (often many), a good mathematical model can properly be formulated, which will provide great insight into the real world problem and may also predict the behaviour of the process under study.

For any physical problem in earth sciences, for example, there are traditionally two ways to deal with it by either theoretical approaches or field observations and experiments. The theoretical approach in terms of mathematical modelling is an idealisation and simplification of the real problem and the theoretical models often extract the essential or major characteristics of the problem. The mathematical equations obtained even for such over-simplified systems are usually very difficult for mathematical analysis. On the other hand, the field studies and experimental approach is usually expensive if not impractical. Apart from financial and practical limitations, other constraining factors in-

clude the inaccessibility of the locations, the range of physical parameters, and time for carrying out various experiments. As computing speed and power have increased dramatically in the last few decades, a practical third way or approach is emerging, which is computational modelling and numerical experimentation based on the mathematical models. It is now widely acknowledged that computational modelling and computer simulations serve as a cost-effective alternative, bridging the gap or complementing the traditional theoretical and experimental approaches to problem solving.

Mathematical modelling is essentially an abstract art of formulating the mathematical models from the corresponding real-world problems. The master of this art requires practice and experience, and it is not easy to teach such skills as the style of mathematical modelling largely depends on each person's own insight, abstraction, type of problems, and experience of dealing with similar problems. Even for the same physical process, different models could be obtained, depending on the emphasis of some part of the process, say, based on your interest in certain quantities in a particular problem, while the same quantities could be viewed as unimportant in other processes and other problems.

### 1.1.2 Model Formulation

Mathematical modelling often starts with the analysis of the physical process and attempts to make an abstract physical model by idealisation and approximations. From this idealised physical model, we can use the various first principles such as the conservation of mass, momentum, energy and Newton's law to translate into mathematical equations. Let us look at the example of the diffusion process of sugar in a glass of water. We know that the diffusion of sugar will occur if there is any spatial difference in the sugar concentration. The physical process is complicated and many factors could affect the distribution of sugar concentration in water, including the temperature, stirring, mass of sugar, type of sugar, how you add the sugar, even geometry of the container and others. We can idealise the process by assuming that the temperature is constant (so as to neglect the effect of heat transfer), and that there is no stirring because stirring will affect the effective diffusion coefficient and introduce the advection of water or even vertices in the (turbulent) water flow. We then choose a representative element volume (REV) whose size is very small compared with the size of the cup so that we can use a single value of concentration to represent the sugar content inside this REV (If this REV is too large, there is considerable variation in sugar concentration inside this REV). We also assume that there is no chemical reaction between sugar and water (otherwise, we are dealing with something else). If you drop

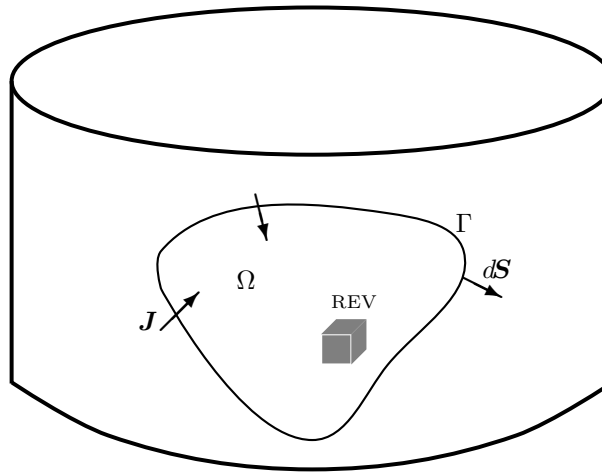



Figure 1.2: Representative element volume (REV).

the sugar into the cup from a considerable height, the water inside the glass will splash and thus fluid volume will change, and this becomes a fluid dynamics problem. So we are only interested in the process after the sugar is added and we are not interested in the initial impurity of the water (or only to a certain degree). With these assumptions, the whole process is now idealised as the physical model of the diffusion of sugar in still water at a constant temperature. Now we have to translate this idealised model into a mathematical model, and in the present case, a parabolic partial differential equation or diffusion equation [These terms, if they sound unfamiliar, will be explained in detail in the book]. Let us look at an example.

---

**Example 1.1:** Let $c$ be the averaged concentration in a representative element volume with a volume $dV$ inside the cup, and let $\Omega$ be an arbitrary, imaginary closed volume $\Omega$ (much larger than our REV but smaller than the container, see Fig. 1.2). We know that the rate of change of the mass of sugar per unit time inside $\Omega$ is

$$\delta_1 = \frac{\partial}{\partial t} \iiint_\Omega c dV,$$

where $t$ is time. As the mass is conserved, this change of sugar content in $\Omega$ must be supplied in or flow out over the surface $\Gamma = \partial\Omega$ enclosing the region $\Omega$. Let $\boldsymbol{J}$ be the flux through the surface, thus the total mass flux

through the whole surface $\Gamma$ is

$$\delta_2 = \iint_\Gamma \boldsymbol{J} \cdot d\boldsymbol{S}.$$

Thus the conservation of total mass in $\Omega$ requires that

$$\delta_1 + \delta_2 = 0,$$

or

$$\frac{\partial}{\partial t} \iiint_\Omega c dV + \iint_\Gamma \boldsymbol{J} \cdot d\boldsymbol{S} = 0.$$

This is essentially the integral form of the mathematical model. Using the Gauss's theorem (discussed later in this book)

$$\iint_\Gamma \boldsymbol{J} \cdot d\boldsymbol{S} = \iiint_\Omega \nabla \cdot \boldsymbol{J} dV,$$

we can convert the surface integral into a volume integral. We thus have

$$\frac{\partial}{\partial t} \iiint_\Omega c dV + \iiint_\Omega \nabla \cdot \boldsymbol{J} dV = 0.$$

Since the domain $\Omega$ is fixed (independent of $t$), we can interchange the differentiation and integration in the first term, we now get

$$\iiint_\Omega \frac{\partial c}{\partial t} dV + \iiint_\Omega \nabla \cdot \boldsymbol{J} dV = \iiint_\Omega [\frac{\partial c}{\partial t} + \nabla \cdot J] dV = 0.$$

Since the enclosed domain $\Omega$ is arbitrary, the above equation should be valid for any shape or size of $\Omega$, therefore, the integrand must be zero. We finally have

$$\frac{\partial c}{\partial t} + \nabla \cdot \boldsymbol{J} = 0.$$

This is the differential form of the mass conservation. It is a partial differential equation. As we know that diffusion occurs from the higher concentration to lower concentration, the rate of diffusion is proportional to the gradient $\nabla c$ of the concentration. The flux $\boldsymbol{J}$ over a unit surface area is given by Fick's law

$$\boldsymbol{J} = -D \nabla c,$$

where $D$ is the diffusion coefficient which depends on the temperature and the type of materials. The negative sign means the diffusion is opposite to the gradient. Substituting this into the mass conservation, we have

$$\frac{\partial c}{\partial t} - \nabla \cdot (D \nabla c) = 0,$$

or

$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c).$$

---

In the simplified case when $D$ is constant, we have

$$\frac{\partial c}{\partial t} = D \nabla^2 c, \tag{1.1}$$

which is the well-known diffusion equation. This equation can be applied to study many phenomena such as heat conduction, pore pressure dissipation, groundwater flow and consolidation if we replace $D$ by the corresponding physical parameters. This will be discussed in greater detail in the related chapters this book.

### 1.1.3 Parameter Estimation

Another important topic in mathematical modelling is the ability to estimate the orders (not the exact numbers) of certain quantities. If we know the order of a quantity and its range of variations, we can choose the right scales to write the mathematical model in the non-dimensional form so that the right mathematical methods can be used to tackle the problem. It also helps us to choose more suitable numerical methods to find the solution over the correct scales. The estimations will often give us greater insight into the physical process, resulting in more appropriate mathematical models. For example, if we want to study plate tectonics, what physical scales (forces and thickness of the mantle) would be appropriate? For a given driving force (from thermal convection or pulling in the subduction zone), could we estimate the order of the plate drifting velocity? Of course, the real process is extremely complicated and it is still an ongoing research area. However, let us do some simple (yet not so naive) estimations.

---

**Example 1.2:** Estimation of plate drifting velocity: we know the drift of the plate is related to the thermal convection, and the deformation is mainly governed by viscous creep (discussed later in this book). The strain rate $\dot{e}$ is linked to the driving stress $\sigma$ by

$$\dot{e} = \frac{\sigma}{\eta},$$

where $\eta$ is the viscosity of the mantle and can be taken as fixed value $\eta = 10^{21}$ Pa s (it depends on temperature). The estimation of $\eta$ will be discussed in Chapter 15.

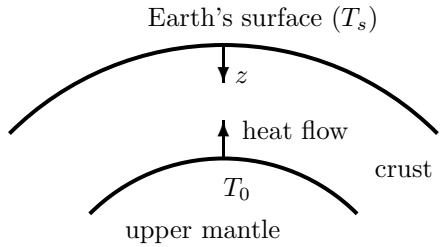



Figure 1.3: Estimation of the rate of heat loss on the Earth's surface.

Let $L$ be the typical scale of the mantle, and $v$ be the averaged drifting velocity. Thus, the strain rate can be expressed as

$$\dot{e} = \frac{v}{L}.$$

Combining this equation with the above creep relationship, we have

$$v = \frac{L\sigma}{\eta}.$$

Using the typical values of $L \approx 3000$ km $\approx 3 \times 10^6$ m, $\sigma \approx 10^6$ Pa, we have

$$v = \frac{L\sigma}{\eta} \approx \frac{3 \times 10^6 \times 10^6}{2 \times 10^{21}} \approx 1.5 \times 10^{-9}\text{m/s} \approx 4.7\text{cm/year}.$$

This value is about right as most plates move in the range of $1 \sim 10$ cm per year. The other interesting thing is that the accurate values of $\sigma$ and $L$ are not needed as long as their product is about the same as $L\sigma \approx 3 \times 10^{12}$, the estimation of $v$ will not change much.

If we use $L \approx 1000$ km $\approx 10^6$ m, then, to produce the same velocity, it requires that $\sigma = 3 \times 10^6$ Pa $\approx 30$ atm, or about 30 atmospheric pressures. Surprisingly, the driving stress for such large motion is not huge. The force could be easily supplied by the pulling force (due to density difference) of the subducting slab in the subduction zone.

---

Let us look at another example to estimate the rate of heat loss at the Earth's surface, and the temperature gradients in the Earth's crust and the atmosphere. We can also show the importance of the sunlight in the heat energy balance of the atmosphere.

**Example 1.3:** We know that the average temperature at the Earth's surface is about $T_s = 300$K, and the thickness of the continental crust varies from $d = 35$km to 70km. The temperature at the upper lithosphere is estimated about $T_0 = 900 \sim 1400$K (very crude estimation). Thus the estimated temperature gradient is about

$$\frac{dT}{dz} = \frac{T_0 - T_s}{d} \approx 9 \sim 31\mathrm{K/km}.$$

The observed values of the temperature gradient around the globe are about $10$ to $30$ K/km. The estimated thermal conductivity $k$ of rocks is about $1.5 \sim 4.5$ W/m K (ignoring the temperature dependence), we can use $k = 3$ W/m K as the estimate for the thermal conductivity of the crust. Thus, the rate of heat loss obeys Fourier's law of conduction

$$q = -k\nabla T = -k\frac{dT}{dz} \approx 0.027 \sim 0.093\mathrm{W/m^2},$$

which is close to the measured average of about 0.07 W/m$^2$. For oceanic crust with a thickness of $6 \sim 7$ km, the temperature gradient (and thus rate of heat loss) could be five times higher at the bottom of the ocean, and this heat loss provides a major part of the energy to the ocean so as to keep it from being frozen.

If this heat loss goes through the atmosphere, then the energy conservation requires that

$$k\frac{dT}{dz}\Big|_{\mathrm{crust}} + k_a\frac{dT}{dh}\Big|_{\mathrm{air}} = 0,$$

where $h$ is the height above the Earth's surface and $k_a = 0.020 \sim 0.025$ W/m K is the thermal conductivity of the air (again, ignoring the variations with the temperature). Therefore, the temperature gradient in the air is

$$\frac{dT}{dh} = -\frac{k}{k_a}\frac{dT}{dz} \approx -3.6 \sim -4.5\mathrm{K/km},$$

if we use $dT/dz = 30$ W/km. The negative sign means the temperature decreases as the height increases. The true temperature gradient in dry air is about 10 K/km in dry air, and $6 \sim 7$K/km in moist air. As the thermal conductivity increases with the humidity, so the gradient decreases with humidity.

Alternatively, we know the effective thickness of the atmosphere is about 50 km (if we define it as the thickness of layers containing 99.9% of the air mass). We know there is no definite boundary between the atmosphere and outer space, and the atmosphere can extend up to several

hundreds of kilometres. In addition, we can also assume that the temperature in space vacuum is about $4$ K and the temperature at the Earth's surface is $300$K, then the temperature gradient in the air is

$$\frac{dT}{dh} \approx \frac{4-300}{50} \approx -6\text{K/km},$$

which is quite close to the true gradient. The higher rate of heat loss (due to higher temperature gradient) means that the heat supplied from the crust is not enough to balance this higher rate. That is where the energy of sunlight comes into play. We can see that estimates of this kind will provide a good insight in the whole process.

Of course the choice of typical values is important in order to get a valid estimation. Such choice will depend on the physical process and the scales we are interested in. The right choice will be perfected by expertise and practice. We will give many worked examples like this in this book.

## 1.2 Mathematical Models

### 1.2.1 Differential Equations

The first step of the mathematical modelling process produces some mathematical equations, often partial differential equations. The next step is to identify the detailed constraints such as the proper boundary conditions and initial conditions so that we can obtain a unique set of solutions. For the sugar diffusion problem discussed earlier, we cannot obtain the exact solution in the actual domain inside the water-filled glass, because we need to know where the sugar cube or grains were initially added. The geometry of the glass also needs to be specified. In fact, this problem needs numerical methods such as finite element methods or finite volume methods. The only possible solution is the long-time behaviour: when $t \to \infty$, we know that the concentration should be uniform $c(z, t \to \infty) \to c_\infty$ (=mass of sugar added/volume of water).

You may say that we know this final state even without mathematical equations, so what is the use of the diffusion equation ? The main advantage is that you can calculate the concentration at any time using the mathematical equation with appropriate boundary and initial conditions, either by numerical methods in most cases or by mathematical analysis in some very simple cases. Once you know the initial and boundary conditions, the whole system history will be determined to a certain degree. The beauty of mathematical models is that many

seemingly diverse problems can be reduced to the same mathematical equation. For example, we know that the diffusion problem is governed by the diffusion equation $\frac{\partial c}{\partial t} = D\nabla^2 c$. The heat conduction is governed by the heat conduction equation

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T, \qquad \kappa = \frac{K}{\rho c_p}, \tag{1.2}$$

where $T$ is temperature and $\kappa$ is the thermal diffusivity. $K$ is thermal conductivity, $\rho$ is the density and $c_p$ is the specific heat capacity. Similarly, the dissipation of the pore pressure $p$ in poroelastic media is governed by

$$\frac{\partial p}{\partial t} = c_v \nabla^2 p, \tag{1.3}$$

where $c_v = k/(S\mu)$ is the consolidation coefficient, $k$ is the permeability of the media, $\mu$ is the viscosity of fluid (water), and $S$ is the specific storage coefficient.

Mathematically speaking, whether it is concentration, temperature or pore pressure, it is the same dependent variable $u$. Similarly, it is just a constant $\kappa$ whether it is the diffusion coefficient $D$, the thermal diffusivity $\alpha$ or the consolidation coefficient $c_v$. In this sense, the above three equations are identical to the following parabolic partial differential equation

$$\frac{\partial u}{\partial t} = \kappa \nabla^2 u. \tag{1.4}$$

Suppose we want to solve the following problem. For a semi-infinite domain shown in Fig. 1.4, the initial condition (whether temperature or concentration or pore pressure) is $u(x, t = 0) = 0$. The boundary condition at $x = 0$ is that $u(x = 0, t) = u_0 =$const at any time $t$. Now the question what is distribution of $u$ versus $x$ at $t$?

Let us summarise the problem. As this problem is one-dimensional, only the $x$-axis is involved, and it is time-dependent. So we have

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}, \tag{1.5}$$

with an initial condition

$$u(x, t = 0) = 0, \tag{1.6}$$

and the boundary condition

$$u(x = 0, t) = u_0. \tag{1.7}$$

Let us start to solve this mathematical problem. How should we start and where to start? Well, there are many techniques to solve these

problems, including the similarity solution technique, Laplace's transform, Fourier's transform, separation of variables and others.

Similarity variable is an interesting and powerful method because it neatly transforms a partial differential equation (PDE) into an ordinary differential equation (ODE) by introducing a similarity variable $\zeta$, then you can use the standard techniques for solving ODEs to obtain the desired solution. We first define a similar variable

$$\zeta = \frac{x^2}{4\kappa t}, \tag{1.8}$$

so that $u(x,t) = u(\zeta) = f(\zeta)$. Using the chain rules of differentiations

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \zeta}\frac{\partial \zeta}{\partial x} = \frac{x}{2\kappa t}\frac{\partial}{\partial \zeta},$$

$$\frac{\partial^2}{\partial x^2} = (\frac{x}{2\kappa t})^2\frac{\partial^2}{\partial \zeta^2} + \frac{1}{2\kappa t}\frac{\partial}{\partial \zeta} = \frac{\zeta}{\kappa t}\frac{\partial^2}{\partial \zeta^2} + \frac{1}{2\kappa t}\frac{\partial}{\partial \zeta},$$

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial \zeta}\frac{\partial \zeta}{\partial t} = -\frac{x^2}{4\kappa t^2}\frac{\partial}{\partial \zeta} = -\frac{\zeta}{t}\frac{\partial}{\partial \zeta}, \tag{1.9}$$

we can write the PDE (1.5) for $u$ as

$$-\frac{\zeta}{t}f' = \kappa \cdot [\frac{\zeta}{\kappa t}f'' + \frac{1}{2\kappa t}f'], \tag{1.10}$$

where $f' = df/d\zeta$. Multiplying both sides by $t/\zeta$,

$$-f' = f''(\zeta) + \frac{1}{2\zeta}f', \qquad \text{or} \qquad \frac{f''}{f'} = -(1 + \frac{1}{2\zeta}). \tag{1.11}$$

Using $(\ln f')' = f''/f'$ and integrating the above equation once, we get

$$\ln f' = -\zeta - \frac{1}{2}\ln \zeta + C, \tag{1.12}$$

where $C$ is an integration constant. This can be written as

$$f' = \frac{Ke^{-\zeta}}{\sqrt{\zeta}}, \tag{1.13}$$

where $K = e^C$. Integrating it again, we obtain

$$u = f(\zeta) = A\text{erf}(\sqrt{\zeta}) + B = A\text{erf}(\frac{x}{\sqrt{4\kappa t}}) + B, \tag{1.14}$$

where

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-\xi^2}d\xi, \tag{1.15}$$

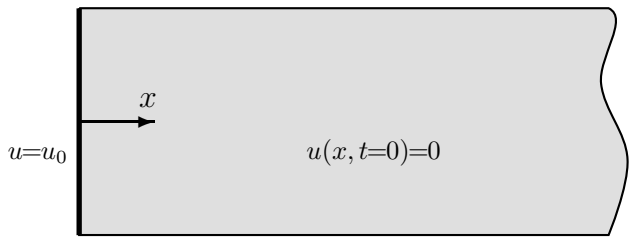



Figure 1.4: Heat transfer near a dyke through a semi-infinite medium.

is the error function and $\xi$ is a dummy variable. $A = K\sqrt{\pi}$ and $B$ are constants that can be determined from appropriate boundary conditions. This is the basic solution in the infinite or semi-infinite domain. The solution is generic because we have not used any of the boundary conditions or initial conditions.

---

**Example 1.4:** For the heat conduction problem near a magma dyke in a semi-infinite domain, we can determine the constants $A$ and $B$. Let $x = 0$ be the centre of the rising magma dyke so that its temperature is constant at the temperature $u_0$ of the molten magma, while the temperature at the far field is $u = 0$ (as we are only interested in the temperature change in this case).

The boundary condition at $x = 0$ requires that

$$A\text{erf}(0) + B = u_0.$$

We know that $\text{erf}(0) = 0$, this means that $B = u_0$. From the initial condition $u(x, t = 0) = 0$, we have

$$A \lim_{t\to 0} \text{erf}(\frac{x}{\sqrt{4\kappa t}}) + u_0 = 0.$$

Since $x/\sqrt{4\kappa t} \to \infty$ as $t \to 0$ and $\text{erf}(\infty) = 1$, we get $A + u_0 = 0$, or $A = -u_0$. Thus the solution becomes

$$u = u_0[1 - \text{erf}(\frac{x}{\sqrt{4\kappa t}})] = u_0\text{erfc}(\frac{x}{\sqrt{4\kappa t}}),$$

where $\text{erfc}(x) = 1 - \text{erf}(x)$ is the complementary error function. The distribution of $u/u_0$ is shown in Fig. 1.5.

---

From the above solution, we know that the temperature variation becomes significant in the region of $x = d$ such that $d/\sqrt{\kappa t} \approx 1$ at a

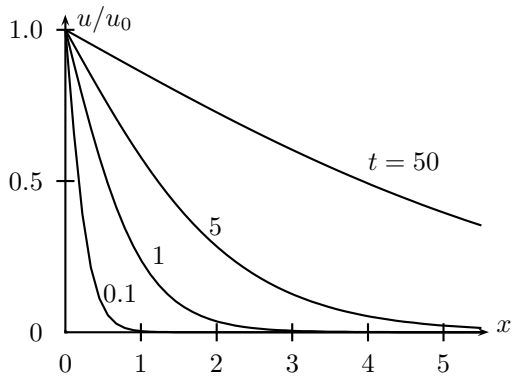



Figure 1.5: Distribution of $u(x,t)/u_0$ with $\kappa = 0.25$.

given time $t$. That is

$$d = \sqrt{\kappa t}, \tag{1.16}$$

which defines a typical length scale. Alternatively, for a given length scale $d$ of interest, we can estimate the time scale $t = \tau$ at which the temperature becomes significant. That is

$$\tau = \frac{d^2}{\kappa}. \tag{1.17}$$

This means that it will take four times longer if the size of the hot body $d$ is doubled. Now let us see what it means in our example. We know that the thermal conductivity is $K \approx 3$ W/m K for rock, its density is $\rho \approx 2700$ Kg/m$^3$ and its specific heat capacity $c_p \approx 1000$ J/kg K. Thus, the thermal diffusivity of solid rock is

$$\kappa = \frac{K}{\rho c_p} \approx \frac{3}{2700 \times 1000} \approx 1.1 \times 10^{-6} \text{m}^2/\text{s}. \tag{1.18}$$

For $d \approx 1$m, the time scale of cooling is

$$\tau = \frac{d^2}{\kappa} \approx \frac{1}{1.1 \times 10^{-6}} \approx 8.8 \times 10^5 \text{ seconds} \approx 10 \text{ days}. \tag{1.19}$$

For a larger hot body $d = 100$ m, then that time scale is $\tau = 10^5$ days or 270 years. This estimate of the cooling time scale is based on the assumption that no more heat is supplied. However, in reality, there is usually a vast magma reservoir below to supply hot magma constantly, and this means that the cooling time is at the geological time scale over millions of years.

### 1.2.2 Functional and Integral Equations

Though most mathematical models are written as partial different equations, however, sometimes it might be convenient to write them in terms of integral equations, and these integral forms can be discretised to obtained various numerical methods. For example, the Fredholm integral equation can be generally written as

$$u(x) + \lambda \int_a^b K(x, \eta) y(\eta) d\eta = v(x) y(x), \tag{1.20}$$

where $u(x)$ and $v(x)$ are known functions of $x$, and $\lambda$ is constant. The kernel $K(x, \eta)$ is also given. The aim is to find the solution $y(x)$. This type of problem can be extremely difficult to solve and analytical solutions exist in only a few very simple cases. We will provide a simple introduction to integral equations later in this book.

Sometimes, the problem you are trying to solve does not give a mathematical model in terms of dependent variance such as $u$ which is a function of spatial coordinates $(x, y, z)$ and time $t$, rather they lead to a functional (or a function of the function $u$); this kind of problem is often linked to the calculus of variations.

For example, finding the shortest path between any given points on the Earth's surface is a complicated geodesic problem. If we idealise the Earth's surface as a perfect sphere, then the shortest path joining any two different points is a great circle through both points. How can we prove this is true? Well, the proof is based on the Euler-Lagrange equation of a functional $\psi(u)$

$$\frac{\partial \psi}{\partial u} = \frac{d}{dx}(\frac{\partial \psi}{\partial u'}), \tag{1.21}$$

where $u$ a function of $x$, $u' = du/dx$, and $\psi$ a function of $u(x)$. The detailed proof will be given later in this book in the chapter dealing with calculus of variations.

### 1.2.3 Statistical Models

Both differential equations and integral equations are the mathematical models for continuum systems. Other systems are discrete and different mathematical models are needed, though they could reduce to certain forms of differential equations if some averaging is carried out. On the other hand, many systems have intrinsic randomness, thus the description and proper modelling require statistical models, or to be more specific, geostatistical models in earth sciences.

For example, suppose that we carried out some field work and made some observations of a specific quantity, say, density of rocks, over a

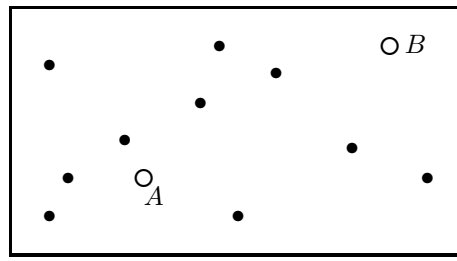



Figure 1.6: Field observations (marked with •) and interpolation for inaccessible locations (marked with ∘).

large area shown in Fig. 1.6. Some locations are physically inaccessible (marked with ∘) and the value at the inaccessible locations can only be estimated. A proper estimation is very important. The question that comes naturally is how to estimate the values at these locations using the observation at other locations? How should we start? As we already have some measured data $\rho_i (i = 1, 2, ..., n)$, the first sensible thing is to use the sample mean or average of $<\rho_i>$ as the approximation to the value at the inaccessible locations. If we do this, then any two inaccessible locations will have the same value (because the sample data do not change). This does not help if there are quite a few inaccessible locations.

Alternatively, we can use the available observed data to construct a surface by interpolation such as linear or cubic splines. There, different inaccessible locations may have different values, which will provide more information about the region. This is obviously a better estimation than the simple sample mean. Thinking along these lines, can we use the statistical information from the sample data to build a statistical model so that we can get a better estimation? The answer is yes. In geostatistics, this is the well-known Kriging interpolation technique which uses the spatial correlation, or semivariogram, among the observation data to estimate the values at new locations. This will be discussed in detail in the chapter about geostatistics.

## 1.3 Numerical Methods

### 1.3.1 Numerical Integration

In the solution (1.14) of problem (1.5), there is a minor problem in the evaluation of the solution $u$. That is the error function $\text{erf}(x)$ because

it is a special function whose integral cannot be expressed as a simple explicit combination of basic functions, it can only be expressed in terms of a quadrature. In order to get its values, we have to either use approximations or numerical integration. You can see that even with seemingly precise solution of a differential equation, it is quite likely that it may involve some special functions.

Let us try to evaluate $\mathrm{erf}(1)$. From advanced mathematics, we know its exact value is $\mathrm{erf}(1) = 0.8427007929...$, but how do we calculate it numerically?

---

**Example 1.5:** In order to estimate $\mathrm{erf}(1)$, we first try to use a naive approach by estimating the area under the curve $f(x) = \frac{2}{\sqrt{\pi}} e^{-x^2}$ in the interval $[0, 1]$ shown in Fig. 1.7. We then divide the interval into 5 equally-spaced thin strips with $h = \Delta x = x_{i+1} - x_i = 1/5 = 0.2$. We have six values of $f_i = f(x_i)$ at $x_i = hi (i = 0, 1, ..., 5)$, and they are

$$f_0 = 1.1284,\ f_1 = 1.084,\ f_2 = 0.9615,$$

$$f_3 = 0.7872,\ f_4 = 0.5950,\ f_5 = 0.4151.$$

Now we can either use the rectangular area under the curve (which underestimates the area) or the area around the curve plus the area under curve (which overestimates the area). Their difference is the tiny area about the curve which could still make some difference. If we use the area under the curve, we have the estimation of the total area as

$$A_1 \approx 0.2(f_1 + f_2 + f_3 + f_4 + f_5) \approx 0.7686.$$

The other approach gives

$$A_2 \approx 0.2(f_0 + f_1 + f_2 + f_3 + f_4) \approx 0.91125.$$

Both are about $8\%$ from the true value $\mathrm{erf}(1) \approx 0.8247$. If we take the average of these two estimates, we get

$$A_3 \approx \frac{A_1 + A_2}{2} \approx 0.8399,$$

which is much better, but still $0.3\%$ from the true value. This average method is essentially equivalent to using $f_i = (f_{i-1} + f_i)/2$ to approximate the value of $f(x)$ in each interval.

---

As you can see from this example, the way you discretise the integrand to estimate the integral numerically can have many variants, subsequently affecting the results significantly. There are much better

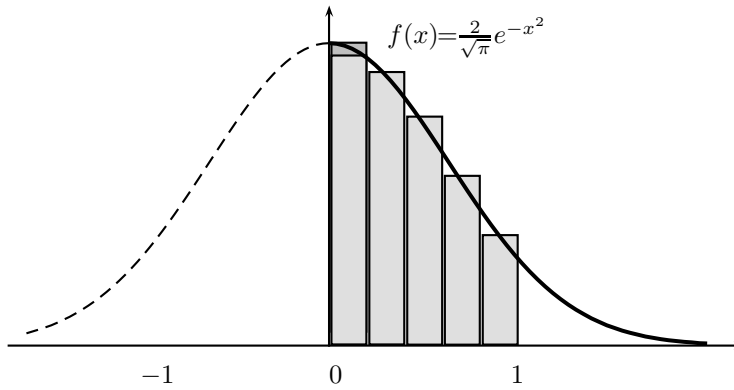



Figure 1.7: Naive numerical integration.

ways to carry out the numerical integration, notably the Gaussian integration which requires only seven points to get the accuracy of about 9th decimal place or 0.0000001% (see Appendix B). All these techniques will be explained in detail in the part dealing with numerical integration and numerical methods.

### 1.3.2 Numerical Solutions of PDEs

The diffusion equation (1.1) is a relatively simple parabolic equation. If we add a reaction term (source or sink) to this equation, we get the classical reaction-diffusion equation

$$\frac{\partial u}{\partial t} = D\nabla^2 u + \gamma u(1-u), \tag{1.22}$$

where $u$ can be concentration and any other quantities. $\gamma u(1-u)$ is the reaction term and $\gamma$ is a constant. This seemingly simple partial differential equation is in fact rather complicated for mathematical analysis because the equation is nonlinear due to the term $-\gamma u^2$. However, numerical technique can be used and it is relatively straightforward to obtain solutions (see the chapter on reaction-diffusion system in this book). This mathematical model can produce intriguing patterns due to its intrinsic instability under appropriate conditions.

In the two-dimensional case, we have

$$\frac{\partial u}{\partial t} = D(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) + \gamma u(1-u). \tag{1.23}$$

Using the finite difference method to be introduced in the second half of this book, we can solve this equation on a 2-D domain. Fig. 1.8

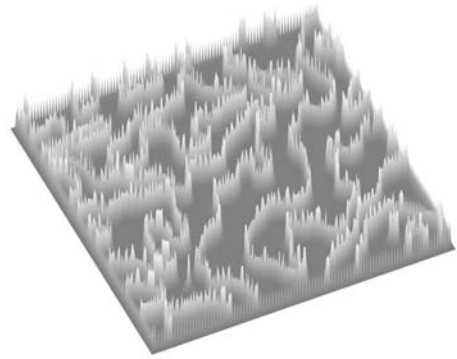


Figure 1.8: Pattern formation of reaction-diffusion equation (1.23) with $D = 0.2$ and $\gamma = 0.5$.

shows the stable pattern generated by Eq.(1.23) with $D = 0.2$ and $\gamma = 0.5$. The initial condition is completely random, say, $u(x, y, t = 0) = \mathsf{rand}(n, n) \in [0, 1]$ where $n \times n$ is the size of the grid used in the simulations. The function $\mathsf{rand}()$ is a random number generator and all the random numbers are in the range of 0 to 1.

We can see that a beautiful and stable pattern forms automatically from an initially random configuration. This pattern formation mechanism has been used to explain many pattern formation phenomena in nature shown in Fig. 1.9, including patterns on zebra skin, tiger skin and sea shell, zebra leaf (green and yellow), and zebra stones. For example, the zebra rocks have reddish-brown and white bands first discovered in Australia. It is believed that the pattern is generated by dissolution and precipitation of mineral bands such as iron oxide as mineral in the fluid percolating through the porous rock.

The instability analysis of pattern formation and the numerical method for solving such nonlinear reaction-diffusion system will be discussed in detail later in this book.

## 1.4 Topics in This Book

So far, we have presented you with a taster of the diverse topics presented in this book. From a mathematical modelling point of view, the topics in earth sciences are vast, therefore, we have to make a decision to select topics and limit the number of chapters so that the book remains concise and yet comprehensive enough to include important

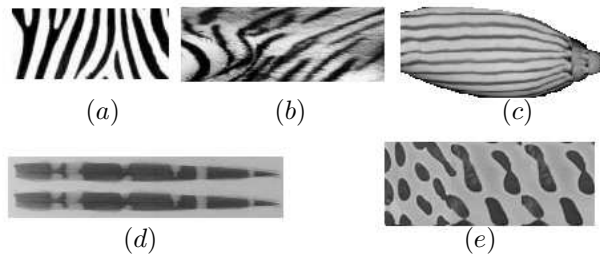



Figure 1.9: Pattern formation in nature: (a) zebra skin; (b) tiger skin; (c) sea shell; (c) zebra grass; and (e) zebra stone.

topics and popular numerical algorithms.

We use a 'theorem-free' approach which is thus informal from the viewpoint of rigorous mathematical analysis. There are two reasons for such an approach: firstly we can focus on presenting the results in a smooth flow, rather than interrupting them by the proof of certain theorems; and secondly we can put more emphasis on developing the analytical skills for building mathematical models and the numerical algorithms for solving mathematical equations.

We also provide dozens of worked examples with step-by-step derivations and these examples are very useful in understanding the fundamental principles and to develop basic skills in mathematical modelling.

The book is organised into three parts: Part I (mathematical methods), Part II (numerical algorithms), and Part III (applications). In Part I, we present you with the fundamental mathematical methods, including calculus and complex variable (Chapter 2), vector and matrix analysis (Chapter 3), ordinary differential equations and integral transform (Chapter 4), and partial differential equations and classic mathematical models (Chapter 5). We then introduce the calculus of variations and integral equations (Chapter 6). The final two chapters (7 and 8) in Part I are about the probability and geostatistics.

In Part II, we first present the root-finding algorithms and numerical integration (Chapter 9), then we move on to study the finite difference and finite volume methods (Chapters 10 and 11), and finite element methods (Chapter 12).

In Part III, we discuss the topics as applications in earth sciences. We first briefly present the reaction-diffusion system (Chapter 13), then present in detail the elasticity, fracture mechanics and poroelasticity (Chapter 14). We end this part by discussing flow in porous media including groundwater flow and pollutant transport (Chapter 15).

There are two appendices at the end of the book. Appendix A

is a summary of the mathematical formulae used in this book, and the second appendix provides some programs (Matlab and Octove) so that readers can experiment with them and carry out some numerical simulations. At the end of each chapter, there is a list of references for further reading.

## References

Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

Gershenfeld N., *Nature of Mathematical Modeling*, Cambridge University Press, (1998).

Kardestruncer H. and Norrie D. H., *Finite Element Handbook*, McGraw-Hill, (1987).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Murch B. W. and Skinner B. J., *Geology Today - Understanding Our Planet*, John Wiley & Sons, (2001).

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

Smith G. D., *Numerical Solution of Partial Differential Equations*, Oxford University Press, (1974).

Wang H. F., *Theory of Linear Poroelasticity: with applications to geomechanics and hydrogeology*, Princeton Univ. Press, (2000).

# Chapter 2

# Calculus and Complex Variables

The preliminary requirements for this book are the pre-calculus foundation mathematics. We assume that the readers are familiar with these preliminaries, therefore, we will only review some of the important concepts of differentiation, integration, Jacobian and multiple integrals.

## 2.1 Calculus

### 2.1.1 Set Theory

#### Definitions

Let us first introduce some of the basic concepts in set theory. A set is any well-defined collection of objects or elements, and the elements are the members or objects in a set. We conventionally use the upper-case letters to denote sets and lower-case letters for elements, and the listed elements are enclosed in the curly brace {}. The membership in a set is denoted using $\in$, thus

$$x \in \mathscr{A}, \tag{2.1}$$

means that '$x$ is a member of the set $\mathscr{A}$', while

$$\boldsymbol{x} \notin \mathscr{A}, \tag{2.2}$$

means that '$x$ is not a member of the set $\mathscr{A}$'. A special set is the empty set or null set which has no element and is denoted by

$$\emptyset = \{\}, \tag{2.3}$$

which should not be confused with a non-empty set which consists of a single element $\{0\}$.

We say that $\mathscr{A}$ is a subset of $\mathscr{B}$ if $a \in \mathscr{A}$ implies that $a \in \mathscr{B}$. That is to say that all the members of $\mathscr{A}$ are also members of $\mathscr{B}$. We denote this relationship as

$$\mathscr{A} \subseteq \mathscr{B}. \tag{2.4}$$

If all the members of $\mathscr{A}$ are also members of $\mathscr{B}$, but there exists at least one element $b$ such that $b \in \mathscr{B}$ while $b \notin \mathscr{A}$, we say $\mathscr{A}$ is a *proper* subset of $\mathscr{B}$, and denote this relationship as

$$\mathscr{A} \subset \mathscr{B}. \tag{2.5}$$

When combining sets, we say that $\mathscr{A}$ union $\mathscr{B}$, denoted by

$$\mathscr{A} \cup \mathscr{B}$$

forms a set of all elements that are in $\mathscr{A}$, *or* $\mathscr{B}$, *or both*. On the other hand, $\mathscr{A}$ intersect $\mathscr{B}$, written as

$$\mathscr{A} \cap \mathscr{B},$$

is the set of all elements that are in both $\mathscr{A}$ *and* $\mathscr{B}$.

A universal set $\Omega$ is the set that consists of all the elements under consideration. The complement set of $\mathscr{A}$ or *not* $\mathscr{A}$, denoted by $\bar{\mathscr{A}}$, is the set of all the elements that are not in $\mathscr{A}$. The set $\mathscr{A} - \mathscr{B}$ or $\mathscr{A}$ minus $\mathscr{B}$ is the set of elements that are in $\mathscr{A}$ and not in $\mathscr{B}$, this is equivalent to removing or substracting from $\mathscr{A}$ all the elements that are in $\mathscr{B}$. This leads to

$$\mathscr{A} - \mathscr{B} = \mathscr{A} \cap \bar{\mathscr{B}}, \tag{2.6}$$

and

$$\bar{\mathscr{A}} = \Omega - \mathscr{A}. \tag{2.7}$$

---

**Example 2.1:** For two sets

$$\mathscr{A} = \{2, 3, 5, 7\}, \ \ \mathscr{B} = \{2, 4, 6, 8, 10\},$$

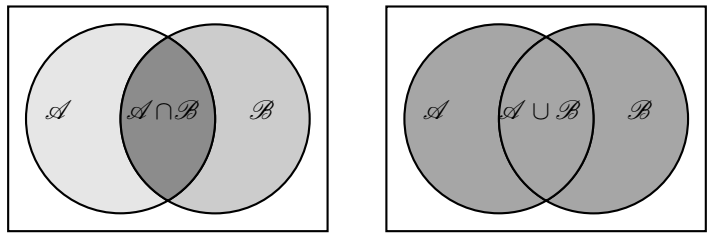
and a universal set

$$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

it is straightforward to check that

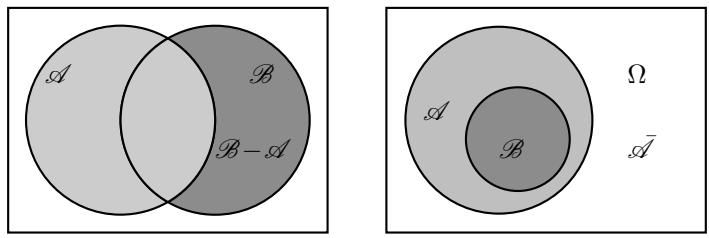
$$\mathscr{A} \subset \Omega, \ \ \mathscr{B} \subset \Omega, \ \ \mathscr{A} - \mathscr{B} = \{3, 5, 7\},$$
$$\mathscr{A} \cup \mathscr{B} = \{2, 3, 4, 5, 6, 7, 8, 10\}, \ \ \mathscr{A} \cap \mathscr{B} = \{2\},$$
$$\bar{\mathscr{A}} = \Omega - \mathscr{A} = \{1, 4, 6, 8, 9, 10\},$$

and finally $\bar{\mathscr{A}} \cap \mathscr{A} = \emptyset$.




(a) (b)

Figure 2.1: Venn diagrams: a) $\mathscr{A} \cap \mathscr{B}$, b) $\mathscr{A} \cup \mathscr{B}$.




(a) (b)

Figure 2.2: Venn diagrams: a) $\mathscr{B} - \mathscr{A}$, b) $\bar{\mathscr{A}} = \Omega - \mathscr{A}$ and $\mathscr{B} \subset \mathscr{A}$.

### Venn Diagrams

The set operations seem too abstract, there is a better way to represent such mathematical operations between sets, that is to use the Venn diagrams as the topological representation. Fig. 2.1 represents the intersect $\mathscr{A} \cap \mathscr{B}$ and union $\mathscr{A} \cup \mathscr{B}$, while Fig. 2.2 represents $\mathscr{B} - \mathscr{A}$, $\bar{\mathscr{A}} = \Omega - \mathscr{A}$ and $\mathscr{B} \subset \mathscr{A}$.

### Special Sets

Some common sets in mathematics are used so often that they deserve special names or notations. These include:

- $\mathscr{N} = \{1, 2, 3, ...\}$ or $\mathscr{N} = \{0, 1, 2, ...\}$ denotes the set of all natural numbers;
- $\mathscr{Z} = \{..., -2, -1, 0, 1, 2, ...\}$ is the set of all integers;
- $\mathscr{P} = \{2, 3, 5, 7, 11, ...\}$ is the set of all primes;
- $\mathscr{Q} = \{\frac{m}{n} : m, n \in \mathscr{Z}, n \neq 0\} = \{\frac{-5}{11}, 2, \frac{7}{19}, ...\}$ is the set of all rational numbers;

- $\mathscr{R}$ is the set of all real numbers consisting of all rational numbers and all irrational numbers such as $\sqrt{2}, \pi, e$;
- $\mathscr{C}$ is the set of complex numbers in the form of $a + bi$ where $a, b \in \mathscr{R}$ and $i = \sqrt{-1}$.

All these sets have an infinite number of elements. From the definitions, it is straightforward to check that

$$\mathscr{P} \subset \mathscr{N} \subset \mathscr{Z} \subset \mathscr{Q} \subset \mathscr{R} \subset \mathscr{C}. \tag{2.8}$$

### 2.1.2 Differentiation and Integration

For a known function $y = f(x)$ or a curve as shown in Figure 2.3, the gradient or slope of the curve at any point $P(x, y)$ is defined as

$$\frac{dy}{dx} \equiv \frac{df(x)}{dx} \equiv f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}, \tag{2.9}$$

on the condition that there exists such a limit at $P$. This gradient or limit is the first derivative of the function $f(x)$ at $P$. If the limit does not exist at a point $P$ when $\Delta x$ approaches zero, then we say that the function is non-differentiable at $P$. By convention, the limit of the infinitesimal change $\Delta x$ is denoted as the differential $dx$. Thus, the above definition can also be written as

$$dy = df = \frac{df(x)}{dx} dx = f'(x)dx, \tag{2.10}$$

which can be used to calculate the change in $dy$ caused by the small change of $dx$. The primed notation $'$ and standard notation $\frac{d}{dx}$ can be used interchangeably, and the choice is purely one of convenience.

The second derivative of $f(x)$ is defined as the gradient of $f'(x)$, that is to say,

$$\frac{d^2y}{dx^2} \equiv f''(x) = \frac{df(x)}{dx}. \tag{2.11}$$

The higher derivatives can be defined in a similar manner. Thus,

$$\frac{d^3y}{dx^3} \equiv f'''(x) = \frac{df''(x)}{dx}, \ \ ..., \ \ \frac{d^ny}{dx^n} \equiv f^{(n)} = \frac{df^{(n-1)}}{dx}. \tag{2.12}$$

In the case of $y' = f'(x_*) = 0$, the point corresponds to a stationary point. The solution of $f'(x_*) = 0$ is also called a first-order critical point. The condition $f''(x_*) = 0$ defines a second-order critical point, called an inflection point.

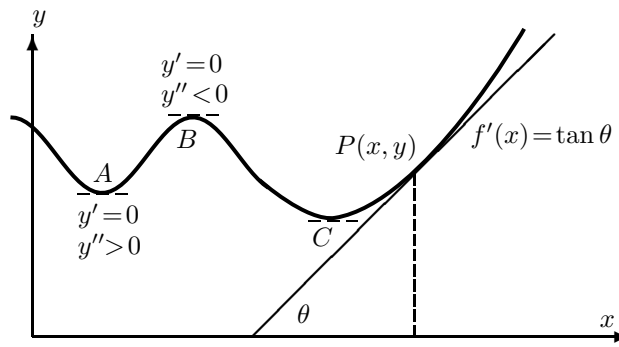



Figure 2.3: Gradient of a function $y = f(x)$

The maxima or minima of a function only occur at stationary points such as $A, B$ and $C$ shown in Fig. 2.3. The local maximum (such as point $B$) occurs at

$$f'(x_*) = 0, \qquad f''(x_*) < 0, \tag{2.13}$$

while the local minima (such as points $A$ and $C$) occurs at

$$f'(x_*) = 0, \qquad f''(x_*) > 0. \tag{2.14}$$

The point $C$ is a global minimum, while point $A$ is just a local minimum. In the case of $f'(x_*) = f''(x_*) = 0$, the point does not mean a minimum or maximum. For example, $y = x^3$, we know that $y'(0) = y''(0) = 0$. It is not a local minimum or maximum, but just an inflection point in this case.

**Differentiation Rules**

If a more complicated function $f(x)$ can be written as a product of two simpler functions $u(x)$ and $v(x)$, we can derive a differentiation rule using the definition from first principles. We have

$$\frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \frac{d[u(x)v(x)]}{dx}$$

$$= \lim_{\Delta x \to 0} \frac{u(x + \Delta x)v(x + \Delta x) - u(x)v(x)}{\Delta x},$$

$$= \lim_{\Delta x \to 0} [u(x+\Delta x)\frac{v(x+\Delta x) - v(x)}{\Delta x} + v(x)\frac{u(x+\Delta x) - u(x)}{\Delta x}]$$

$$= u(x)\frac{dv}{dx} + \frac{du}{dx}v(x), \tag{2.15}$$

which can be written in a compact form using primed notations

$$f'(x) = (uv)' = u'v + uv'. \tag{2.16}$$

If we differentiate this equation again and again, we can get Leibnitz's Theorem for differentiations

$$\frac{d^n(uv)}{dx^n} = u^{(n)}v + nu^{(n-1)}v' + ... + \binom{n}{r}u^{(n-r)}v^{(r)} + ... + uv^{(n)}, \tag{2.17}$$

where the coefficients are the same as the binomial coefficients

$${}^nC_r \equiv \binom{n}{r} = \frac{n!}{r!(n-r)!}. \tag{2.18}$$

If a function $f(x)$ [for example, $f(x) = e^{\sin(x)}$] can be written as a function of another function $g(x)$, or $f(x) = f[g(x)]$ [for example, $f(x) = e^{g(x)}$ and $g(x) = \sin(x)$], then we have

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta g}\frac{\Delta g}{\Delta x}, \tag{2.19}$$

which leads to the following chain rule

$$f'(x) = \frac{df}{dg}\frac{dg}{dx}, \tag{2.20}$$

or

$$\{f[g(x)]\}' = f'[g(x)] \cdot g'(x). \tag{2.21}$$

In our example, we have $f'(x) = (e^{\sin(x)})' = e^{\sin(x)}\cos(x)$. If we use $1/v$ instead of $v$ in Eq.(2.16) and $(1/v)' = -v'/v^2$, we have the following differentiation rule for quotients:

$$(\frac{u}{v})' = \frac{u'v - uv'}{v^2}. \tag{2.22}$$

---

**Example 2.2:** Assume an ideal mountain has the shape of a Gaussian function

$$f(x) = e^{-ax^2}, \quad -\infty < x < \infty, \quad a > 0,$$

at what location is the slope steepest? We know that

$$f'(x) = -2axe^{-ax^2},$$

and

$$f''(x) = \frac{df'(x)}{dx} = -2ae^{-ax^2} + 4a^2x^2e^{-ax^2}.$$

The steepest (or maximum) slope or first derivative $f'(x)$ occurs at the location where the second-derivative $f''(x)$ is zero. That is

$$f''(x) = (-2a + 4a^2x^2)e^{-ax^2} = 0.$$

Since $\exp(-ax^2) > 0$, we have

$$-2a + 4a^2x^2 = 0,$$

or

$$x = \pm\frac{1}{\sqrt{2a}},$$

so we have two solutions.

---

The derivatives of various functions are listed in Table 2.1.

Table 2.1: First Derivatives

| $f(x)$ | $f'(x)$ |
|---|---|
| $x^n$ | $nx^{n-1}$ |
| $e^x$ | $e^x$ |
| $a^x (a > 0)$ | $a^x \ln a$ |
| $\ln x$ | $\frac{1}{x}$ |
| $\log_a x$ | $\frac{1}{x \ln a}$ |
| $\sinh x$ | $\cosh x$ |
| $\cosh x$ | $\sinh x$ |
| $\tan x$ | $\sec^2 x$ |
| $\sin^{-1} x$ | $\frac{1}{\sqrt{1-x^2}}$ |
| $\cos^{-1} x$ | $-\frac{1}{\sqrt{1-x^2}}$ |
| $\tan^{-1} x$ | $\frac{1}{1+x^2}$ |

### Implicit Differentiation

The above differentiation rules still apply in the case when there is no simple explicit function form $y = f(x)$ as a function of $x$ only. Let us look at an example.

**Example 2.2:** To find $dy/dx$ given

$$y^2 - \sin(x)e^y = 0.$$

In this case, we can differentiate the equation term by term with respect to $x$ so that we can obtain the derivative $dy/dx$ which is in general a function of both $x$ and $y$. Now we have

$$2y\frac{dy}{dx} - \cos(x)e^y - \sin(x)e^y\frac{dy}{dx} = 0,$$

which leads to

$$\frac{dy}{dx} = \frac{\cos(x)e^y}{2y - \sin(x)e^y}.$$

---

### Integration

Integration can be viewed as the inverse of differentiation. The integration $F(x)$ of a function $f(x)$ satisfies

$$\frac{dF(x)}{dx} = f(x), \tag{2.23}$$

or

$$F(x) = \int_{x_0}^{x} f(\xi)d\xi, \tag{2.24}$$

where $f(x)$ is called the integrand, and the integration starts from $x_0$ (arbitrary) to $x$. In order to avoid any potential confusion, it is conventional to use a dummy variable (say, $\xi$) in the integrand. As we know, the geometrical meaning of the first derivative is the gradient of the function $f(x)$ at a point $P$, the geometrical representation of an integral

$$\int_a^b f(\xi)d\xi$$

(with lower integration limit $a$ and upper integration limit $b$) is the area under the curve $f(x)$ enclosed by $x$-axis in the region $x \in [a, b]$. In this case, the integral is called a definite integral as the limits are given. For the definite integral, we have

$$\int_a^b f(x)dx = \int_{x_0}^b f(x)dx - \int_{x_0}^a f(x)dx = F(b) - F(a). \tag{2.25}$$

The difference $F(b) - F(a)$ is often written in a compact form $F\Big|_a^b \equiv F(b) - F(a)$. As $F'(x) = f(x)$, we can also write the above equation as

$$\int_a^b f(x)dx = \int_a^b F'(x)dx = F(b) - F(a). \tag{2.26}$$

Since the lower limit $x_0$ is arbitrary, the change or shift of the lower limit will lead to an arbitrary constant $c$. When the lower limit is not explicitly given, the integral is called an indefinite integral

$$\int f(x)dx = F(x) + c, \tag{2.27}$$

where $c$ is the constant of integration. The integrals of some of the common functions are listed in Table 2.2.

Table 2.2: Integrals

| $f(x)$ | $\int f(x)dx$ |
|---|---|
| $x^n (n \neq -1)$ | $\frac{x^{n+1}}{n+1}$ |
| $\frac{1}{x}$ | $\ln \lvert x \rvert$ |
| $e^x$ | $e^x$ |
| $\frac{1}{a^2+x^2}$ | $\frac{1}{a} \tan^{-1} \frac{x}{a}$ |
| $\frac{1}{a^2-x^2}$ | $\frac{1}{2a} \ln \frac{a+x}{a-x}$ |
| $\frac{1}{x^2-a^2}$ | $\frac{1}{2a} \ln \frac{x-a}{x+a}$ |
| $\frac{1}{\sqrt{a^2-x^2}}$ | $\sin^{-1} \frac{x}{a}$ |
| $\frac{1}{\sqrt{x^2+a^2}}$ | $\ln(x + \sqrt{x^2+a^2})$ [or $\sinh^{-1} \frac{x}{a}$] |
| $\frac{1}{\sqrt{x^2-a^2}}$ | $\ln(x + \sqrt{x^2-a^2})$ [or $\cosh^{-1} \frac{x}{a}$] |
| $\sinh x$ | $\cosh x$ |
| $\cosh x$ | $\sinh x$ |
| $\tanh x$ | $\ln \cosh x$ |

**Integration by Parts**

From the differentiation rule $(uv)' = uv' + u'v$, we get

$$uv' = (uv)' - u'v. \tag{2.28}$$

Integrating both sides, we have

$$\int u \frac{dv}{dx} dx = uv - \int \frac{du}{dx} v dx, \tag{2.29}$$

in the indefinite form. It can also be written in the definite form as

$$\int_a^b u\frac{dv}{dx}dx = [uv]\Big|_a^b + \int_a^b v\frac{du}{dx}dx. \tag{2.30}$$

The integration by parts is a very powerful method for evaluating integrals. Many complicated integrands can be rewritten as a product of two simpler functions so that their integrals can be easily obtained using integration by parts.

---

**Example 2.3:** The integral of $I = \int x^{99} \ln x \; dx$ can be obtained by setting $v' = x^{99}$ and $u = \ln x$. Hence, $v = \frac{x^{100}}{100}$ and $u' = \frac{1}{x}$. We now have

$$I = \int x \ln x dx = \frac{x^{100}\ln x}{100} - \int \frac{x^{100}}{100}\frac{1}{x}dx = \frac{x^{100}\ln x}{100} - \frac{x^{100}}{10000}.$$

In general, we have

$$\int x^n \ln x = \frac{x^{n+1}\ln x}{n+1} - \frac{x^{n+1}\ln x}{(n+1)^2}.$$

for $n \neq -1$.

---

Other important methods of integration include the substitution and reduction methods. Readers can refer to any book that is dedicated to advanced calculus.

### Taylor Series and Power Series

From

$$\int_a^b f(x)dx = F(b) - F(a), \tag{2.31}$$

and $\frac{dF}{dx} = F' = f(x)$, we have

$$\int_{x_0}^{x_0+h} f'(x)dx = f(x_0+h) - f(x_0), \tag{2.32}$$

which means that

$$f(x_0+h) = f(x_0) + \int_{x_0}^{x_0+h} f'(x)dx. \tag{2.33}$$

If $h$ is not too large or $f'(x)$ does not vary dramatically, we can approximate the integral as

$$\int_{x_0}^{x_0} f'(x)dx \approx f'(x_0)h. \tag{2.34}$$

Thus, we have the first-order approximation to $f(x_0 + h)$

$$f(x_0 + h) \approx f(x_0) + hf'(x_0). \tag{2.35}$$

This is equivalent to saying that any change from $x_0$ to $x_0 + h$ is approximated by a linear term $hf'(x_0)$. If we repeat the procedure for $f'(x)$, we have

$$f'(x_0 + h) \approx f'(x_0) + hf''(x_0), \tag{2.36}$$

which is a better approximation than $f'(x_0 + h) \approx f'(x_0)$. Following the same procedure for higher order derivatives, we can reach the $n$-th order approximation

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \frac{h^3}{3!}f'''(x_0)$$

$$+ ... + \frac{h^n}{n!}f^{(n)}(x_0) + R_{n+1}(h), \tag{2.37}$$

where $R_{n+1}(h)$ is the error of this approximation and the notation means that the error is about the same order as $(n+1)$-th term in the series. This is the well-known Taylor theorem and it has many applications. In deriving this formula, we have implicitly assumed that all the derivatives $f'(x), f''(x), ..., f^{(n)}(x)$ exist. In almost all the applications we meet, this is indeed the case. For example, $\sin(x)$ and $e^x$, all the orders of the derivatives exist. If we continue the process to infinity, we then reach the infinite power series and the error $\lim_{n\to\infty} R_{n+1} \to 0$ if the series converges. The end results are the Maclaurin series. For example,

$$e^x = 1 + x + \frac{x^2}{2!} + ... + \frac{x^n}{n!} + ..., \quad (x \in \mathcal{R}), \tag{2.38}$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ..., \quad (x \in \mathcal{R}), \tag{2.39}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ..., \quad (x \in \mathcal{R}), \tag{2.40}$$

and

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - ..., \quad x \in (-1, 1]. \tag{2.41}$$

### 2.1.3 Partial Differentiation

The derivative defined earlier is for function $f(x)$ which has only one independent variable $x$, and the gradient will generally depend on the location $x$. For functions $f(x, y)$ of two variables $x$ and $y$, their gradient will depend on both $x$ and $y$ in general. In addition, the gradient or

rate of change will also depend on the direction (along $x$-axis or $y$-axis or any other directions). For example, the function $f(x, y) = x(y - 1)$ has different gradients at $(0, 0)$ along $x$-axis and $y$-axis. The gradients along the positive $x$- and $y$- directions are called the partial derivatives with respect to $x$ and $y$, respectively. They are denoted as $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, respectively.

The partial derivative of $f(x, y)$ with respect to $x$ can be calculated assuming that $y$ =constant. Thus, we have

$$\frac{\partial f(x,y)}{\partial x} \equiv f_x \equiv \frac{\partial f}{\partial x}|_y = \lim_{\Delta x \to 0, y=\text{const}} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}. \quad (2.42)$$

Similarly, we have

$$\frac{\partial f(x,y)}{\partial y} \equiv f_y \equiv \frac{\partial f}{\partial y}|_x = \lim_{\Delta y \to 0, x=\text{const}} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}. \quad (2.43)$$

The notation $\frac{\partial}{\partial x}\big|_y$ emphasises that the fact that $y$ is held constant. The subscript notation $f_x$ (or $f_y$) emphasizes that the derivative is carried out with respect to $x$ (or $y$). Mathematicians like to use the subscript forms as they are simpler notations and can be easily generalised. For example,

$$f_{xx} = \frac{\partial^2 f}{\partial x^2}, \qquad f_{xy} = \frac{\partial^2 f}{\partial x \partial y}. \quad (2.44)$$

Since $\Delta x \Delta y = \Delta y \Delta x$, we have $f_{xy} = f_{yx}$.

For any small change $\Delta f = f(x + \Delta x, y + \Delta y) - f(x, y)$ due to $\Delta x$ and $\Delta y$, the total infinitesimal change $df$ can be written as

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy. \quad (2.45)$$

If $x$ and $y$ are functions of another independent variable $\xi$, then the above equation leads to the following chain rule

$$\frac{df}{d\xi} = \frac{\partial f}{\partial x}\frac{dx}{d\xi} + \frac{\partial f}{\partial y}\frac{dy}{d\xi}, \quad (2.46)$$

which is very useful in calculating the derivatives in parametric form or for change of variables. If a complicated function $f(x)$ can be written in terms of simpler functions $u$ and $v$ so that $f(x) = g(x, u, v)$ where $u(x)$ and $v(x)$ are known functions of $x$, then we have the generalised chain rule

$$\frac{dg}{dx} = \frac{\partial g}{\partial x} + \frac{\partial g}{\partial u}\frac{du}{dx} + \frac{\partial g}{\partial v}\frac{dv}{dx}. \quad (2.47)$$

When differentiating an integral

$$\Phi(x) = \int_a^b \phi(x, y)dy, \tag{2.48}$$

with fixed integration limits $a$ and $b$, we have

$$\frac{\partial \Phi(x)}{\partial x} = \int_a^b \frac{\partial \phi(x, y)}{\partial x} dy. \tag{2.49}$$

When differentiating the integrals with the limits being functions of $x$,

$$I(x) = \int_{v(x)}^{u(x)} \psi(x, \tau)d\tau = \Psi[x, u(x)] - \Psi[x, v(x)], \tag{2.50}$$

the following formula is useful:

$$\frac{dI}{dx} = \int_{v(x)}^{u(x)} \frac{\partial \psi}{\partial x} d\tau + [\psi(x, u(x))\frac{du}{dx} - \psi(x, v(x))\frac{dv}{dx}]. \tag{2.51}$$

This formula can be derived using the chain rule

$$\frac{dI}{dx} = \frac{\partial I}{\partial x} + \frac{\partial I}{\partial u}\frac{du}{dx} + \frac{\partial I}{\partial v}\frac{dv}{dx}, \tag{2.52}$$

where $\frac{\partial I}{\partial u} = \psi(x, u(x))$, and $\frac{\partial I}{\partial v} = -\psi(x, v(x))$.

### 2.1.4 Multiple Integrals

As the integration of a function $f(x)$ corresponds to the area enclosed under the function between integration limits, this can extend to the double integral and multiple integrals. For a function $f(x, y)$, the double integral is defined as

$$F = \int_\Omega f(x, y)dA, \tag{2.53}$$

where $dA$ is the infinitesimal element of the area, and $\Omega$ is the region for integration. The simplest form of $dA$ is $dA = dxdy$ in Cartesian coordinates. In order to emphasise the double integral in this case, the integral is often written as

$$I = \iint_\Omega f(x, y)dxdy. \tag{2.54}$$

**Example 2.4:** The area moment of inertia of a thin rectangular plate, with the width $a$ and the depth $b$, is defined by

$$I = \iint_\Omega y^2 dS = \iint_\Omega y^2 dxdy.$$

The plate can be divided into four equal parts, and we have

$$I = 4\int_0^{a/2}[\int_0^{b/2} y^2 dy]dx = 4\int_0^{a/2} \frac{1}{3}(\frac{b}{2})^3 dx = \frac{b^3}{6}\int_0^{a/2} dx = \frac{ab^3}{12}.$$

### 2.1.5 Jacobian

Sometimes it is necessary to change variables when evaluating an integral. For a simple one-dimensional integral, the change of variables from $x$ to a new variable $v$ (say) leads to $x = x(v)$. This is relatively simple as $dx = \frac{dx}{dv}dv$, and we have

$$\int_{x_a}^{x_b} f(x)dx = \int_a^b f(x(v))\frac{dx}{dv}dv, \tag{2.55}$$

where the integration limits change so that $x(a) = x_a$ and $x(b) = x_b$. Here the extra factor $dx/dv$ in the integrand is referred to as the Jacobian.

For a double integral, it is more complicated. Assuming $x = x(\xi, \eta)$, $y = y(\xi, \eta)$, we have

$$\iint f(x, y)dxdy = \iint f(\xi\eta)|J|d\xi d\eta, \tag{2.56}$$

where $J$ is the Jacobian. That is

$$J \equiv \frac{\partial(x, y)}{\partial(\xi, \eta)} = \begin{vmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{vmatrix} = \begin{vmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{vmatrix}. \tag{2.57}$$

The notation $\partial(x, y)/\partial(\xi, \eta)$ is just a useful shorthand. This is equivalent to saying that the change of the infinitesimal area $dA = dxdy$ becomes

$$dxdy = |\frac{\partial(x, y)}{\partial(\xi, \eta)}|d\xi d\eta = |\frac{\partial x}{\partial \xi}\frac{\partial y}{\partial \eta} - \frac{\partial x}{\partial \eta}\frac{\partial y}{\partial \xi}|d\xi d\eta. \tag{2.58}$$

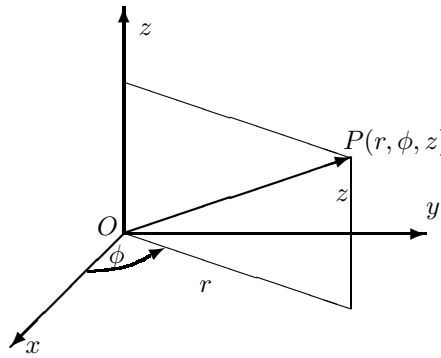



Figure 2.4: Cylindrical polar coordinates.

---

**Example 2.5:** When transforming from $(x, y)$ to polar coordinates $(r, \theta)$, we have the following relationships

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

Thus, the Jacobian is

$$J = \frac{\partial(x,y)}{\partial(r,\theta)} = \frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial x}{\partial \theta}\frac{\partial y}{\partial r}$$

$$= \cos\theta \times r\cos\theta - (-r\sin\theta) \times \sin\theta = r[\cos^2\theta + \sin^2\theta] = r.$$

Thus, an integral in $(x, y)$ will be transformed into

$$\iint \phi(x,y)dxdy = \iint \phi(r\cos\theta, r\sin\theta)rdrd\theta.$$

---

In a similar fashion, the change of variables in triple integrals gives

$$V = \iiint_{\Omega} \phi(x,y,z)dxdydz = \iiint_{\omega} \psi(\xi,\eta,\zeta)|J|d\xi d\eta d\zeta, \qquad (2.59)$$

and

$$J \equiv \frac{\partial(x,y,z)}{\partial(\xi,\eta,\zeta)} = \begin{vmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} & \frac{\partial z}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} & \frac{\partial z}{\partial \eta} \\ \frac{\partial x}{\partial \zeta} & \frac{\partial y}{\partial \zeta} & \frac{\partial z}{\partial \zeta} \end{vmatrix}. \qquad (2.60)$$

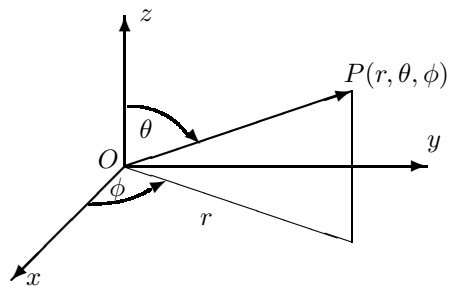



Figure 2.5: Spherical polar coordinates.

For cylindrical polar coordinates $(r, \phi, z)$ as shown in Figure 2.4, we have

$$x = r\cos\phi, \qquad y = r\sin\phi, \qquad z = z. \tag{2.61}$$

The Jacobian is therefore

$$J = \frac{\partial(x, y, z)}{\partial(r, \phi, z)} = \begin{vmatrix} \cos\phi & \sin\phi & 0 \\ -r\sin\phi & r\cos\phi & 0 \\ 0 & 0 & 1 \end{vmatrix} = r. \tag{2.62}$$

For spherical polar coordinates $(r, \theta, \phi)$ as shown in Figure 2.5, where $\theta$ is the zenithal angle between the $z$-axis and the position vector $\mathbf{r}$, and $\phi$ is the azimuthal angle, we have

$$x = r\sin\theta\cos\phi, \qquad y = r\sin\theta\sin\phi, \qquad z = r\cos\theta. \tag{2.63}$$

Therefore, the Jacobian is

$$J = \begin{vmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ r\cos\theta\cos\phi & r\cos\theta\sin\phi & -r\sin\theta \\ -r\sin\theta\sin\phi & r\sin\theta\cos\phi & 0 \end{vmatrix} = r^2\sin\theta. \tag{2.64}$$

Thus, the volume element change in the spherical system is

$$dxdydz = r^2\sin\theta dr d\theta d\phi. \tag{2.65}$$

## 2.2 Complex Variables

Although all the quantities are real variables in the physical world, however, it is sometimes desirable or even necessary to use complex variables in mathematical modelling. In fact, the techniques based on complex variables are among the most powerful methods for mathematical analysis and solutions of mathematical models.

### 2.2.1 Complex Numbers and Functions

Mathematically speaking, a complex number $z$ is a generalised set or the order pair of two real numbers $(a, b \in \mathcal{R})$, written in the form of

$$z = a + ib, \qquad i^2 = -1, \qquad a, b \in \mathcal{R}, \tag{2.66}$$

which consists of the real part $\Re(z) = a$ and the imaginary part $\Im(z) = b$. It can also be written as the order pair of real numbers $(a, b)$. The addition and substraction of a two complex numbers are defined as

$$(a + ib) \pm (c + id) = (a \pm c) + i(b \pm d). \tag{2.67}$$

The multiplication and division of two complex numbers are in the similar way for polynomial expansions.

$$(a + ib) \cdot (c + id) = (ac - bd) + i(ad + bc), \tag{2.68}$$

and

$$\frac{a + ib}{c + id} = \frac{ac + bd}{c^2 + d^2} + i\frac{bc - ad}{c^2 + d^2}. \tag{2.69}$$

Two complex numbers are equal $a + ib = c + id$ if and only if $a = c$ and $b = d$. The complex conjugate or simply conjugate $\bar{z}$ of $z = a + ib$ is defined as

$$\bar{z} = a - ib. \tag{2.70}$$

The order pair $(a, b)$, similar to a vector, implies that a geometrical representation of a complex number $a + ib$ by the point in an ordinary Euclidean plane with $x$-axis being the real axis and $y$-axis being the imaginary axis $(iy)$. This plane is called the complex plane. The vector representation starts from $(0, 0)$ to the point $(a, b)$. The length of the vector is called the magnitude or modulus or the absolute value of the complex number

$$r = |z| = \sqrt{a^2 + b^2}. \tag{2.71}$$

The angle $\theta$ that the vector makes with the positive real axis is called the argument,

$$\theta = \arg z. \tag{2.72}$$

In fact, we may replace $\theta$ by $\theta + 2n\pi$ $(n \in \mathcal{N})$. The value range $-\pi < \theta \leq \pi$ is called the principal argument of $z$, and it is usually denoted by Arg$z$. In the complex plane, the complex number can be written as

$$z = re^{i\theta} = r\cos(\theta) + ir\sin(\theta). \tag{2.73}$$

This polar form of $z$ and its geometrical representation can result in the Euler's formula which is very useful in the complex analysis

$$e^{i\theta} = \cos(\theta) + i\sin(\theta). \tag{2.74}$$

The Euler formula can be proved using the power series. For any $z \in \mathcal{C}$, we have the power series

$$e^z = 1 + z + \frac{z^2}{2!} + ... + \frac{z^n}{n!} + ..., \tag{2.75}$$

and for a special case $z = i\theta$, we have

$$e^{i\theta} = 1 + i\theta - \frac{\theta^2}{2!} + \frac{i\theta^3}{3!} - ... = (1 - \frac{\theta^2}{2!} + ...) + i(\theta - \frac{\theta^3}{3!} + ...). \tag{2.76}$$

Using the power series

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - ..., \qquad \cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - ..., \tag{2.77}$$

we get

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{2.78}$$

For $\theta = \pi$, this leads to a very interesting formula

$$e^{i\pi} + 1 = 0. \tag{2.79}$$

For two complex numbers $z_1 = r_1 e^{i\alpha_1}$ and $z_2 = r_2 e^{i\alpha_2}$, it is straightforward to show that

$$z_1 z_2 = r_1 r_2 e^{i(\alpha_1+\alpha_2)} = r_1 r_2[\cos(\alpha_1 + \alpha_2) + i\sin(\alpha_1 + a_2)]. \tag{2.80}$$

Generally, the de Moivre's formula can be obtained by repetitive use of the above formula $n$ times with $\alpha_1 = \alpha_2 = \theta$

$$[\cos(\theta) + i\sin(\theta)]^n = \cos(n\theta) + i\sin(n\theta). \tag{2.81}$$

### 2.2.2 Analytic Functions

Any function of real variables can be extended to the function of complex variables in the same form while treating the real numbers $x$ as $x + i0$. For example, $f(x) = x^2, x \in \mathcal{R}$ becomes $f(z) = z^2, z \in \mathcal{C}$. Any complex function $f(z)$ can be written as

$$f(z) = f(x + iy) = \Re(f(z)) + i\Im(f(z)) = u(x, y) + iv(x, t), \tag{2.82}$$

where $u(x, y)$ and $v(x, y)$ are real-valued functions of two real variables.

A function $f(z)$ is called analytic at $z_0$ if $f'(z)$ exists for all $z$ in some $\epsilon-$neighborhood of $z_0$ or differentiable in some open disk $|z - z_0| < \epsilon$. If $f(z) = u + iv$ is analytic at every point in a domain $\Omega$, then $u(x, y)$ and $v(x, y)$ satisfies the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{2.83}$$

Conversely, if $u$ and $v$ in $f(z) = u + iv$ satisfy the Cauchy-Riemann equation at all points in a domain, then the complex function $f(z)$ is analytic in the same domain. For example, the elementary power function $w = z^n, (n > 1)$ is analytic on the whole plane, $w = \rho e^{i\phi}, z = re^{i\theta}$, then

$$\rho = r^n, \phi = n\theta. \tag{2.84}$$

The logarithm is also an elementary function $w = \ln z$

$$\ln z = \ln |z| + i \arg(z) = \ln r + i(\theta + w\pi k), \tag{2.85}$$

which has infinitely many values, due to the multiple values of $\theta$, with the difference of $2\pi ik, (k = 0, \pm 1, \pm 2, ...)$. If we use the principal argument Arg$z$, then we have the principal logarithm function

$$\text{Ln}(z) = \ln |z| + \text{Arg}z. \tag{2.86}$$

If we differentiate the Cauchy-Riemann equations and use $\partial^2 u/\partial x \partial y = \partial^2 u/\partial y \partial x$, we have

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y}, \tag{2.87}$$

and

$$\frac{\partial^2 v}{\partial y^2} = -\frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial^2 v}{\partial x \partial y}. \tag{2.88}$$

Adding these two, we have

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{2.89}$$

A similar argument for $v$ leads to the following theorem: For given analytic function $f(z) = u + iv$, then both $u$ and $v$ satisfy the Laplace equations

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0, \qquad \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0. \tag{2.90}$$

This is to say, both real and imaginary parts of an analytic function are harmonic.

## 2.3 Complex Integrals

Given a function $f(z)$ that is continuous on a piecewise smooth curve $\Gamma$, then the integral over $\Gamma$ is $\int_\Gamma f(z)dz$, called a contour or line integral of $f(z)$. This integral has similar properties to the real integral

$$\int_\Gamma [\alpha f(z) + \beta g(z)]dz = \alpha \int_\Gamma f(z)dz + \beta \int_\Gamma g(z)dz. \tag{2.91}$$

If $F(z)$ is analytic and $F'(z) = f(z)$ is continuous along a curve $\Gamma$, then

$$\int_a^b f(z)dz = F[z(b)] - F[z(a)]. \tag{2.92}$$

### 2.3.1 Cauchy's Integral Theorem

We say a path is simply closed if its end points and initial points coincide and the curve does not cross itself. For an analytic function $f(z) = u(x, y) + iv(x, y)$, the integral on a simply closed path

$$I = \int_\Gamma (u + iv)(dx + idy)] = \int_\Gamma (udx - vdy) + i\int_\Gamma (vdx + udy). \tag{2.93}$$

By using the Green theorem, this becomes

$$I = \int_\Omega (-\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x})dxdy + i\int_\Omega (\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y})dxdy. \tag{2.94}$$

From the Cauchy-Riemann equations, we know that both integrals are zero. Thus, we have Cauchy's Integral Theorem, which states that the integral of any analytic function $f(z)$ on a simply closed path $\Gamma$ in a simply connected domain $\Omega$ is zero. That is

$$\int_\Gamma f(z)dz = 0.$$

This theorem is very important as it has interesting consequences. If the closed path is decomposed into two paths with reverse directions $\Gamma_1$ and $\Gamma_2$, then $\Gamma_1$ and $-\Gamma_2$ form a closed path, which leads to

$$\int_{\Gamma_1} f(z)dz = \int_{\Gamma_2} f(z)dz. \tag{2.95}$$

That is to say that the integral over any curve between two points is independent of the path. This property becomes very useful for evaluation of integrals. The integral of $f(z)/(z - z_0)$ over any simply closed path $\Gamma$ enclosing a point $z_0$ in the domain $\Omega$ is,

$$\frac{1}{2\pi i}\oint_\Gamma \frac{f(z)}{z - z_0}dz = f(z_0).$$

Similarly, for higher derivatives, we have

$$\oint_\Gamma \frac{f(z)}{(z - z_0)^{n+1}}dz = \frac{2\pi i f^{(n)}(z_0)}{n!}.$$

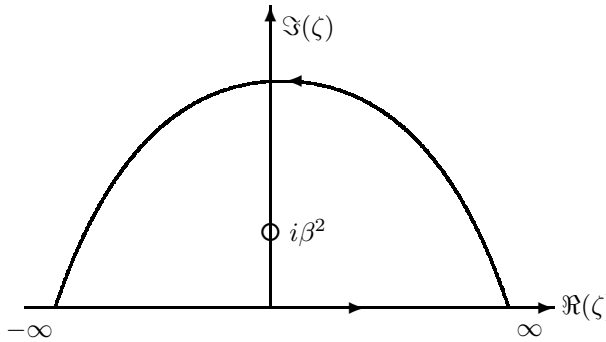



Figure 2.6: Contour for the integral $I(\alpha, \beta)$.

### 2.3.2 Residue Theorem

A function $f(z)$ has a pole or singularity of order $n$ at $z = z_0$ if $f(z)$ is not analytic at $z = z_0$ but $(z - z_0)^n f(z)$ is analytic at $z = z_0$. This suggests that $f(z)$ can be expanded as a power series, called Laurant series

$$f(z) = \sum_{k=-n}^{\infty} \alpha_k (z - z_0)^k, \tag{2.96}$$

where $\alpha_k$ are the coefficients. The most important coefficient is probably $\alpha_{-1}$ which is called the residue of $f(z)$ at the pole $z = z_0$. If $f(z)$ has a pole of order $N$ at $z_0$, the following formula gives a quick way to calculate the residue

$$\text{Res} f(z)|_{z_0} = \frac{1}{(N-1)!} \lim_{z \to z_0} \frac{d^{N-1}[(z - z_0)^N f(z)]}{dz^{N-1}}. \tag{2.97}$$

For any analytic $f(z)$ function in a domain $\Omega$ except isolated singularities at finite points $z_1, z_2, ..., z_N$, the residue theorem states

$$\oint_\Gamma f(z)dz = 2\pi i \sum_{k=1}^{N} \text{Res} f(z)|_{z_k},$$

where $\Gamma$ is a simple closed path enclosing all these isolated points or poles. The residue theorem serves as a powerful tool for calculating some real integrals and summation of series, especially when the integrand is a function of sin and cos that can be changed into complex integral. The real integral $\int_{-\infty}^{\infty} \psi(x)dx$ becomes $2\pi i$ multiplying the sum of the residues of $\psi(x)$ at the poles in the upper half-space. Let us look at an example.

**Example 2.6:** In order to evaluate the integral

$$I(\alpha,\beta)=\int_{-\infty}^{\infty}\frac{e^{i\alpha^2\zeta}}{x^2+\beta^4}d\zeta,$$

it is necessary to construct a contour (see Figure 2.6). As the function $\phi=e^{i\alpha^2\zeta}/(\beta^4+\zeta^2)$ has two poles $\zeta=+i\beta^2$ and $-i\beta^2$ from $\beta^4+\zeta^2=0$, and only one pole $\zeta=+i\beta^2$ is in the upper half plane, we can construct a contour to encircle the pole at $\zeta=i\beta^2$ by adding an additional arc at the infinity ($\zeta\to\infty$) on the upper half plane. Combining the arc with the horizontal line from the integral limits from $-\infty$ to $\infty$ along the $\zeta$-axis, a contour is closed. Hence, we have

$$\phi=\frac{e^{i\alpha^2\zeta}/(\zeta+i\beta^2)}{\zeta-i\beta^2}=\frac{f(\zeta)}{\zeta-i\beta^2},$$

where $f(\zeta)=e^{i\alpha^2\zeta}/(\zeta+i\beta^2)$. Using the residue theorem, we have

$$I=2\pi i[f(\zeta=i\beta^2)]=2\pi i\frac{e^{-\alpha^2\beta^2}}{i\beta^2+i\beta^2}=\pi\frac{e^{-\alpha^2\beta^2}}{\beta^2}.$$

In a special case when $\alpha=0$, we have $\int_{-\infty}^{\infty}\frac{1}{\zeta^2+\beta^4}d\zeta=\frac{\pi}{\beta^2}$.

# References

Abramowitz M. and Stegun I. A., *Handbook of Mathematical Functions*, Dover Publication, (1965).

Courant R. and Hilbert, D., *Method of Mathematical Physics*, 2 volumes, Wiley-Interscience, New York, (1953, 1962).

Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, 3rd Edition, Cambridge University Press (2006).

# Chapter 3

# Vectors and Matrices

Many quantities such as force, velocity, and deformation in nature are vectors that have both a magnitude and a direction. The manipulation of vectors is often associated with matrices. In this chapter, we will first discuss the basics of vector and matrix analysis, then we will discuss tensors.

## 3.1 Vectors

A vector $\mathbf{x}$ is a set of ordered numbers $\mathbf{x} = (x_1, x_2, ..., x_n)$, where its components $x_1, x_2, ..., x_n$ are real numbers. All these vectors forms a $n$-dimensional vector space $\mathcal{V}^n$. To add two vectors $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, ..., y_n)$, simply add their corresponding components,

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, ..., x_n + y_n), \tag{3.1}$$

and the sum is also a vector. The addition of vectors has commutability ($\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$) and associativity [$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$]. Zero vector $\mathbf{0}$ is a special case in that all its components are zeros. The multiplication of a vector $\mathbf{x}$ with a scalar or constant $\alpha$ is carried out by the multiplication of each component,

$$\alpha\mathbf{y} = (\alpha y_1, \alpha y_2, ..., \alpha y_n). \tag{3.2}$$

Thus, $-\mathbf{y} = (-y_1, -y_2, ..., -y_n)$. In addition, $(\alpha\beta)\mathbf{y} = \alpha(\beta\mathbf{y})$ and $(\alpha + \beta)\mathbf{y} = \alpha\mathbf{y} + \beta\mathbf{y}$.

Two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$ are said to be linearly independent if $\alpha\mathbf{a} + \beta\mathbf{b} = 0$ implies that $\alpha = \beta = 0$. If $\alpha, \beta$ are not all zeros, then these two vectors are linearly dependent. Two linearly dependent vectors are parallel ($\mathbf{a} /\!\!/ \mathbf{b}$) to each other. Three linearly dependent vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are in the same plane.

### 3.1.1 Dot Product and Norm

The dot product or inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + ... + x_n y_n = \sum_{i=1}^{n} x_i y_i, \tag{3.3}$$

which is a real number. The length or norm of a vector $x$ is the root of the dot product of the vector itself,

$$|\mathbf{x}| = \|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{3.4}$$

When $\|\mathbf{x}\| = 1$, then it is a unit vector. It is straightforward to check that the dot product has the following properties:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}, \qquad \mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}, \tag{3.5}$$

and

$$(\alpha \mathbf{x}) \cdot (\beta \mathbf{y}) = (\alpha\beta)\mathbf{x} \cdot \mathbf{y}, \tag{3.6}$$

where $\alpha, \beta$ are constants.

If $\theta$ is the angle between two vectors $\mathbf{a}$ and $\mathbf{b}$, then the dot product can also be written

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos(\theta), \qquad 0 \leq \theta \leq \pi. \tag{3.7}$$

If the dot product of these two vectors is zero or $\cos(\theta) = 0$ (i.e., $\theta = \pi/2$), then we say that these two vectors are orthogonal.

Rearranging Eq.(3.7), we obtain a formula to calculate the angle $\theta$ between two vectors

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}. \tag{3.8}$$

Since $\cos(\theta) \leq 1$, then we get the useful Cauchy-Schwartz inequality:

$$\|\mathbf{a} \cdot \mathbf{b}\| \leq \|\mathbf{a}\| \, \|\mathbf{b}\|. \tag{3.9}$$

Any vector $\mathbf{a}$ in a $n$-dimensional vector space $\mathcal{V}^n$ can be written as a combination of a set of $n$ independent basis vectors or orthogonal spanning vectors $\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}$, so that

$$\mathbf{a} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + ... + \alpha_n \mathbf{e}_n = \sum_{i=1}^{n} \alpha_i \mathbf{e}_i, \tag{3.10}$$

where the coefficients/scalars $\alpha_1, \alpha_2, ..., \alpha_n$ are the components of $\mathbf{a}$ relative to the basis $\mathbf{e_1}, \mathbf{e_2}..., \mathbf{e_n}$. The most common basis vectors are the orthogonal unit vectors. In a three-dimensional case, they are $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0$, $\mathbf{k} = (0, 0, 1)$ for three $x$-, $y$-, $z$-axis, and thus $\mathbf{x} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$. The three unit vectors satisfy $\mathbf{i} \cdot \mathbf{j} = \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{i} = 0$.

### 3.1.2 Cross Product

The dot product of two vectors is a scalar or a number. On the other hand, the cross product or outer product of two vectors is a new vector

$$\mathbf{c} = \mathbf{a} \times \mathbf{b} = (x_2y_3 - x_3y_2)\mathbf{i} + (x_3y_1 - x_1y_3)\mathbf{j} + (x_1y_2 - x_2y_1)\mathbf{k}, \quad (3.11)$$

this is usually written as

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix}$$

$$= \begin{vmatrix} x_2 & x_3 \\ y_2 & y_3 \end{vmatrix} \mathbf{i} + \begin{vmatrix} x_3 & x_1 \\ y_3 & y_1 \end{vmatrix} \mathbf{j} + \begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix} \mathbf{k}. \quad (3.12)$$

The angle between $\mathbf{a}$ and $\mathbf{b}$ can also be expressed as

$$\sin\theta = \frac{\|\mathbf{a} \times \mathbf{b}\|}{\|\mathbf{a}\| \, \|\mathbf{b}\|}. \quad (3.13)$$

In fact, the norm $\|\mathbf{a} \times \mathbf{b}\|$ is the area of the parallelogram formed by $\mathbf{a}$ and $\mathbf{b}$. The vector $\mathbf{c} = \mathbf{a} \times \mathbf{b}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, following a right-hand rule. It is straightforward to check that the cross product has the following properties:

$$\mathbf{x} \times \mathbf{y} = -\mathbf{y} \times \mathbf{x}, \quad (\mathbf{x} + \mathbf{y}) \times \mathbf{z} = \mathbf{x} \times \mathbf{z} + \mathbf{y} \times \mathbf{z}, \quad (3.14)$$

and

$$(\alpha\mathbf{x}) \times (\beta\mathbf{y}) = (\alpha\beta)\mathbf{x} \times \mathbf{y}. \quad (3.15)$$

A very special case is $\mathbf{a} \times \mathbf{a} = \mathbf{0}$. For unit vectors, we have

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}, \qquad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \qquad \mathbf{k} \times \mathbf{i} = \mathbf{j}. \quad (3.16)$$

---

**Example 3.1:** For two 3-D vectors $\mathbf{a} = (2, -3, 1)$ and $\mathbf{b} = (2, 5, 0)$, their dot product is

$$\mathbf{a} \cdot \mathbf{b} = 2 \times 2 + (-3) \times 5 + 1 \times 0 = -11.$$

As their moduli are

$$||\mathbf{a}|| = \sqrt{2^2 + (-3)^2 + 1^2} = \sqrt{14}, \quad ||\mathbf{b}|| = \sqrt{2^2 + 5^2 + 0} = \sqrt{29},$$

we can calculate the angle $\theta$ between the two vectors. We have

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||||\mathbf{b}||} = \frac{-11}{\sqrt{14}\sqrt{29}},$$

or

$$\theta = \cos^{-1} \frac{-11}{\sqrt{406}} \approx 123.09^\circ.$$

Their cross product is

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = (-3 \times 0 - 1 \times 5, 1 \times 2 - 2 \times 0, 2 \times 5 - (-3) \times 2)$$
$$= (-5, 2, 16).$$

The vector $\mathbf{v}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ because

$$\mathbf{a} \cdot \mathbf{v} = 2 \times (-5) + (-3) \times 2 + 1 \times 16 = 0,$$

and

$$\mathbf{b} \cdot \mathbf{v} = 2 \times (-5) + 5 \times 2 + 0 \times 16 = 0.$$

---

### 3.1.3 Differentiation of Vectors

The differentiation of a vector is carried out over each component and treating each component as the usual differentiation of a scalar. Thus, from a position vector

$$\mathbf{P}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}, \tag{3.17}$$

we can write its velocity as

$$\mathbf{v} = \frac{d\mathbf{P}}{dt} = \dot{x}(t)\mathbf{i} + \dot{y}(t)\mathbf{j} + \dot{z}(t)\mathbf{k}, \tag{3.18}$$

and acceleration as

$$\mathbf{a} = \frac{d^2\mathbf{P}}{dt^2} = \ddot{x}(t)\mathbf{i} + \ddot{y}(t)\mathbf{j} + \ddot{z}(t)\mathbf{k}, \tag{3.19}$$

where $\dot{()} = d()/dt$. Conversely, the integral of $\mathbf{v}$ is

$$\mathbf{P} = \int \mathbf{v}dt + \mathbf{c}, \tag{3.20}$$

where $\mathbf{c}$ is a constant.

From the basic definition of differentiation, it is easy to check that the differentiation of vectors has the following properties:

$$\frac{d(\alpha\mathbf{a})}{dt} = \alpha\frac{d\mathbf{a}}{dt}, \quad \frac{d(\mathbf{a} \cdot \mathbf{b})}{dt} = \frac{d\mathbf{a}}{dt} \cdot \mathbf{b} + \mathbf{a} \cdot \frac{d\mathbf{b}}{dt}, \tag{3.21}$$

and

$$\frac{d(\mathbf{a} \times \mathbf{b})}{dt} = \frac{d\mathbf{a}}{dt} \times \mathbf{b} + \mathbf{a} \times \frac{d\mathbf{b}}{dt}. \tag{3.22}$$

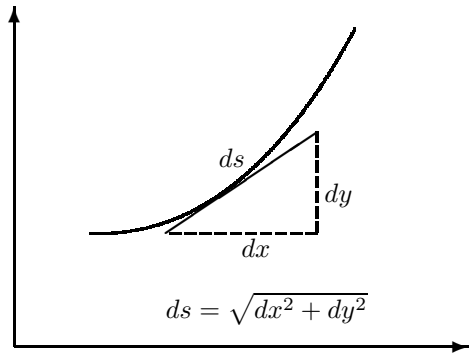



Figure 3.1: Arc length along a curve.

### 3.1.4 Line Integral

An important class of integrals in this context is the line integral which integrates along a curve $\mathbf{r}(x, y, z) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. For example, in order to calculate the arc length $L$ of curve $\mathbf{r}$ as shown in Figure 3.1, we have to use the line integral.

$$L = \int_{s_0}^{s} ds = \int_{s_0}^{s} \sqrt{dx^2 + dy^2} = \int_{x_0}^{x} \sqrt{1 + (\frac{dy}{dx})^2} dx. \tag{3.23}$$

---

**Example 3.2:** The arc length of the parabola $y(x) = \frac{1}{2}x^2$ from $x = 0$ to $x = 1$ is given by

$$L = \int_0^1 \sqrt{1 + y'^2} dx = \int_0^1 \sqrt{1 + x^2} dx$$

$$= \frac{1}{2}[x\sqrt{1 + x^2} + \ln(x + \sqrt{1 + x^2})]\Big|_0^1 = \frac{1}{2}[\sqrt{2} - \ln(\sqrt{2} - 1)] \approx 1.15.$$

---

### 3.1.5 Three Basic Operators

Three important operators commonly used in vector analysis, especially in the formulation of the finite element methods, are the gradient operator (grad or $\nabla$), the divergence operator (div or $\nabla\cdot$) and the curl operator (curl or $\nabla\times$).

Sometimes, it is useful to calculate the directional derivative of a scalar $\phi$ at the point $(x, y, z)$ in the direction of $\mathbf{n}$

$$\frac{\partial \phi}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla \phi = \frac{\partial \phi}{\partial x} \cos(\alpha) + \frac{\partial \phi}{\partial y} \cos(\beta) + \frac{\partial \phi}{\partial z} \cos(\gamma), \tag{3.24}$$

where $\mathbf{n} = (\cos\alpha, \cos\beta, \cos\gamma)$ is a unit vector and $\alpha, \beta, \gamma$ are the directional angles. Generally speaking, the gradient of any scalar function $\phi$ of $x, y, z$ can be written in a similar way,

$$\text{grad}\phi = \nabla \phi = \frac{\partial \phi}{\partial x}\mathbf{i} + \frac{\partial \phi}{\partial y}\mathbf{j} + \frac{\partial \phi}{\partial z}\mathbf{k}. \tag{3.25}$$

This is equivalent to applying the del operator $\nabla$ to the scalar function $\phi$

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}. \tag{3.26}$$

The direction of the gradient operator on a scalar field gives a vector field. The gradient operator has the following properties:

$$\nabla(\alpha\psi + \beta\phi) = \alpha\nabla\psi + \beta\nabla\phi, \qquad \nabla(\psi\phi) = \psi\nabla\phi + \phi\nabla\psi, \tag{3.27}$$

where $\alpha, \beta$ are constants and $\psi, \phi$ are scalar functions.

For a vector field

$$\mathbf{u}(x, y, z) = u_1(x, y, z)\mathbf{i} + u_2(x, y, z)\mathbf{j} + u_3(x, y, z)\mathbf{k}, \tag{3.28}$$

the application of the operator $\nabla$ can lead to either a scalar field or vector field depending on how the del operator applies to the vector field. The divergence of a vector field is the dot product of the del operator $\nabla$ and $\mathbf{u}$

$$\text{div } \mathbf{u} = \nabla \cdot \mathbf{u} = \frac{\partial \mathrm{u}_1}{\partial \mathrm{x}} + \frac{\partial \mathrm{u}_2}{\partial \mathrm{y}} + \frac{\partial \mathrm{u}_3}{\partial \mathrm{z}}, \tag{3.29}$$

and the curl of $\mathbf{u}$ is the cross product of the del operator and the vector field $\mathbf{u}$

$$\text{curl } \mathbf{u} = \nabla \times \mathbf{u} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ u_1 & u_2 & u_3 \end{vmatrix}. \tag{3.30}$$

One of the most common operators in engineering and science is the Laplacian operator

$$\nabla^2\phi \equiv \nabla \cdot (\nabla\phi) = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2}, \tag{3.31}$$

for Laplace's equation

$$\Delta\phi \equiv \nabla^2\phi = 0. \tag{3.32}$$

### 3.1.6 Some Important Theorems

The Green theorem is an important theorem, especially in fluid dynamics and finite element analysis. For a vector field $\mathbf{Q} = u\mathbf{i} + v\mathbf{j}$ in a 2-D region $\Omega$ with the boundary $\Gamma$ and the unit outer normal $\mathbf{n}$ and unit tangent $\mathbf{t}$. The theorems connecting the integrals of divergence and curl with other integrals can be written as Gauss's theorem:

$$\iiint_{\Omega} (\nabla \cdot \mathbf{Q}) d\Omega = \iint_{S} \mathbf{Q} \cdot \mathbf{n} dS, \tag{3.33}$$

and Stokes's theorem:

$$\iint_{S} (\nabla \times \mathbf{Q}) \cdot \mathbf{k} d\mathbf{S} = \oint_{\Gamma} \mathbf{Q} \cdot \mathbf{t} d\Gamma = \oint_{\Gamma} \mathbf{Q} \cdot d\mathbf{r}. \tag{3.34}$$

In our simple 2-D case, this becomes

$$\oint (udx + vdy) = \iint_{\Omega} (\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}) dxdy. \tag{3.35}$$

For any scalar functions $\psi$ and $\phi$, the useful Green's first identity can be written as

$$\oint_{\partial\Omega} \psi \nabla \phi d\Gamma = \int_{\Omega} (\psi \nabla^2 \phi + \nabla \psi \cdot \nabla \phi) d\Omega, \tag{3.36}$$

where $d\Omega = dxdydz$. By using this identity twice, we get Green's second identity

$$\oint_{\partial\Omega} (\psi \nabla \phi - \phi \nabla \psi) d\Gamma = \int_{\Omega} (\psi \nabla^2 \phi - \phi \nabla^2 \psi) d\Omega. \tag{3.37}$$

## 3.2 Matrix Algebra

### 3.2.1 Matrix

Matrices are widely used in almost all engineering subjects. A matrix is a table or array of numbers or functions arranged in rows and columns. The elements or entries of a matrix $\mathbf{A}$ are often denoted as $a_{ij}$. A matrix $\mathbf{A}$ has $m$ rows and $n$ columns,

$$\mathbf{A} = [a_{ij}] = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1j} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2j} & ... & a_{2n} \\ \vdots & \vdots & & a_{ij} & ... & \vdots \\ a_{m1} & a_{m2} & ... & a_{mj} & ... & a_{mn} \end{pmatrix}, \tag{3.38}$$

we say the size of $\mathbf{A}$ is $m$ by $n$, or $m \times n$. $\mathbf{A}$ is square if $m = n$. For example,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} e^x & \sin x \\ -i\cos x & e^{i\theta} \end{pmatrix}, \tag{3.39}$$

and

$$\mathbf{u} = \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \tag{3.40}$$

where $\mathbf{A}$ is a $2 \times 3$ matrix, $\mathbf{B}$ is a $2 \times 2$ square matrix, and $\mathbf{u}$ is a $3 \times 1$ column matrix or column vector.

The addition (substraction) of two matrices $\mathbf{A}$ and $\mathbf{B}$ is only possible if they have the same size $m \times n$, and their sum, which is also $m \times n$, is obtained by adding corresponding entries

$$\mathbf{C} = \mathbf{A} + \mathbf{B}, \qquad c_{ij} = a_{ij} + b_{ij}, \tag{3.41}$$

where $(i = 1, 2, ..., m; j = 1, 2, ..., n)$. We can multiply a matrix $\mathbf{A}$ by a scalar $\alpha$ and this is equivalent to multiplying each entry by $\alpha$. The product of two matrices is only possible if the number of columns of $\mathbf{A}$ is the same as the number of rows of $\mathbf{B}$. That is to say, if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times r$, then the product $\mathbf{C}$ is $m \times r$,

$$c_{ij} = (AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{3.42}$$

If $\mathbf{A}$ is a square matrix, then we have $\mathbf{A}^r = \overbrace{AA...A}^{r}$. The multiplications of matrices are generally not commutive, i.e., $\mathbf{AB} \neq \mathbf{BA}$. However, the multiplication has associativity

$$\mathbf{A}(\mathbf{uv}) = (\mathbf{Au})\mathbf{v}, \qquad \mathbf{A}(\mathbf{u} + \mathbf{v}) = \mathbf{Au} + \mathbf{Av}. \tag{3.43}$$

The transpose (denoted by $\mathbf{A}^T$) of $\mathbf{A}$ is obtained by switching the position of rows and columns, and thus $\mathbf{A}^T$ will be $n \times m$ if $\mathbf{A}$ is $m \times n$, $(a^T)_{ij} = a_{ji}, (i = 1, 2, ..., m; j = 1, 2, ..., n)$. Generally,

$$(\mathbf{A}^T)^T = \mathbf{A}, \qquad (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \tag{3.44}$$

The differentiation and integral of a matrix are done on each member element. For example, for a $2 \times 2$ matrix

$$\frac{d\mathbf{A}}{dt} \equiv \dot{\mathbf{A}} = \begin{pmatrix} \frac{da_{11}}{dt} & \frac{da_{12}}{dt} \\ \frac{da_{21}}{dt} & \frac{da_{22}}{dt} \end{pmatrix}, \tag{3.45}$$

and

$$\int \mathbf{A}dt = \begin{pmatrix} \int a_{11}dt & \int a_{12}dt \\ \int a_{21}dt & \int a_{22}dt \end{pmatrix}. \tag{3.46}$$

A diagonal matrix $\mathbf{A}$ is a square matrix whose every entry off the main diagonal is zero ($a_{ij} = 0$ if $i \neq j$). Its diagonal elements or entries may or may not have zeros. For example, the matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.47}$$

is a $3 \times 3$ identity or unitary matrix. In general, we have

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}. \tag{3.48}$$

A zero or null matrix $\mathbf{0}$ is a matrix with all of its elements being zero.

### 3.2.2 Determinant

The determinant of a square matrix $\mathbf{A}$ is a number or scalar obtained by the following recursive formula or the cofactor or Laplace expansion by column or row. For example, expanding by row $k$, we have

$$\det(\mathbf{A}) = |\mathbf{A}| = \sum_{j=1}^{n} (-1)^{k+j} a_{kj} M_{kj}, \tag{3.49}$$

where $M_{ij}$ is the determinant of a minor matrix of $\mathbf{A}$ by deleting row $i$ and column $j$. For a simple $2 \times 2$ matrix, its determinant simply becomes

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{3.50}$$

The determinant has the following properties:

$$|\alpha\mathbf{A}| = \alpha|\mathbf{A}|, \qquad |\mathbf{A}^T| = |\mathbf{A}|, \qquad |\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|, \tag{3.51}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the same size ($n \times n$).

A $n \times n$ square matrix is singular if $|\mathbf{A}| = 0$, and is nonsingular if and only if $|\mathbf{A}| \neq 0$. The trace of a square matrix $\mathrm{tr}(\mathbf{A})$ is defined as the sum of the diagonal elements,

$$\mathrm{tr}(\mathbf{A}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} a_{ii} = a_{11} + a_{22} + ... + a_{nn}. \tag{3.52}$$

The rank of a matrix $\mathbf{A}$ is the number of linearly independent vectors forming the matrix. Generally, the rank of $\mathbf{A}$ is $\mathrm{rank}(\mathbf{A}) \leq \min(\mathrm{m}, \mathrm{n})$. For a $n \times n$ square matrix $\mathbf{A}$, it is nonsingular if $\mathrm{rank}(\mathbf{A}) = n$.

### 3.2.3 Inverse

The inverse matrix $\mathbf{A}^{-1}$ of a square matrix $\mathbf{A}$ is defined as

$$\mathbf{A^{-1}A} = \mathbf{AA^{-1}} = \mathbf{I}. \tag{3.53}$$

It is worth noting that the unit matrix $\mathbf{I}$ has the same size as $\mathbf{A}$ and thus is a $n \times n$ matrix. The inverse of a square matrix exists if and only if $\mathbf{A}$ is nonsingular or $\det(\mathbf{A}) \neq 0$. From the basic definitions, it is straightforward to prove that the inverse of a matrix has the following properties

$$(\mathbf{A^{-1}})^{-1} = \mathbf{A}, \qquad (\mathbf{A^T})^{-1} = (\mathbf{A^{-1}})^{\mathbf{T}}, \tag{3.54}$$

and

$$(\mathbf{AB})^{-1} = \mathbf{B^{-1}A^{-1}}. \tag{3.55}$$

### 3.2.4 Matrix Exponential

Sometimes, we need to calculate $\exp[\mathbf{A}]$, where $\mathbf{A}$ is a square matrix. In this case, we have to deal with matrix exponentials. The exponential of a square matrix $\mathbf{A}$ is defined as

$$e^{\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!}\mathbf{A}^n = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + .... \tag{3.56}$$

where $\mathbf{I}$ is a unity matrix with the same size as $\mathbf{A}$, and $\mathbf{A}^2 = \mathbf{AA}$ and so on. This (rather odd) definition in fact provides a method to calculate the matrix exponential. The matrix exponentials are very useful in solving systems of differential equations.

---

**Example 3.3:** For a simple matrix

$$\mathbf{A} = \begin{pmatrix} t & 0 \\ 0 & t \end{pmatrix},$$

we have

$$e^{\mathbf{A}} = \begin{pmatrix} e^t & 0 \\ 0 & e^t \end{pmatrix}.$$

Similarly, we have

$$\mathbf{A} = \begin{pmatrix} t & t \\ t & t \end{pmatrix}, \qquad e^{\mathbf{A}} = \begin{pmatrix} \frac{1}{2}(1+e^{2t}) & \frac{1}{2}(e^{2t}-1) \\ \frac{1}{2}(e^{2t}-1) & \frac{1}{2}(1+e^{2t}) \end{pmatrix}.$$

---

As you see, it is quite complicated but still straightforward to calculate the matrix exponentials. Fortunately, it can be easily done using a computer. By using the power expansions and the basic definition, we can prove the following useful identities

$$e^{t\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!}(t\mathbf{A})^n = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2}\mathbf{A}^2 + ..., \tag{3.57}$$

$$\ln(\mathbf{IA}) \equiv \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n!}\mathbf{A}^n = \mathbf{A} - \frac{1}{2}\mathbf{A}^2 + \frac{1}{3}\mathbf{A}^3 + ..., \tag{3.58}$$

where $t \in \Re$ is a real number.

### 3.2.5 Solution of linear systems

A linear system of $m$ equations for $n$ unknowns

$$\begin{aligned} a_{11}u_1 + a_{12}u_2 + ... + a_{1n}u_n &= b_1, \\ a_{21}u_1 + a_{22}u_2 + ... + a_{2n}u_n &= b_2, \\ \vdots \qquad\qquad & \vdots \\ a_{m1}u_1 + a_{m2}u_2 + ... + a_{mn}u_n &= b_n, \end{aligned} \tag{3.59}$$

can be written in the compact form as

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & & \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \tag{3.60}$$

or simply

$$\mathbf{Au} = \mathbf{b}. \tag{3.61}$$

In the case of $m = n$, we multiply both sides by $\mathbf{A}^{-1}$ (this is only possible when $m = n$),

$$\mathbf{A^{-1}Au} = \mathbf{A^{-1}b}, \tag{3.62}$$

we obtain the solution of this equation as

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{b}. \tag{3.63}$$

A special case of the above equation is when $\mathbf{b} = \lambda\mathbf{u}$, and this becomes an eigenvalue problem. An eigenvalue $\lambda$ and corresponding eigenvector $\mathbf{v}$ of a square matrix $\mathbf{A}$ satisfy

$$\mathbf{Av} = \lambda\mathbf{v}, \tag{3.64}$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}. \tag{3.65}$$

Any nontrivial solution requires

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} - \lambda & ... & a_{2n} \\ \vdots & \vdots & & \\ a_{n1} & a_{n2} & ... & a_{nn} - \lambda \end{vmatrix} = 0, \tag{3.66}$$

which is equivalent to

$$\lambda^n + \alpha_{n-1}\lambda^{n-1} + ... + \alpha_0 = (\lambda - \lambda_1)(\lambda - \lambda_2)...(\lambda - \lambda_n) = 0. \tag{3.67}$$

In general, the characteristic equation has $n$ solutions. Eigenvalues have interesting connections with the matrix,

$$\mathrm{tr}(\mathbf{A}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a_{ii}} = \lambda_1 + \lambda_2 + ... + \lambda_\mathrm{n}. \tag{3.68}$$

For a symmetric square matrix, the two eigenvectors for two distinct eigenvalues $\lambda_i$ and $\lambda_j$ are orthogonal $\mathbf{v^T v} = 0$.

---

**Example 3.4:** For a simple $2 \times 2$ matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 5 \\ 2 & 4 \end{pmatrix},$$

its eigenvalues can be determined by

$$\begin{vmatrix} 1 - \lambda & 5 \\ 2 & 4 - \lambda \end{vmatrix} = 0,$$

or

$$(1 - \lambda)(4 - \lambda) - 2 \times 5 = 0,$$

which is equivalent to

$$(\lambda + 1)(\lambda - 6) = 0.$$

Thus, the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 6$. The trace of $\mathbf{A}$ is $\mathrm{tr}(\mathbf{A}) = A_{11} + A_{22} = 1 + 4 = 5 = \lambda_1 + \lambda_2$. In order to obtain the eigenvector for each eigenvalue, we assume

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

For the eigenvalue $\lambda_1 = -1$, we plug this into

$$|\mathbf{A} - \lambda\mathbf{I}|\mathbf{v} = \mathbf{0},$$

and we have

$$\begin{vmatrix} 1-(-1) & 5 \\ 2 & 4-(-1) \end{vmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0, \qquad \text{or} \qquad \begin{vmatrix} 2 & 5 \\ 2 & 5 \end{vmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0,$$

which is equivalent to $2v_1 + 5v_2 = 0$, or $v_1 = -\frac{5}{2}v_2$. This equation has infinite solutions, each corresponds to the vector parallel to the unit eigenvector. As the eigenvector should be normalised so that its modulus is unity, this additional condition requires

$$v_1^2 + v_2^2 = 1, \qquad \text{or} \qquad (\frac{-5v_2}{2})^2 + v_2^2 = 1.$$

We have $v_1 = -5/\sqrt{29}$, $v_2 = 2/\sqrt{29}$. Thus, we have the first set of eigenvalue and eigenvector

$$\lambda_1 = -1, \qquad \mathbf{v}_1 = \begin{pmatrix} -\frac{5}{\sqrt{29}} \\ \frac{2}{\sqrt{29}} \end{pmatrix}. \tag{3.69}$$

Similarly, the second eigenvalue $\lambda_2 = 6$ gives

$$\begin{vmatrix} 1-6 & 5 \\ 2 & 4-6 \end{vmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0.$$

Using the normalisation condition $v_1^2 + v_2^2 = 1$, the above equation has the following solution $\lambda_2 = 6, \qquad \mathbf{v}_2^T = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$.

---

### 3.2.6 Gauss-Seidel Iteration

For a linear system $\mathbf{Au} = \mathbf{b}$, the solution $\mathbf{u} = \mathbf{A^{-1}b}$ generally involves the inversion of a large matrix. The direct inversion becomes impractical if the matrix is very large (say, if $n > 10000$). Many efficient algorithms have been developed for solving such systems. Gauss elimination and Gauss-Seidel iteration are just two examples.

Gauss-Seidel iteration method provides an efficient way to solve the linear matrix equation $\mathbf{Au} = \mathbf{b}$ by splitting $\mathbf{A}$ into

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}, \tag{3.70}$$

where $\mathbf{L}, \mathbf{D}, \mathbf{U}$ are the lower triangle, diagonal and upper triangle matrices of the $\mathbf{A}$, respectively. The $n$ step iteration is updated by

$$(\mathbf{L} + \mathbf{D})\mathbf{u}^{(n)} = \mathbf{b} - \mathbf{U}\mathbf{u}^{(n-1)}. \tag{3.71}$$

This procedure starts from an initial vector $\mathbf{u}^{(0)}$ (usually, $\mathbf{u}^{(0)} = \mathbf{0}$) stops if a prescribed criterion is reached. There are other iteration methods such as relaxation method and Newton-Raphson method, and readers can go into this in greater detail in a more advanced book on linear algebra.

## 3.3 Tensors

Many physical quantities such as stress and strain are tensors. Vectors are essentially first-order tensors. Tensors are the extension of vectors, and they can have any number of dimensions and any orders, though most commonly used tensors are second order tensors.

### 3.3.1 Notations

In tensor analysis, the summation convention and notations for subscripts are widely used. Any lowercase subscript that appears exactly twice in any term of an expression means that sum is over all possible values of the subscript. This convention is also called Einstein's summation or the index form. For example, in the three-dimensional case, we have

$$\alpha_i x_i \equiv \sum_{i=1}^{3} a_i x_i = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3. \tag{3.72}$$

$$A_{ij} B_{jk} \equiv \sum_{j=1}^{3} A_{ij} B_{jk} = A_{i1} B_{1k} + A_{i2} B_{2k} + A_{i3} B_{3k}. \tag{3.73}$$

$$\nabla \cdot \mathbf{u} \equiv \frac{\partial u_i}{\partial x_i} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}. \tag{3.74}$$

The Kronecker delta $\delta_{ij}$ which is a unity tensor (like the unity matrix $\mathbf{I}$ in matrix analysis), is defined as

$$\delta_{ij} = \begin{cases} 1 & (\text{if } \ i = j), \\ 0 & (\text{if } \ i \neq j). \end{cases} \tag{3.75}$$

Similar to $\delta_{ij}$, the three subscripts Levi-Civita symbol (not a tensor) is defined as

$$\epsilon_{ijk} = \begin{cases} +1 & (\text{if } \ \mathrm{i,j,k} \ \text{ is an even permutation of } 1,2,3), \\ -1 & (\text{if } \ \mathrm{i,j,k} \ \text{ is an odd permutation of } 1,2,3), \\ 0 & (\text{otherwise}). \end{cases} \tag{3.76}$$

Both $\delta_{ij}$ and $\epsilon_{ijk}$ are related by

$$\epsilon_{ijk} \epsilon_{kpq} = \delta_{ip} \delta_{jq} - \delta_{iq} \delta_{jp}. \tag{3.77}$$

Using the summation conventions, the matrix equation $\mathbf{Ax} = \mathbf{b}$ can alternatively be written as

$$A_{ij} x_j = b_i, \qquad (i = 1, 2, ..., n). \tag{3.78}$$

### 3.3.2 Tensors

When changing the basis from the standard Cartesian $\mathbf{e_1} = \mathbf{i}$, $\mathbf{e_2} = \mathbf{j}$, $\mathbf{e_3} = \mathbf{k}$ to a new basis $\mathbf{e'_1}, \mathbf{e'_2}, \mathbf{e'_3}$, a position vector $\mathbf{x} = (x_1, x_2, x_3)$ in the old bases is related to the new vector $\mathbf{x}' = (x'_1, x'_2, x'_3)$ in the new bases by a coefficient matrix $S_{ij}$. $S_{ij}$ can be the rotation, translation, enlargement or any of their combinations. For a rotation with an angle of $\theta$ around a fixed axis, $S_{ij}$ becomes

$$S_{ij} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.79}$$

The orthogonality of $S_{ij}$ requires that $\mathbf{S}\mathbf{S}^{\mathbf{T}} = \mathbf{S}^{\mathbf{T}}\mathbf{S} = \mathbf{I}$ or

$$S_{ij}S_{jk} = \delta_{ik}, \qquad S_{ki}S_{kj} = \delta_{ij}. \tag{3.80}$$

If the components $u_i$ of any variable $\mathbf{u}$ are transformed to the components $u'_i$ in a new basis in the same manner as

$$u'_i = S_{ij}u_j, \qquad u_i = S_{ji}u'_j, \tag{3.81}$$

then $u_i$ are said to form a first-order Cartesian tensor (or vector in this case). If components of a variable $\sigma_{ij}$ are transformed as

$$\sigma'_{ij} = S_{ip}S_{jq}\sigma_{pq}, \qquad \sigma_{ij} = S_{pi}S_{qj}\sigma'_{pq}, \tag{3.82}$$

we say these components form a second-order tensor such as stresses and strains.

In a similar fashion, higher-order tensors can be defined, and for each order increase, then there is one $S_{ij}$ extra in the product for transforming, but no subscripts are allowed to appear more than twice.

$$\tau'_{ij\ldots k} = S_{ip}S_{jq}\ldots S_{kr}\tau_{pq\ldots r}, \qquad \tau_{ij\ldots k} = S_{pi}S_{qj}\ldots S_{rk}\tau'_{pq\ldots r}. \tag{3.83}$$

In a similar way to multi-dimensional matrices, two tensors can be added or subtracted component-by-component if and only if they are the tensors of the same order. For second-order tensors, a tensor $\tau_{ij}$ is said to be symmetric if $\tau_{ij} = \tau_{ji}$, and antisymmetric if $\tau_{ij} = -\tau_{ji}$. An interesting property of a tensor $\tau_{ij}$ is that it can always be written as a sum of a symmetric tensor and an antisymmetric tensor

$$\tau_{ij} = \frac{1}{2}(\tau_{ij} + \tau_{ji})[\text{sym.}] + \frac{1}{2}(\tau_{ij} - \tau_{ji})[\text{antisym.}]. \tag{3.84}$$

All the formulae in vector analysis can be rewritten in the tensor forms using the summation convention, known as index forms. For example, we can write

$$\nabla \times (\nabla \times \mathbf{u})_i = \epsilon_{ijk}\epsilon_{kpq}\frac{\partial u_q}{\partial x_j x_p}, \tag{3.85}$$

which is convenient for proving theorems.

---

**Example 3.5:** The dot product of two vectors can be written as

$$\mathbf{u} \cdot \mathbf{v} = u_i v_i = \delta_{ij} u_i v_j,$$

while the Laplace operator is equivalent to

$$\nabla^2 \psi = \frac{\partial^2 \psi}{\partial x_i \partial x_i} = \delta_{ij} \frac{\partial \psi}{\partial x_i \partial x_j}.$$

Similarly, the divergence theorem can be rewritten as the following form

$$\int_V \frac{\partial u_i}{\partial x_i} dV = \oint_S u_i n_i dS.$$

---

The tensor forms are sometimes useful in proving complex relationships among vectors and tensors. They also become handy for the implementation of numerical algorithms.

# References

Arfken G., *Mathematical Methods for Physicists*, Academic Press, (1985).

Courant R. and Hilbert, D., *Methods of Mathematical Physics*, 2 volumes, Wiley-Interscience, New York, (1962).

Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, Cambridge University Press, (2006).

# Chapter 4

# ODEs and Integral Transforms

Most mathematical models in applied mathematics, physics and earth sciences are formulated in terms of differential equations. If the variables or quantities (such as velocity, temperature, pressure) change with other independent variables such as spatial coordinates and time, their relationship can in general be written as a differential equation or even a set of differential equations.

## 4.1 Ordinary Differential Equations

An ordinary differential equation (ODE) is a relationship between a function $y(x)$ of an independent variable $x$ and its derivatives $y'$, $y''$, ..., $y^{(n)}$. It can be written in a generic form

$$\Psi(x, y, y', y'', ..., y^{(n)}) = 0, \tag{4.1}$$

where $\Psi$ is a function of $x, y, ..., y^{(n)}$. The solution of the equation is a function $y = f(x)$, satisfying the equation for all $x$ in a given domain $\Omega$. The order of the differential equation is equal to the order $n$ of the highest derivative in the equation. Thus, the Riccati equation :

$$y' + a(x)y^2 + b(x)y = c(x), \tag{4.2}$$

is a first-order ODE, and the following equation of Euler-type

$$x^2y'' + a_1xy' + a_0y = 0, \tag{4.3}$$

is a second order. The degree of an equation is defined as the power to which the highest derivative occurs. Therefore, both the Riccati equation and the Euler equation are of the first degree.

An equation is called linear if it can be arranged into the form

$$a_n(x)y^{(n)} + ... + a_1(x)y' + a_0(x)y = \phi(x), \tag{4.4}$$

where all the coefficients depend on $x$ only, not on $y$ or any of its derivatives. If any of the coefficients is a function of $y$ or any of its derivatives, then the equation is nonlinear. If the right-hand side is zero or $\phi(x) \equiv 0$, the equation is homogeneous. It is called nonhomogeneous if $\phi(x) \neq 0$.

The solution of an ordinary differential equation is not always easy, and it is usually very complicated for nonlinear equations. Even for linear equations, the solutions can be found in a straightforward way for only a few simple cases. The solution of a differential equation generally falls into three types: closed form, series form and integral form. A closed form solution is the type of solution that can be expressed in terms of elementary functions and some arbitrary constants. Series solutions are the ones that can be expressed in terms of a series when a closed-form is not possible for certain type of equations. The integral form of solutions or quadrature is sometimes the only form of solution that is possible. If all these forms are not possible, the alternatives are to use approximate and numerical solutions.

### 4.1.1 First-Order ODEs

#### Linear ODEs

A first-order linear differential equation can generally be written as

$$y' + a(x)y = b(x), \tag{4.5}$$

where $a(x)$ and $b(x)$ are the known functions of $x$. Multiplying both sides of the equation by $\exp[\int a(x)dx]$, called the integrating factor, we have

$$y'e^{\int a(x)dx} + a(x)ye^{\int a(x)dx} = b(x)e^{\int a(x)dx}, \tag{4.6}$$

which can be written as

$$[ye^{\int a(x)dx}]' = b(x)e^{\int a(x)dx}. \tag{4.7}$$

By simple integration, we have

$$ye^{\int a(x)dx} = \int b(x)e^{\int a(x)dx}dx + C. \tag{4.8}$$

So its solution becomes

$$y(x) = e^{-\int a(x)dx} \int b(x)e^{\int a(x)dx}dx + Ce^{-\int a(x)dx}, \tag{4.9}$$

where $C$ is an integration constant. For example, from $y'(x) - y(x) = e^{-x}$, we have $a(x) = -1$ and $b = e^{-x}$, so the solution is

$$y(x) = e^{-\int(-1)dx} \int e^{-x} e^{\int(-1)dx} + Ce^{-\int(-1)dx}$$

$$= e^{x} \int e^{-2x} dx + Ce^{x} = -\frac{1}{2}e^{-x} + Ce^{x}. \tag{4.10}$$

**Nonlinear ODEs**

For some nonlinear first order ordinary differential equations, sometimes a transform or change of variables can convert it into the standard first order linear equation (4.5). This is better demonstrated by an example.

---

**Example 4.1:** The Bernoulli's equation can be written in the generic form

$$y' + p(x)y = q(x)y^n, \qquad n \neq 1. \tag{4.11}$$

In the case of $n = 1$, it reduces to a standard first order linear ordinary differential equation. By dividing both sides by $y^n$ and using the change of variables

$$u(x) = \frac{1}{y^{n-1}}, \qquad u' = \frac{(1-n)y'}{y^n}, \tag{4.12}$$

we have

$$u' + (1-n)p(x)u = (1-n)q(x), \tag{4.13}$$

which is a standard first order linear differential equation whose general solution is given earlier in (4.9).

In the simpler case when $p(x) = 2x$, $q(x) = -1$ and $n = 2$, we have

$$u' - 2xu = 1, \qquad u(x) = \frac{1}{y(x)}.$$

For the initial condition $y(0) = 1$, we have $u(0) = 1$. Using solution (4.9), we have

$$u(x) = \frac{\sqrt{\pi}}{2} e^{x^2} \mathrm{erf}(x) + Ae^{x^2},$$

where $A$ is the integration constant. The requirement of $u(0) = 1$ leads to $A = 1$. Thus, the solution for $y(x)$ becomes

$$y(x) = \frac{2e^{-x^2}}{(\sqrt{\pi}\mathrm{erf}(x) + 2)}.$$

We will compare this solution with Euler's scheme in Chapter 10.

---

### 4.1.2 Higher-Order ODEs

Higher order ODEs are more complicated to solve even for the linear equations. For the special case of higher-order ODEs where all the coefficients $a_n, ..., a_1, a_0$ are constants,

$$a_n y^{(n)} + ... + a_1 y' + a_0 y = f(x), \tag{4.14}$$

its general solution $y(x)$ consists of two parts: the complementary function $y_c(x)$ and the particular integral or particular solution $y_p^*(x)$. We have

$$y(x) = y_c(x) + y_p^*(x). \tag{4.15}$$

The complementary function which is the solution of the linear homogeneous equation with constant coefficients can be written in a generic form

$$a_n y_c^{(n)} + a_{n-1} y_c^{(n-1)} + ... + a_1 y_c' + a_0 = 0. \tag{4.16}$$

Assuming $y = Ae^{\lambda x}$ where $A$ is a constant, we get the characteristic equation as a polynomial

$$a_n \lambda^n + a_{n-1} \lambda^{(n-1)} + ... + a_1 \lambda + a_0 = 0, \tag{4.17}$$

which has $n$ roots in the general case. Then, the solution can be expressed as the summation of various terms $y_c(x) = \sum_{k=1}^{n} c_k e^{\lambda_k x}$ if the polynomial has $n$ distinct zeros $\lambda_1, ...\lambda_n$. For complex roots, and complex roots always occur in pairs $\lambda = r \pm i\omega$, the corresponding linearly independent terms can then be replaced by $e^{rx}[A\cos(\omega x) + B\sin(\omega x)]$.

The particular solution $y_p^*(x)$ is any $y(x)$ that satisfies the original inhomogeneous equation (4.14). Depending on the form of the function $f(x)$, the particular solutions can take various forms. For most of the combinations of basic functions such as $\sin x, \cos x$, $e^{kx}$, and $x^n$, the method of the undetermined coefficients is widely used. For $f(x) = \sin(\alpha x)$ or $\cos(\alpha x)$, then we can try $y_p^* = A\sin\alpha x + B\sin\alpha x$. We then substitute it into the original equation (4.14) so that the coefficients $A$ and $B$ can be determined. For a polynomial $f(x) = x^n (n = 0, 1, 2, ...., N)$, we then try $y_p^* = A + Bx + ... + Qx^n$ (polynomial). For $f(x) = e^{kx}x^n$, $y_p^* = (A + Bx + ...Qx^n)e^{kx}$. Similarly, $f(x) = e^{kx}\sin\alpha x$ or $f(x) = e^{kx}\cos\alpha x$, we can use $y_p^* = e^{kx}(A\sin\alpha x + B\cos\alpha x)$. More general cases and their particular solutions can be found in various textbooks.

A very useful technique is to use the method of differential operator $D$. A differential operator $D$ is defined as

$$D \equiv \frac{d}{dx}. \tag{4.18}$$

Since we know that $De^{\lambda x} = \lambda e^{\lambda x}$ and $D^n e^{\lambda x} = \lambda^n e^{\lambda x}$, so they are equivalent to $D \mapsto \lambda$, and $D^n \mapsto \lambda^n$. Thus, any polynomial $P(D)$ will map to $P(\lambda)$. On the other hand, the integral operator $D^{-1} = \int dx$ is just the inverse of the differentiation. The beauty of the differential operator form is that one can factorise it the same as for the polynomial, then solve each factor separately. The differential operator is very useful in finding out both the complementary functions and particular integral.

---

**Example 4.2:** To find the particular integral for the equation

$$y'''' + 2y = 17e^{2x}, \tag{4.19}$$

we get

$$(D^5 + 2)y_p^* = 17e^{2x}, \tag{4.20}$$

or

$$y_p^* = \frac{17}{D^5 + 2} e^{2x}. \tag{4.21}$$

Since $D^5 \mapsto \lambda^5 = 2^5$, we have

$$y_p^* = \frac{17e^{2x}}{2^5 + 2} = \frac{e^{2x}}{2}. \tag{4.22}$$

---

This method also works for $\sin x, \cos x, \sinh x$ and others, and this is because they are related to $e^{\lambda x}$ via $\sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta})$ and $\cosh x = (e^x + e^{-x})/2$.

Higher order differential equations can conveniently be written as a system of differential equations. In fact, an $n$th-order linear equation can always be written as a linear system of $n$ first-order differential equations. A linear system of ODEs is more suitable for mathematical analysis and numerical integration.

### 4.1.3 Linear System

For a linear $n$ order equation (4.16), it can always be written as a linear system

$$\frac{dy}{dx} = y_1, \quad \frac{dy_1}{dx} = y_2, \ ..., \ \frac{dy_{n-1}}{dx} = y_{n-1},$$

$$a_n(x)y'_{n-1} = -a_{n-1}(x)y_{n-1} + ... + a_1(x)y_1 + a_0(x)y + \phi(x), \tag{4.23}$$

which is a system for $u = [y \ y_1 \ y_2 \ ... \ y_{n-1}]^T$. If the independent variable $x$ does not appear explicitly in $y_i$, then the system is said to be autonomous with important properties. For simplicity and in keeping

with the convention, we use $t = x$ and $\dot{u} = du/dt$ in our following discussion. A general linear system of $n$-th order can be written as

$$\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \vdots \\ \dot{u}_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \tag{4.24}$$

or

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}. \tag{4.25}$$

If $\mathbf{u} = \mathbf{v}\exp(\lambda t)$, then this becomes an eigenvalue problem,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}, \tag{4.26}$$

which will have non-null solution only if

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0. \tag{4.27}$$

### 4.1.4 Sturm-Liouville Equation

One of the commonly used second-order ordinary differential equations is the Sturm-Liouville equation in the interval $x \in [a, b]$

$$\frac{d}{dx}[p(x)\frac{dy}{dx}] + q(x)y + \lambda r(x)y = 0, \tag{4.28}$$

with the boundary conditions

$$y(a) + \alpha y'(a) = 0, \qquad y(b) + \beta y'(b) = 0, \tag{4.29}$$

where the known function $p(x)$ is differentiable, and the known functions $q(x), r(x)$ are continuous. The parameter $\lambda$ to be determined can only take certain values $\lambda_n$, called the eigenvalues, if the problem has solutions. For the obvious reason, this problem is called Sturm-Liouville eigenvalue problem.

Sometimes, it is possible to transform a nonlinear equation into a standard Sturm-Liouville equation, and this is better demonstrated by an example.

---

**Example 4.3:** The Riccati equation can be written in the generic form

$$y' = p(x) + q(x)y + r(x)y^2, \qquad r(x) \neq 0.$$

If $r(x) = 0$, then it reduces to a first order linear ODE. By using the transform

$$y(x) = -\frac{u'(x)}{r(x)u(x)},$$

or

$$u(x) = e^{-\int r(x)y(x)dx},$$

we have

$$u'' - P(x)u' + Q(x)u = 0,$$

where $P(x) = -r'(x)/r(x) + q(x)$ and $Q(x) = r(x)p(x)$.

For each eigenvalue $\lambda_n$, there is a corresponding solution $\psi_{\lambda_n}$, called eigenfunctions. The Sturm-Liouville theory states that for two different eigenvalues $\lambda_m \neq \lambda_n$, their eigenfunctions are orthogonal. That is

$$\int_a^b \psi_{\lambda_m}(x)\psi_{\lambda_n}(x)r(x)dx = 0, \text{or } \int_a^b \psi_{\lambda_m}(x)\psi_{\lambda_n}(x)r(x)dx = \delta_{mn}.$$

It is possible to arrange the eigenvalues in an increasing order

$$\lambda_1 < \lambda_2 < ... < \lambda_n < ... \to \infty.$$

Now let us study a real-world problem using differential equations. Many fluid flow problems are relative to flow through a pipe, including the water flow through a pipe, oil in an oil pipeline, and magma flow in a dyke and others. Let us look at the Poiseuille flow in a cylindrical pipe.

**Example 4.4:** The laminar flow of a viscous fluid through a pipe with a radius $r = a$ under a pressure gradient (see Fig. 4.1) $\nabla p = \Delta P/L = (P_o - P_i)/L$ where $P_i$ and $P_o$ $(< P_i)$ are the pressures at inlet and outlet, respectively. $L$ is the length of the pipe. The drag force is balanced by pressure change, and this leads to the following second-order ordinary differential equation

$$\frac{\Delta P}{L} = \eta \frac{1}{r}\frac{d}{dr}[r\frac{dv(r)}{dr}],$$

where $\eta$ is the viscosity of the fluid. This equation implies that the flow velocity $v$ is not uniform, it varies with $r$. Integrating the above equation twice, we have

$$v(r) = \frac{\Delta P}{4\eta L}r^2 + A\ln r + B,$$

where $A$ and $B$ are integrating constants. The velocity must be finite at $r = 0$, which means that $A = 0$. The no-slip boundary $v = 0$ at $r = a$ requires that

$$\frac{\Delta P}{4\eta L}a^2 + B = 0.$$

Thus, the velocity profile is

$$v(r) = -\frac{\Delta P}{4\eta L}(a^2 - r^2).$$

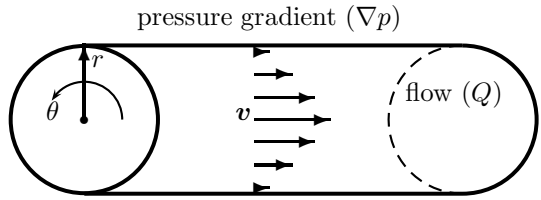



Figure 4.1: Flow through a pipe under pressure gradient.

Now the total flow rate $Q$ down the pipe is given by integrating the flow over the whole cross section. We have

$$Q = \int_0^a 2\pi r v(r) dr = -\frac{\pi \Delta P}{2\eta L} \int_0^a (a^2 r - r^3) dr = -\frac{\pi \Delta P}{8\eta L} a^4. \quad (4.30)$$

Here the negative sign means the flow down the pressure gradient. We can see that the flow rate is proportional to the pressure gradient, inversely proportional to the viscosity. Double the radius of the pipe, and the flow rate will increase to 16 times.

## 4.2 Integral Transforms

Mathematical transform is a method of changing one kind of function and equation into another, often simpler or more easily solveable kind. Integral transform is a mathematical operator that produces a new function $F(s)$ by integrating the product of an existing function $f(t)$ and a kernel function $K(t, s)$ between suitable limits

$$F(s) = \int K(t, s) f(t) dt. \quad (4.31)$$

In the Laplace transform, the kernel is simply $\exp(-st)$ and integration limit is from 0 to $\infty$. In the Fourier transform, the kernel is $\exp(\pm ist)$ with a normalised factor.

Fourier transform maps a function in the time domain such as a signal into another function in the frequency domain, which is commonly used in signal processing. Laplace transform is a very powerful tool in solving differential equations. Here, we will focus on the three major transforms: Fourier, Laplace and Wavelet commonly encountered in engineering and computational sciences.

### 4.2.1 Fourier Series

For a function $f(t)$ on an interval $t \in [-T, T]$ where $T > 0$ is a finite constant or half period, the Fourier series is defined as

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n[\cos(\frac{n\pi t}{T}) + b_n \sin(\frac{n\pi t}{T})], \tag{4.32}$$

where

$$a_0 = \frac{1}{T}\int_{-T}^{T} f(t)dt, \qquad a_n = \frac{1}{T}\int_{-T}^{T} f(t)\cos(\frac{n\pi t}{T})dt, \tag{4.33}$$

and

$$b_n = \frac{1}{T}\int_{-T}^{T} f(T)\sin(\frac{n\pi t}{T})dt, \quad (n = 1, 2, ...). \tag{4.34}$$

Here $a_n$ and $b_n$ are the Fourier coefficients of $f(t)$ on $[-T, T]$. The function $f(t)$ can be continuous or piecewise continuous with a finite number of jump discontinuity. For a jump discontinuity at $t = t_0$, if $f'(t_0-)$ and $f'(t_0+)$ both exist with $f(t_0-) \neq f(t_0+)$, then the Fourier series converge to $[f(t_0-) + f(t_0+)]/2$. Fourier series in general tends to converge slowly. In order for a function $f(x)$ to be expanded, it must satisfy the Dirichlet conditions: $f(x)$ must be periodic with at most a finite number of discontinuities, and/or a finite number of minima or maxima within one period. In addition, the integral of $|f(x)|$ must converge. For example, these conditions suggest that $\ln(x)$ cannot be expanded into a Fourier series in the interval $[0, 1]$ as $\int_0^1 |\ln x|dx$ diverges.

The $n$th term of the Fourier series, $a_n \cos(n\pi t/T) + b_n \sin(n\pi t/T)$, is called the $n$th harmonic. The energy of the $n$th harmonic is defined by $A_n^2 = a_n^2 + b_n^2$, and the sequence of $A_n^2$ forms the energy or power spectrum of the Fourier series.

From the coefficient $a_n$ and $b_n$, one can easily see that $b_n = 0$ for an even function $f(-t) = f(t)$. Similarly, $a_0 = a_n = 0$ for an odd function $f(-t) = -f(t)$. In both cases, only one side $[0, T]$ of the integration is used due to the symmetry. Thus, for even function $f(t)$, we have the Fourier cosine series on [0,T]

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(\frac{n\pi t}{T}). \tag{4.35}$$

For odd function $f(t)$, we have the Fourier sine series

$$f(t) = \sum_{n=1}^{\infty} \sin(\frac{n\pi t}{T}). \tag{4.36}$$

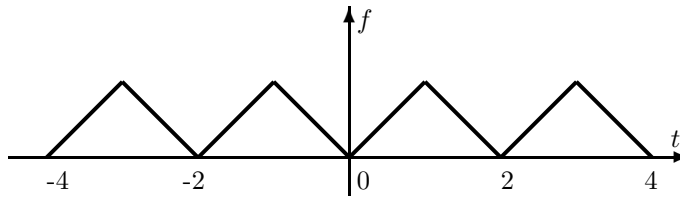



Figure 4.2: Triangular wave with a period of 2.

---

**Example 4.5:** The triangular wave is defined by $f(t) = |t|$ for $t \in [-1, 1]$ with a period of 2 or $f(t + 2) = f(t)$ shown in Fig. 4.2. Using the coefficients of the Fourier series, we have

$$a_0 = \int_{-1}^{1} |t| dt = \int_{-1}^{0} (-t) dt + \int_{0}^{1} t dt = 1.$$

Since both $|t|$ and $\cos(n\pi t)$ are even functions, we have for any $n \geq 1$,

$$a_n = \int_{-1}^{1} |t| \cos(n\pi t) dt = 2 \int_{0}^{1} t \cos(n\pi t) dt$$

$$= 2\frac{t}{n\pi} \sin(n\pi t)\Big|_0^1 - \frac{2}{n\pi} \int_0^1 \sin(n\pi t) dt = \frac{2}{n^2\pi^2}[\cos(n\pi) - 1].$$

Because $|t| \sin(n\pi t)$ is an odd function, we have

$$b_n = \int_{-1}^{1} |t| \sin(n\pi t) dt = 0.$$

Hence, the Fourier series for the triangular wave can be written as

$$f(t) = \frac{1}{2} + 2 \sum_{n=1}^{\infty} \frac{\cos(n\pi) - 1}{n^2\pi^2} \cos(n\pi t) = \frac{1}{2} + \frac{4}{\pi^2} \sum_{n=1,3,5,\ldots}^{\infty} \frac{(-1)^n}{n^2} \cos(n\pi t).$$

The first few terms, $f_n(t) = 1/2 + 4/\pi^2 \cos(\pi t)$, are shown in Fig. 4.3 where we can see that only a few terms are needed to produce a very good approximation.

---

Here we can see that the triangular wave with derivative discontinuity can be approximated well by two or three terms. This makes it

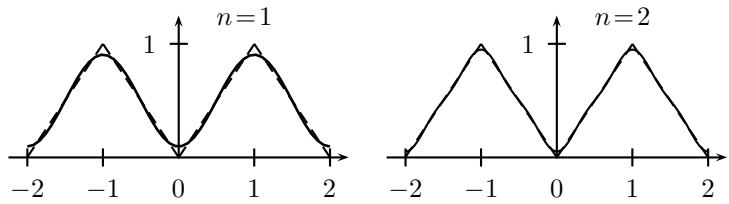



Figure 4.3: Fourier series for the triangular wave $f(t) = |t|, t \in [-1, 1]$ : (a) first two terms ($n$=1); (b) first three terms ($n = 2$).

easy for any mathematical analysis. Fourier series are widely applied in signal processing.

Now let us look at a real-world example by studying the Milankovitch cycles in climate changes. Milankovitch theory explains paleoclimate fluctuations and occurrence of the Ice Ages very well. The Milankovitch cycles, named after the Serbian scientist M. Milankovitch who studied the effect of the Earth's orbital motion on the climate in a pioneer paper published in 1941, refer to the collective effect on the climate change due to the changes in the Earth's orbital movements (see Fig. 4.4).

There are three major components in the orbital changes: precession of the perihelion, obliquity (or wobble of the Earth's axis of rotation), and eccentricity (or shape of the Earth's orbit). Because of the interaction of the Sun, the Moon, and other planets (mainly Jupiter and Saturn) with the Earth, each of the three components usually has multiple harmonic components. Here we will outline the theory.

The precession of the perihelion has a number of harmonic components, ranging from 19 to 23.7 thousand years (kyrs), though the weighted averaged is about 21 kyrs. The tilting of the Earth's axis of rotation varies from about 21.5° to 24.5° with periods from 29 to 53.6 kyrs. The averaged period is about 41.6 kyrs. The increase of obliquity will lead to the increase of the amplitude of the seasonal cycle in insolation. At the same time, the precession or wobble of this axis (relative to fixed stars) completes a big circle in about 26 kyrs, though it is about 21 kyrs if calculated relative to the perihelion. This wobble is mainly caused by the differential gravitational force due to the fact that the Earth is not a perfect sphere and it has an equatorial bulge.

The change of eccentricity varies from $e = 0.005$ to 0.06 with periods ranging from 94.9 to 412.9 kyrs. Two major components are a long period of 412.9 kyrs and an averaging short period of 110.7 kyrs, and the latter is close to the 100 kyrs cycles of ice ages. All these harmonic

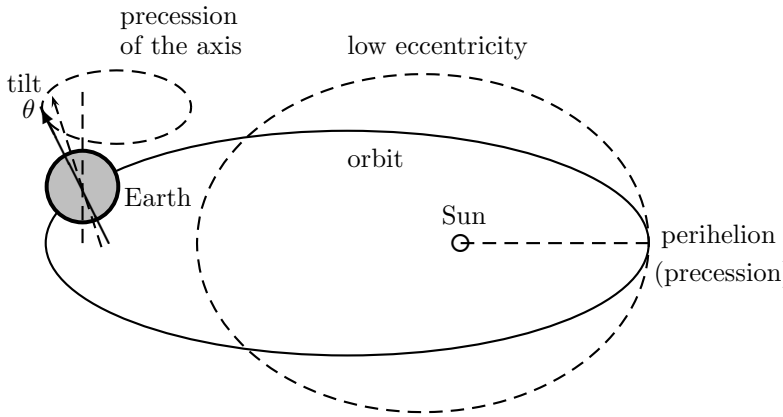



Figure 4.4: Milankovitch cycles of the Earth's orbital elements.

components interact and result in a complicated climate pattern.

---

**Example 4.6:** From the detailed calculation by Berger (1977) based on Milankovitch's theory, we can write the precession as

$$p \approx p_0 + \tilde{p}[0.42\sin(\frac{2\pi t}{19.1}) + 0.28\sin(\frac{2\pi t}{22.4}) + 0.30\sin(\frac{2\pi t}{23.7})],$$

where we have used the approximated averaged periods. $\tilde{p}$ is the averaged amplitude of the precession and $p_0$ is the initial value. In writing this equation, we have implicitly assumed that the phase shift between different harmonic components is negligible, and many components with similar periods have been combined into a single major component.

Similarly, the obliquity can be expressed as

$$\theta \approx \theta_0 + \tilde{\theta}[0.06\sin(\frac{2\pi t}{29}) + 0.80\sin(\frac{2\pi t}{41}) + 0.14\sin(\frac{2\pi t}{53.6})],$$

where $\tilde{\theta} \approx 1.5^\circ$ is the averaged amplitude of tilting and $\theta_0 \approx 23^\circ$ is the mean angle. The current tilting is about $23.44^\circ$.

The variation of the eccentricity is

$$e \approx e_0 + \tilde{e}[0.22\sin(\frac{2\pi t}{95}) + 0.50\sin(\frac{2\pi t}{125}) + 0.28\sin(\frac{2\pi t}{412.9})],$$

where $\tilde{e} \approx 0.0275$ is the averaged amplitude of eccentricity, and $e_0 \approx 0.0325$ is the mean eccentricity. The present eccentricity of the Earth's orbit is about 0.017. Although the variation of $e$ is small, it still results in

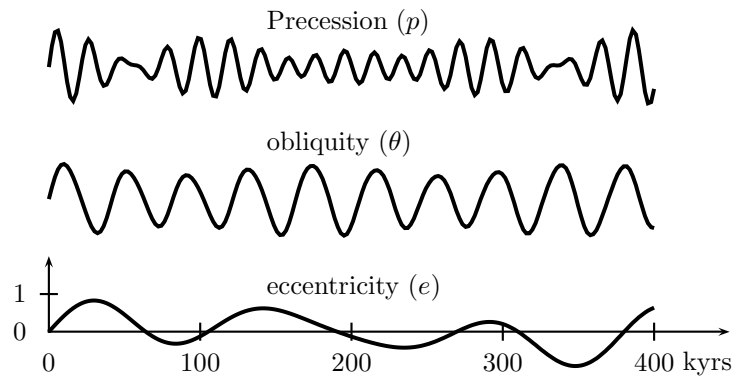



Figure 4.5: Milankovitch cycles: (a) precession of perihelion; (b) obliquity; and (c) eccentricity.

a change of distance of the order of about 5 million kilometers (aphelion minus perihelion), or about 3% of the average distance from the Earth to the Sun, which will result in about 6% change in solar energy reaching the Earth as the energy flux is inversely proportional to the distance.

These variations are shown in Fig. 4.5 and their amplitude spectra are shown in Fig. 4.6.

### 4.2.2 Fourier Integral

For the Fourier coefficients of a function defined on the whole real axis $[-\infty, \infty]$:

$$a(\omega_n) = \int_{-T}^{T} f(t)\cos(\omega_n t)dt, \qquad b(\omega_n) = \int_{-T}^{T} f(t)\sin(\omega_n t)dt, \quad (4.37)$$

where $\omega_n = \frac{n\pi}{T}$ under the limits of $T \to \infty$ and $\omega_n \to 0$, we have $a_0 \to 0$ if $\int_{\infty}^{\infty} |f(t)| < \infty$. In this case, the Fourier series becomes the Fourier integral

$$f(t) = \int_0^{\infty} [a(\omega)\cos(\omega t) + b(\omega)\sin(\omega t)]d\omega, \quad (4.38)$$

where

$$a(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(t)\cos(\omega t)dt, \; b(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(t)\sin(\omega t)dt. \quad (4.39)$$

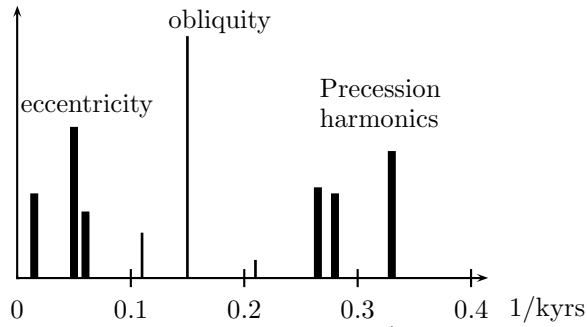


Figure 4.6: Spectra of Milankovitch cycles (relative amplitudes).

Following similar discussions above, even functions lead to Fourier cosine integrals and odd functions lead to Fourier sine integrals.

### 4.2.3 Fourier Transforms

The Fourier transform $\mathcal{F}[f(t)]$ of $f(t)$ is defined as

$$F(\omega) = \mathcal{F}[f(t)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt, \tag{4.40}$$

and the inverse Fourier transform can be written as

$$f(t) = \mathcal{F}^{-1}[F(\omega)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega, \tag{4.41}$$

where $\exp[i\omega t] = \cos(\omega t) + i \sin(\omega t)$. The Fourier transform has the following properties:

$$\mathcal{F}[f(t) + g(t)] = \mathcal{F}[f(t)] + \mathcal{F}[g(t)], \qquad \mathcal{F}[\alpha f(t)] = \alpha \mathcal{F}[f(t)], \tag{4.42}$$

and

$$\mathcal{F}[(-it)^n f(t)] = \frac{d^n F(\omega)}{d\omega^n}, \qquad \mathcal{F}[f^{(n)}(t)] = (i\omega)^n F(\omega), \tag{4.43}$$

if $f(t \to \pm\infty) = f'(t \to \pm\infty) = ... = f^{(n-1)}(t \to \pm\infty) \to 0$. There are some variations of the transforms such as the Fourier sine transform and the Fourier cosine transform. The Fourier transforms of some common functions are listed in the following table 4.1.

The most useful Fourier transform is the discrete form because signal processing is digital in form. The discrete Fourier transform (DFT)

Table 4.1: Fourier Transforms

| f(t) | $F(\omega) = \mathcal{F}[f(t)]$ |
|---|---|
| $f(t-t_0)$ | $F(\omega)e^{-i\omega t_0}$ |
| $f(t)e^{-i\omega_0 t}$ | $F(\omega-\omega_0)$ |
| $\delta(t)$ | $1/\sqrt{2\pi}$ |
| $1$ | $\sqrt{2\pi}\delta(\omega)$ |
| $\text{sign}(t)$ | $\frac{2}{i\omega}$ |
| $e^{-\alpha\|t\|}$ | $\frac{2\alpha}{\alpha^2+\omega^2}$ |
| $e^{-(\alpha t)^2}\ \ (\alpha>0)$ | $\frac{1}{\sqrt{2}\alpha}e^{-\frac{\omega^2}{4\alpha^2}}$ |
| $f(\alpha t)$ | $\frac{1}{\|\alpha\|}F(\frac{\omega}{\alpha})$ |
| $\frac{1}{\alpha^2+t^2}$ | $\sqrt{\frac{\pi}{2}}\frac{e^{-\alpha\|\omega\|}}{\alpha}$ |
| $\cos(\omega_0 t)$ | $\sqrt{\frac{\pi}{2}}[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]$ |
| $\sin(\omega_0)$ | $i\sqrt{\frac{\pi}{2}}[\delta(\omega+\omega_0)-\delta(\omega-\omega_0)]$ |
| $\frac{\sin\alpha x}{x}\ \ (\alpha>0)$ | $\sqrt{\frac{\pi}{2}},\ \ (\|\omega\|<\alpha); 0,\ \ (\|\omega\|>\alpha)$ |

for periodic discrete function or signal $x(n)$ with a period $N$ is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-i\frac{2\pi kn}{N}}, \tag{4.44}$$

and inverse transform or signal reconstruction is

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{i\frac{2\pi kn}{N}}. \tag{4.45}$$

A periodic signal $x(n+N) = x(n)$ has a periodic spectrum $X[k+N] = X[k]$. The discrete Fourier transform consists of $N$ multiplications and $N-1$ additions for each $X[k]$, thus for $N$ values of $k$, the computational complexity is of $O(N^2)$. However, if $N = 2^m$ ($m \in \mathcal{N}$), many of the DFT calculations are not necessary. In fact, by rearranging the formula, one can get the complexity of $O(N\log_2 N)$. This type of algorithm is called Fast Fourier Transform (FFT). There is a vast amount of literature about the signal processing such as FFT, filter design and signal reconstruction.

### 4.2.4 Laplace Transforms

The Laplace transform $\mathcal{L}[f(t)]$ of a function $f(t)$ is defined as

$$F(s) = \mathcal{L}[f(t)] = \int_0^\infty f(t)e^{-st}dt, \tag{4.46}$$

where $s > 0$. The inverse Laplace transform $\mathcal{L}^{-1}[F(s)]$ is $f(t)$ or $f(t) = \mathcal{L}^{-1}[F(s)]$. The Laplace transform of most simple functions can be obtained by direct integration. For simple functions $t$ and $e^{\alpha t}$, we have

$$\mathcal{L}[t] = \int_0^\infty te^{-st}dt = \int_0^\infty \frac{1}{s}e^{-st}dt + \left[-\frac{t}{s}e^{-st}\right]_0^\infty = \frac{1}{s^2}.$$

$$\mathcal{L}[e^{\alpha t}] = \int_0^\infty e^{\alpha t}e^{-st}dt = \left[-\frac{1}{s-\alpha}e^{-(s-\alpha)t}\right]_0^\infty = \frac{1}{s-\alpha}.$$

Conversely, $\mathcal{L}^{-1}[\frac{1}{s^2}] = t, \qquad \mathcal{L}^{-1}[\frac{1}{s-\alpha}] = e^{\alpha t}$.

However, the inverse of Laplace transform is usually more complicated. It often involves the partial fractions of polynomials and use of a combination of Laplace transform rules. From the definition, it is straightforward to prove that the Laplace transform has the following properties:

$$\mathcal{L}[\alpha f(t) + \beta g(t)] = \alpha\mathcal{L}[f(t)] + \beta\mathcal{L}[g(t)], \tag{4.47}$$

$$\mathcal{L}[e^{\alpha t}f(t)] = F(s-\alpha), \quad s > \alpha, \tag{4.48}$$

$$\mathcal{L}[f(t-\alpha)] = e^{-\alpha s}\mathcal{L}[f(t)], \tag{4.49}$$

$$\mathcal{L}[f'(t)] = s\mathcal{L}[f(t)] - f(0), \qquad \mathcal{L}[\int_0^t f(\tau)d\tau] = \frac{1}{s}\mathcal{L}[f], \tag{4.50}$$

The Laplace transform pairs of common functions are listed below in table 4.2.

---

**Example 4.7:** In order to obtain the Laplace transform of $f(t) = \cosh\omega t$, we shall first write

$$f(t) = \cosh\omega t = \frac{1}{2}(e^{\omega t} + e^{-\omega t}).$$

Then, we have

$$\mathcal{L}[f(t)] = F(s) = \int_0^\infty [\frac{1}{2}(e^{\omega t} + e^{-\omega t})]e^{-st}dt$$

$$= \frac{1}{2}[\int_0^\infty e^{-(s-\omega)t}dt + \int_0^\infty e^{-(s+\omega)t}dt] = \frac{1}{2}[\frac{1}{s-\omega} + \frac{1}{s+\omega}] = \frac{s}{s^2-\omega^2}.$$

---

Both Fourier and Laplace transforms follow the convolution theorem. For two functions $f$ and $g$, their convolution $f * g$ is given by

$$f * g = \int_0^t f(t-\alpha)g(\alpha)d\alpha. \tag{4.51}$$

Table 4.2: Laplace Transform

| Function $f(t)$ | Laplace Transform $F(s)$ |
|---|---|
| $1$ | $\frac{1}{s}$ |
| $\delta(t)$ | $1$ |
| $t^n, n > 0$ | $\frac{n!}{s^{n+1}}$ |
| $\cos(\alpha t)$ | $\frac{s}{s^2+\alpha^2}$ |
| $\sin(\alpha t)$ | $\frac{\alpha}{s^2+\alpha^2}$ |
| $e^{\alpha t}$ | $\frac{1}{s-\alpha}$ |
| $t^{1/2}$ | $\frac{1}{2}(\frac{\pi}{s^3})^{1/2}$ |
| $t^{-1/2}$ | $\sqrt{\frac{\pi}{s}}$ |
| $t^n f(t)$ | $(-1)^n \frac{d^n F(s)}{ds^n}$ |
| $\cos(\alpha t + \beta)$ | $\frac{s\cos(\beta)-\alpha\sin(\beta)}{s^2+\alpha^2}$ |
| $\sinh(\alpha t)$ | $\frac{\alpha}{s^2-\alpha^2}$ |
| $\cosh(\alpha t)$ | $\frac{s}{s^2-\alpha^2}$ |
| $\operatorname{erfc}(\frac{\alpha}{2\sqrt{t}})$ | $\frac{1}{s}e^{-\alpha\sqrt{s}}$ |
| $\frac{1}{\sqrt{\pi t}}e^{-\frac{\alpha^2}{4t}}$ | $\frac{1}{\sqrt{s}}e^{-\alpha\sqrt{s}}$ |

and their Laplace transforms follow the convolution theorem

$$\mathcal{L}[f(t) * g(t)] = F(s)G(s), \tag{4.52}$$

$$\mathcal{L}^{-1}[F(s)G(s)] = \int_0^t f(t-\alpha)g(\alpha)d\alpha. \tag{4.53}$$

The Fourier transform has the similar property

$$f(t)*g(t) = \int_{-\infty}^{\infty} f(t)g(t-u)du, \qquad \mathcal{F}[f(t)*g(t)] = F(\omega)G(\omega). \tag{4.54}$$

### 4.2.5 Wavelets

Fourier transform is an ideal tool for studying the stationary time signal whose properties are statistically invariant over time. In Fourier transform, the stationary signal is decomposed into linear combinations of sine and cosine waves

$$\frac{1}{\sqrt{2\pi}}, \quad \frac{1}{\sqrt{\pi}}\cos(nt), \quad \frac{1}{\sqrt{\pi}}\sin(nt), \quad (n = 1, 2, ...). \tag{4.55}$$

For non-stationary signals whose frequencies $f = \omega/2\pi$ vary with time, the Fourier transform does not work well. In addition, in the Fourier

transform there is a tradeoff between frequency resolution and time resolution, that is $\Delta\omega\Delta t \geq \frac{1}{2}$, which is similar to the Heisenberg uncertainty principle for spatial and velocity intervals. The wavelet transform is an alternative approach to the Fourier transform to overcome the resolution problem using the Mother wavelet $\psi$ or prototype for generating the other windows functions, and all the used windows are in the form of either dilated/compressed or shifted. As a result, the wavelet transform is very powerful in dealing with non-stationary signals. In the wavelet transform, a transient signal is decomposed into elementary components of wavelets or wavelet packets. There are three major type of wavelets: Grossmann-Morlet wavelets, Daubechies wavelets and Gabor-Malvar wavelets. We start to define wavelets with a real-valued function $\psi(t)$ $(t \in \mathcal{R})$ as the generator wavelet or mother wavelet. The function $\psi$ is both well localised, decreasing rapidly as $t \to \infty$ and oscillating in a wavy manner. To generator other wavelets, $\psi(\alpha, \beta, t)$ is used by translating in time and change of scales.

Grossmann-Morlet wavelets are of the form

$$\frac{1}{\alpha}\psi(\frac{t-\beta}{\alpha}), \quad \alpha > 0, \qquad a, b \in \mathcal{R}, \tag{4.56}$$

where $\psi$ a generator wavelet. The Daubechies wavelets have the form

$$2^{n/2}\psi(2^n t - m), \qquad m, n \in \mathcal{Z}. \tag{4.57}$$

The Gabor-Malvar wavelets are in the form

$$w(t-m)\cos[\pi(n+\frac{1}{2})(t-m)], \qquad m \in \mathcal{Z}, n \in N. \tag{4.58}$$

The literature on the wavelet processing is vast, and readers can find more details in more specialised books.

# References

Berger A. L., Long term variations of the Earth's orbital elements, *Celestial Mechanics*, **15**, 53-74 (1977).

Carrrier G. F. and Pearson C. E., *Partial Differential Equations: Theory and Technique*, 2nd Edition, Academic Press, (1988).

Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, Cambridge University Press, (2006).

# Chapter 5

# PDEs and Solution Techniques

## 5.1 Partial Differential Equations

Partial differential equations are much more complicated compared with ordinary differential equations. There is no universal solution technique for nonlinear equations, even numerical simulations are usually not straightforward. Thus, we will mainly focus on the linear partial differential equations and equations of special interest to the earth sciences.

A partial differential equation (PDE) is a relationship containing one or more partial derivatives. Similar to the ordinary differential equation, the highest $n$th partial derivative is referred to as the order $n$ of the partial differential equation. The general form of a partial differential equation can be written as

$$\psi(u, x, y, ..., \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}, \frac{\partial^2 u}{\partial x \partial y}, ...) = 0. \tag{5.1}$$

where $u$ is the dependent variable, and $x, y, ...$ are the independent variables.

A simple example of partial differential equations is the linear first-order partial differential equation, which can be written as

$$a(x, y)\frac{\partial u}{\partial x} + b(x, y)\frac{\partial u}{\partial y} = f(x, y). \tag{5.2}$$

for two independent variables and one dependent variable $u$. If the right-hand side is zero or simply $f(x, y) = 0$, then the equation is said

to be homogeneous. The equation is said to be linear if $a, b$ and $f$ are functions of $x, y$ only, not $u$ itself.

For simplicity in notation in the studies of PDEs, compact subscript forms are often used in the literature. They are

$$u_x \equiv \frac{\partial u}{\partial x}, \quad u_y \equiv \frac{\partial u}{\partial y}, \quad u_{xx} \equiv \frac{\partial^2 u}{\partial x^2}, \quad u_{xy} \equiv \frac{\partial^2 u}{\partial x \partial y}, \quad \ldots \tag{5.3}$$

and thus we can write (5.2) as

$$au_x + bu_y = f. \tag{5.4}$$

In the rest of the chapters in this book, we will use this notation whenever no confusion occurs.

### 5.1.1 First-Order PDEs

A first order linear partial differential equation can be written as

$$a(x, y)u_x + b(x, y)u_y = f(x, y), \tag{5.5}$$

which can be solved using the method of characteristics in terms of a parameter $s$

$$\frac{dx}{ds} = a, \quad \frac{dy}{ds} = b, \quad \frac{du}{ds} = f, \tag{5.6}$$

which essentially forms a system of first-order ordinary differential equations. The simplest example of first-order linear partial differential equations is the first-order hyperbolic equation

$$u_t + cu_x = 0, \tag{5.7}$$

where $c$ is a constant. It has a general solution

$$u = \psi(x - ct), \tag{5.8}$$

which is a travelling wave along the $x$-axis with a constant speed $c$. If the initial shape is $u(x, 0) = \psi(x)$, then $u(x, t) = \psi(x - ct)$ at time $t$, therefore the shape of the wave does not change with time though its position is constantly changing.

### 5.1.2 Classification of Second-Order PDEs

A linear second-order partial differential equation can be written in the generic form in terms of two independent variables $x$ and $y$,

$$au_{xx} + bu_{xy} + cu_{yy} + gu_x + hu_y + ku = f, \tag{5.9}$$

where $a, b, c, g, h, k$ and $f$ are functions of $x$ and $y$ only. If $f(x, y, u)$ is also a function of $u$, then we say that this equation is quasi-linear.

If $\Delta = b^2 - 4ac < 0$, the equation is elliptic. One famous example is the Laplace equation $u_{xx} + u_{yy} = 0$.

If $\Delta > 0$, it is hyperbolic. A good example is the wave equation $u_{tt} = c^2 u_{xx}$.

If $\Delta = 0$, it is parabolic. Diffusion and heat conduction are of the parabolic type $u_t = \kappa u_{xx}$.

## 5.2 Classic Mathematical Models

Three types of classic partial differential equations are widely used and they occur in a vast range of applications. In fact, almost all books or studies on partial differential equations will have to deal with these three types of basic partial differential equations.

### 5.2.1 Laplace's and Poisson's Equation

In heat transfer problems, the steady state of heat conduction with a source is governed by the Poisson equation

$$k\nabla^2 u = f(x, y, t), \qquad (x, y) \in \Omega, \tag{5.10}$$

or

$$u_{xx} + u_{yy} = q(x, y, t), \tag{5.11}$$

for two independent variables $x$ and $y$. Here $k$ is thermal diffusivity and $f(x, y, t)$ is the heat source. $\Omega$ is the domain of interest, usually a physical region. If there is no heat source ($q = f/\kappa = 0$), it becomes the Laplace equation. The solution of a function is said to be harmonic if it satisfies Laplace's equation.

In order to determine the temperature $u$ completely, the appropriate boundary conditions are needed. A simple boundary condition is to specify the temperature $u = u_0$ on the boundary $\partial\Omega$. This type of problem is the Dirichlet problem.

On the other hand, if the temperature is not known, but the gradient $\partial u/\partial \boldsymbol{n}$ is known on the boundary where $\boldsymbol{n}$ is the outward-pointing unit normal, this forms the Neumann problem. Furthermore, some problems may have a mixed type of boundary conditions in the combination of

$$\alpha u + \beta \frac{\partial u}{\partial \boldsymbol{n}} = \gamma,$$

which naturally occurs as a radiation or cooling boundary condition.

### 5.2.2 Parabolic Equation

Time-dependent problems, such as diffusion and transient heat conduction, are governed by the parabolic equation

$$u_t = ku_{xx}. \tag{5.12}$$

Written in the $n$-dimensional case $x_1 = x, x_2 = y, x_3 = z, ...$, it can be extended to the reaction-diffusion equation

$$u_t = k\nabla^2 u + f(u, x_1, .., x_n, t). \tag{5.13}$$

### 5.2.3 Wave Equation

The vibration of strings and travelling seismic waves are governed by the hyperbolic wave equation.

The 1-D wave equation in its simplest form is

$$u_{tt} = c^2 u_{xx}, \tag{5.14}$$

where $c$ is the velocity of the wave. Using a transformation of the pair of independent variables

$$\xi = x + ct, \tag{5.15}$$

and

$$\eta = x - ct, \tag{5.16}$$

for $t > 0$ and $-\infty < x < \infty$, the wave equation can be written as

$$u_{\xi\eta} = 0. \tag{5.17}$$

Integrating twice and substituting back in terms of $x$ and $t$, we have

$$u(x, t) = f(x + ct) + g(x - ct), \tag{5.18}$$

where $f$ and $g$ are functions of $x + ct$ and $x - ct$, respectively. We can see that the solution is composed of two independent waves. One wave moves to the right and one travels to the left at the same constant speed $c$.

## 5.3 Other Mathematical Models

We have shown examples of the three major equations of second-order linear partial differential equations. There are other equations that occur frequently in mathematical physics, engineering and earth sciences. We will give a brief description of some of these equations.

### 5.3.1 Elastic Wave Equation

A wave in an elastic isotropic homogeneous solid is governed by the following equation in terms of displacement $\mathbf{u}$,

$$\rho\frac{\partial^2\mathbf{u}}{\partial t^2} = \mu\nabla^2\mathbf{u} + (\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) + \mathbf{f}, \tag{5.19}$$

where $\rho$ is density, $\lambda$ and $\mu$ are Lamé constants, and $\mathbf{f}$ is body force. Such an equation can describe two types of wave: transverse wave (S wave) and longitudinal or dilatational wave (P wave). The speed of the longitudinal wave is

$$v_p = \sqrt{(\lambda+2\mu)/\rho}, \tag{5.20}$$

and the transverse wave has the speed

$$v_s = \sqrt{\mu/\rho}. \tag{5.21}$$

### 5.3.2 Reaction-Diffusion Equation

The reaction-diffusion equation is an extension of heat conduction with a source $f$

$$u_t = D\Delta u + f(x,y,z,u), \tag{5.22}$$

where $D$ is the diffusion coefficient and $f$ is the reaction rate. One example is the combustion equation

$$u_t = Du_{xx} + Que^{-\lambda/u}, \tag{5.23}$$

where $Q$ and $\lambda$ are constants.

### 5.3.3 Navier-Stokes Equations

The Navier-Stokes equations for incompressible flow in the absence of body forces can be written, in terms of the velocity $\boldsymbol{u}$ and the pressure $p$, as

$$\nabla\cdot\mathbf{u} = 0, \qquad \rho[\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u}] = \mu\nabla^2\mathbf{u} - \nabla p, \tag{5.24}$$

where $\rho$ and $\mu$ are the density of the fluid and its viscosity, respectively. In computational fluid dynamics, most simulations are mainly related to these equations. We can define the Reynolds number as $\mathrm{Re} = \rho UL/\mu$ where $U$ is the typical velocity and $L$ is the length scale.

In the limit of $\mathrm{Re} \ll 1$, we have the Stokes flow governed by

$$\mu\nabla^2\mathbf{u} = \nabla p. \tag{5.25}$$

In the other limit of $\mathrm{Re} \gg 1$, we have the inviscous flow

$$\nabla\cdot\mathbf{u} = 0, \qquad \rho[\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u}] = -\nabla p, \tag{5.26}$$

where there is still a nonlinear term $(\mathbf{u}\cdot\nabla)\mathbf{u}$.

### 5.3.4 Groundwater Flow

The general equation for three-dimensional groundwater flow is

$$S_\sigma \frac{\partial p}{\partial t} = \frac{k}{\mu}\nabla^2 p - S_\sigma B \frac{\partial \sigma}{\partial t} + Q, \tag{5.27}$$

where $\sigma = \sigma_{kk}/3$ is the mean stress, $p$ is the pore water pressure, and $Q$ is source or sink term. $S_\sigma$ is the specific storage coefficient and $B$ is the Skempton constant. $k$ is the permeability of the porous medium and $\mu$ is the viscosity of water. This can be considered as the inhomogeneous diffusion equation for pore pressure.

## 5.4 Solution Techniques

Each type of equation usually requires different solution techniques. However, there are some methods that work for most of the linearly partial differential equations with appropriate boundary conditions on a regular domain. These methods include separation of variables, method of series expansion and transform methods such as the Laplace and Fourier transforms.

### 5.4.1 Separation of Variables

The separation of variables attempts a solution of the form

$$u = X(x)Y(y)Z(z)T(t), \tag{5.28}$$

where $X(x), Y(y), Z(z), T(t)$ are functions of $x, y, z, t$, respectively. By determining these functions that satisfy the partial differential equation and the required boundary conditions in terms of eigenvalue problems, the solution of the original problem is then obtained.

As a classic example, we now try to solve the 1-D heat conduction equation in the domain $x \in [0, L]$ and $t \geq 0$

$$u_t = k u_{xx}, \tag{5.29}$$

with the initial value and boundary conditions

$$u(0,t) = 0, \qquad \frac{\partial u(x,t)}{\partial x}\Big|_{x=L} = 0, \qquad u(x,0) = \psi(x). \tag{5.30}$$

Letting $u(x,t) = X(x)T(t)$, we have

$$\frac{X''(x)}{X} = \frac{T'(t)}{kT}. \tag{5.31}$$

As the left-hand side depends only on $x$ and the right hand side only depends on $t$, therefore, both sides must be equal to the same constant, and the constant can be assumed to be $-\lambda^2$. The negative sign is just for convenience because we will see below that the finiteness of the solution $T(t)$ requires that eigenvalues $\lambda^2 > 0$ or $\lambda$ are real. Hence, we now get two ordinary differential equations

$$X''(x) + \lambda^2 X(x) = 0, \qquad T'(t) + k\lambda^2 T(t) = 0, \tag{5.32}$$

where $\lambda$ is the eigenvalue. The solution for $T(t)$ is

$$T = A_n e^{-\lambda^2 kt}. \tag{5.33}$$

The basic solution for $X(x)$ is simply

$$X(x) = \alpha \cos \lambda x + \beta \sin \lambda x. \tag{5.34}$$

So the fundamental solution for $u$ is

$$u(x,t) = (\alpha \cos \lambda x + \beta \sin \lambda x) e^{-\lambda^2 kt}, \tag{5.35}$$

where we have absorbed the coefficient $A_n$ into $\alpha$ and $\beta$ because they are the undetermined coefficients anyway. As the value of $\lambda$ varies with the boundary conditions, it forms an eigenvalue problem. The general solution for $u$ should be derived by superposing solutions of (5.35), and we now have

$$u = \sum_{n=1}^{\infty} X_n T_n = \sum_{n=1}^{\infty} (\alpha_n \cos \lambda_n x + \beta_n \sin \lambda_n x) e^{-\lambda_n^2 kt}. \tag{5.36}$$

From the boundary condition $u(0,t) = 0$ at $x = 0$, we have

$$0 = \sum_{n=1}^{\infty} \alpha_n e^{-\lambda_n^2 kt}, \tag{5.37}$$

which leads to $\alpha_n = 0$ since $\exp(-\lambda^2 kt) > 0$.

From $\left.\frac{\partial u}{\partial x}\right|_{x=L} = 0$, we have

$$\lambda_n \cos \lambda_n L = 0, \tag{5.38}$$

which requires

$$\lambda_n L = \frac{(2n-1)\pi}{2}, \qquad (n = 1, 2, ...). \tag{5.39}$$

Therefore, $\lambda$ cannot be continuous, and it only takes an infinite number of discrete values, called eigenvalues.

Each eigenvalue $\lambda = \lambda_n = \frac{(2n-1)\pi}{2L}, (n = 1, 2, ...)$ has a corresponding eigenfunction $X_n = \sin(\lambda_n x)$. Substituting into the solution for $T(t)$, we have

$$T_n(t) = A_n e^{-\frac{[(2n-1)\pi]^2}{4L^2}kt}. \tag{5.40}$$

By expanding the initial condition into a Fourier series so as to determine the coefficients, we have

$$u(x,t) = \sum_{n=1}^{\infty} \beta_n \sin(\frac{(2n-1)\pi x}{2L}) e^{-[\frac{(2n-1)\pi}{2L}]^2 kt},$$

$$\beta_n = \frac{2}{L}\int_0^L \psi(x) \sin[\frac{(2n-1)\pi x}{2L}]dx. \tag{5.41}$$

---

**Example 5.1:** In the special case when initial condition $u(x, t = 0) = \psi = u_0$ is constant, the requirement for $u = u_0$ at $t = 0$ becomes

$$u_0 = \sum_{n=1}^{\infty} \beta_n \sin\frac{(2n-1)\pi x}{2L}. \tag{5.42}$$

Using the orthogonal relationships

$$\int_0^L \sin\frac{m\pi x}{L}\sin\frac{n\pi x}{L}dx = 0, \qquad m \neq n,$$

and

$$\int_0^L (\sin\frac{n\pi x}{L})^2 dx = \frac{L}{2}, \qquad (n = 1, 2, ...),$$

and multiplying both sides of Eq.(5.42) by $\sin[(2n-1)\pi x/2L]$, we have the integration

$$\beta_n \frac{L}{2} = \int_0^L \sin\frac{(2n-1)\pi x}{2L} u_0 dx = \frac{2u_0 L}{(2n-1)\pi}, \quad (n = 1, 2, ...),$$

which leads to

$$\beta_n = \frac{4u_0}{(2n-1)\pi}, \qquad n = 1, 2, ...,$$

and thus the solution becomes

$$u = \frac{4u_0}{\pi}\sum_{n=1}^{\infty} \frac{1}{(2n-1)} e^{-\frac{(2n-1)^2\pi^2 kt}{4L^2}} \sin\frac{(2n-1)\pi x}{2L}. \tag{5.43}$$

This solution is essentially the same as the classical heat conduction problem discussed by Carslaw and Jaeger in 1959. This same solution can also be obtained using the Fourier series of $u_0$ in $0 < x < L$.

---

### 5.4.2 Laplace Transform

The integral transform can reduce the number of the independent variables. For the 1-D time-dependent case, it transforms a partial differential equation into an ordinary differential equation. By solving the ordinary differential equation and inverting it back, we can obtain the solution for the original partial differential equation. As an example, we now solve the heat conduction problem over a semi-infinite interval $[0, \infty)$,

$$u_t = k u_{xx}, \qquad u(x,0) = 0, \;\; u(0,t) = T_0. \tag{5.44}$$

---

**Example 5.2:** Let $\bar{u}(x,s) = \int_0^\infty u(x,t) e^{-st} dt$ be the Laplace transform of $u(x,t)$, then Eq.(5.44) becomes

$$s\bar{u} = k\frac{d^2\bar{u}}{dx^2}, \qquad \bar{u}_{x=0} = \frac{T_0}{s},$$

which is an ordinary differential equation whose general solution can be written as

$$\bar{u} = Ae^{-\sqrt{\frac{s}{k}}x} + Be^{\sqrt{\frac{s}{k}}x}.$$

The finiteness of the solution as $x \to \infty$ requires that $B = 0$, and the boundary condition at $x = 0$ leads to

$$\bar{u} = \frac{T_0}{s} e^{-\sqrt{\frac{s}{k}}x}.$$

By using the inverse Laplace transform, we have

$$u = T_0 \text{erfc}(\frac{x}{2\sqrt{kt}}),$$

where erfc$(x)$ is the complementary error function.

---

### 5.4.3 Fourier Transform

The Fourier transform works in a similar manner to the Laplace transform. The famous example is the classical wave equation

$$u_{tt} = c^2 u_{xx}, \qquad u(x,0) = \psi(x), \;\; u_t(x,0) = 0. \tag{5.45}$$

Let $\bar{u}(\omega, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x,t) e^{i\omega x} dx$ be the Fourier transform of $u(x,t)$. This transforms the PDE problem into an ODE

$$\frac{d^2\bar{u}}{dt^2} = -c^2\omega^2\bar{u}, \;\; \bar{u}(\omega,0) = \bar{\psi}(\omega), \;\; \frac{d\bar{u}(\omega,0)}{dt} = 0.$$

The general solution in terms of the parameter $\omega$ is

$$\bar{u}(\omega, t) = \bar{\psi}(\omega)\cos(c\omega t).$$

By using the inverse Fourier transform, we finally have

$$\begin{aligned} u(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \bar{\psi}(\omega)\cos(c\omega t)e^{-i\omega x}d\omega \\ &= \frac{1}{2}[\psi(x+ct) + \psi(x-ct)], \end{aligned} \tag{5.46}$$

which implies two travelling waves: one travels along the $x$-axis and the other along the negative $x$-axis direction.

### 5.4.4 Similarity Solution

Sometimes, the diffusion equation

$$u_t = \kappa u_{xx}, \tag{5.47}$$

can be solved by using the so-called similarity method by defining a similar variable

$$\eta = \frac{x}{\sqrt{\kappa t}} \tag{5.48}$$

or

$$\zeta = \frac{x^2}{\kappa t}. \tag{5.49}$$

One can assume that the solution to the equation has the form

$$u = (\kappa t)^{\alpha} f[\frac{x^2}{(\kappa t)^{\beta}}]. \tag{5.50}$$

By substituting it into the diffusion equation, the coefficients $\alpha$ and $\beta$ can be determined. For most applications, one can assume $\alpha = 0$ so that $u = f(\zeta)$. In this case, we have

$$4\zeta u'' + 2u' + \zeta\beta(\kappa t)^{\beta-1}u' = 0, \tag{5.51}$$

where $u' = du/d\zeta$. In deriving this equation, one has to use the chain rules of differentiations $\frac{\partial}{\partial x} = \frac{\partial}{\partial \zeta}\frac{\partial \zeta}{\partial x}$ and $\frac{\partial}{\partial t} = \frac{\partial}{\partial \zeta}\frac{\partial \zeta}{\partial t}$.

Since the original equation does not have time-dependent terms explicitly, this means that all the exponents for any $t$-terms must be zero. Therefore, we have

$$\beta = 1. \tag{5.52}$$

Now, the diffusion equation becomes

$$\zeta f''(\zeta) = -(\frac{1}{2} + \frac{\zeta}{4})f'. \tag{5.53}$$

Using $(\ln f')' = f''/f'$ and integrating the above equation once, we get

$$f' = \frac{Ke^{-\zeta/4}}{\sqrt{\zeta}}. \tag{5.54}$$

Integrating it again and using the substitution $\zeta = 4\xi^2$, we obtain

$$u = A\int_0^{\xi} e^{-\xi^2} d\xi = C\text{erf}(\frac{x}{\sqrt{4\kappa t}}) + D, \tag{5.55}$$

where $C$ and $D$ are constants that can be determined from appropriate boundary conditions.

---

**Example 5.3:** For the same problem as (5.44), the boundary condition as $x \to \infty$ implies that $C + D = 0$, while $u(0,t) = T_0$ means that $D = -C = T_0$. Therefore, we finally have

$$u = T_0[1 - \text{erf}(\frac{x}{\sqrt{4\kappa t}})] = T_0\text{erfc}(\frac{x}{\sqrt{4\kappa t}}).$$

---

### 5.4.5 Change of Variables

In some cases, the partial differential equation may not be written in the standard form; however, it can be converted into a known standard equation by a change of variables. For example, the following simple reaction-diffusion equation

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} - \alpha u, \tag{5.56}$$

describes the heat conduction along a wire with a heat loss term $-\alpha u$. Carslaw and Jaeger show that it can be transformed into a standard equation of heat conduction using the following change of variables

$$u = ve^{-\alpha t}, \tag{5.57}$$

where $v$ is the new variable. By simple differentiations, we have

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial t}e^{-\alpha t} - \alpha v e^{-\alpha t} = \frac{\partial v}{\partial t}e^{-\alpha t} - \alpha u, \qquad \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x^2}e^{-\alpha t}, \tag{5.58}$$

we have

$$\frac{\partial u}{\partial t} = \underbrace{\frac{\partial v}{\partial t}e^{-\alpha t}} - \alpha u = k\frac{\partial^2 u}{\partial x^2} - \alpha u = \underbrace{k\frac{\partial^2 v}{\partial x^2}e^{-\alpha t}} - \alpha u, \tag{5.59}$$

which becomes

$$\frac{\partial v}{\partial t}e^{-\alpha t} = k\frac{\partial^2 v}{\partial x^2}e^{-\alpha t}. \tag{5.60}$$

After dividing both sides by $e^{-\alpha t} > 0$, we have

$$\frac{\partial v}{\partial t} = k\frac{\partial^2 v}{\partial x^2}, \tag{5.61}$$

which is the standard heat conduction equation for $v$.

For given initial (usually constant) and boundary conditions (usually zero), we can use all the techniques for solving the standard equation to get solutions. However, for some boundary conditions such as $u = u_0$, a more elaborate form of change of variables is needed. Crank introduced Danckwerts's method by using the following transform

$$u = \alpha \int_0^t ve^{-\alpha\tau}d\tau + ve^{-\alpha t}. \tag{5.62}$$

Noting that $\frac{\partial u}{\partial t} = \alpha ve^{-\alpha t} - \alpha ve^{-\alpha t} + \frac{\partial v}{\partial t}e^{-\alpha t}$, it is straightforward to show

$$\frac{\partial u}{\partial t} + \alpha u = k\frac{\partial^2 u}{\partial x^2}. \tag{5.63}$$

For the boundary condition $u = u_0$, we have $v = v_0 = u_0$, and this is because

$$u = u_0 = \alpha v_0 \int_0^t e^{-\alpha\tau}d\tau + v_0 e^{-\alpha t} = v_0 - v_0 e^{-\alpha t} + v_0 e^{-\alpha t} = v_0, \tag{5.64}$$

which is the same boundary condition for $u$.

There are other important methods for solving partial differential equations. These include Green's function, series methods, asymptotic methods, approximate methods, perturbation methods and naturally the numerical methods.

# References

Carrrier G. F. and Pearson C. E., *Partial Differential Equations: Theory and Technique*, 2nd Edition, Academic Press, (1988).

Carslaw H. S. and Jaeger J. C., *Conduction of Heat in Solids*, 2nd Edition, Oxford University Press, (1986).

Crank J., *Mathematics of Diffusion*, Clarendon Press, Oxford, (1970).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, Cambridge University Press, (2006).

# Chapter 6

# Calculus of Variations

The calculus of variations is important in many optimisation problems and computational sciences, especially in the formulation of the finite element methods. In this chapter, we will briefly touch on these topics.

The main aim of the calculus of variations is to find a function that makes an integral stationary, making the value of the integral a local maximum or minimum. For example, in mechanics we may want to find the shape $y(x)$ of a rope or chain when suspended under its own weight from two fixed points. In this case, the calculus of variations provides a method for finding the function $y(x)$ so that the curve $y(x)$ minimises the gravitational potential energy of the hanging rope system.

## 6.1 Euler-Lagrange Equation

### 6.1.1 Curvature

Before we proceed to the calculus of variations, let us first discuss an important concept, namely the curvature of a curve. In general, a curve $y(x)$ can be described in a parametric form in terms of a vector $\mathbf{r}(s)$ with a parameter $s$ which is the arc length along the curve measured from a fixed point. The curvature $\kappa$ of a curve is defined as the rate at which the unit tangent $\mathbf{t}$ changes with respect to $s$. The change of arc length is

$$\frac{ds}{dx} = \sqrt{1 + (\frac{dy}{dx})^2} = \sqrt{1 + y'^2}. \tag{6.1}$$

We have the curvature

$$\frac{d\mathbf{t}}{ds} = \kappa\, \mathbf{n} = \frac{1}{\rho}\mathbf{n}, \tag{6.2}$$

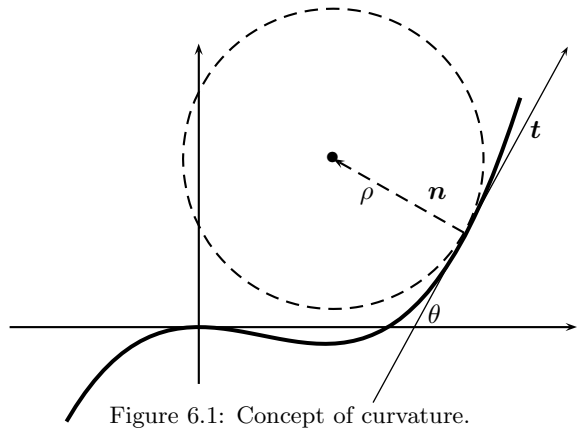



Figure 6.1: Concept of curvature.

where $\rho$ is the radius of the curvature, and $\mathbf{n}$ is the principal normal shown in Fig. 6.1.

As the direction of the tangent is defined by the angle $\theta$ made with the $x$-axis by $\mathbf{t}$, we have $\tan\theta = y'$. Hence, the curvature becomes

$$\kappa = \frac{d\theta}{ds} = \frac{d\theta}{dx}\frac{dx}{ds}. \tag{6.3}$$

From $\theta = \tan^{-1} y'(x)$, we have

$$\frac{d\theta}{dx} = [\tan^{-1}(y')]' = \frac{y''}{(1+y'^2)}. \tag{6.4}$$

Using the expression for $ds/dx$, the curvature can be written in terms of $y(x)$, and we get

$$\kappa = |\frac{d^2\mathbf{r}}{ds^2}| = \left|\frac{y''}{[1+(y')^2]^{3/2}}\right|. \tag{6.5}$$

---

**Example 6.1:** We know that the equation of a circle centered at $(0,0)$ with a radius $r$ is

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

In order to calculate the curvature at any point $(x,y)$ along the circle, we have to calculate $y'$ and $y''$. Using implicit differentiation with respect to $x$, we have

$$2x + 2y'y = 0,$$

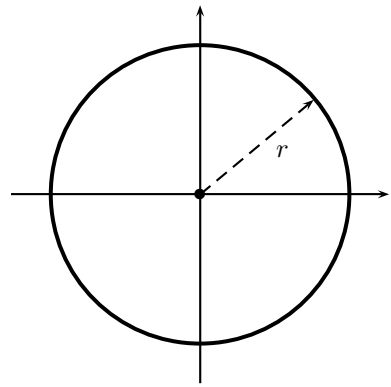



Figure 6.2: The radius of curvature of a circle is its radius $r$.

which leads to

$$y' = -\frac{x}{y}.$$

The second derivative is

$$y'' = -\frac{x^2 + y^2}{y^3}.$$

Using these expressions and $x^2 + y^2 = r^2$, we have the curvature

$$\kappa = |\frac{y''}{[1 + (y')^2]^{3/2}}| = \left|-\frac{r^2/y^3}{[1 + (-x/y)^2]^{3/2}}\right|$$

$$= \left|-\frac{1}{y\sqrt{\frac{r^2}{y^2}}}\right| = \frac{1}{r}.$$

Indeed, the curvature of a circle is everywhere $1/r$. Thus, the radius of curvature is the radius of the circle $\rho = 1/\kappa = r$.

### 6.1.2 Euler-Lagrange Equation

Since the calculus of variations is always related to some minimisation or maximisation, we can in general assume that the integrand $\psi$ of the integral is a function of the shape or curve $y(x)$ (shown in Figure 6.3), its derivative $y'(x)$ and the spatial coordinate $x$ (or time $t$, depending on the context). For the integral

$$I = \int_a^b \psi(x, y, y')dx, \tag{6.6}$$

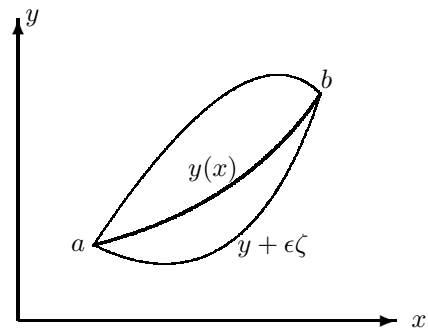



Figure 6.3: Variations in the path $y(x)$.

where $a$ and $b$ are fixed, the aim is to find the solution of the curve $y(x)$ such that it makes the value of $I$ stationary or optimal. In this sense, $I[y(x)]$ is a function of the function $y(x)$, and thus it is referred to as a functional.

Here, stationary means that a small first-order change in $y(x)$ will only lead to the second-order changes in the values of $I[y(x)]$, and consequently, the change $\delta I$ of $I$ should be virtually zero due to any small variations in the function $y(x)$. Translating this into mathematical language, we suppose that $y(x)$ has a small change of magnitude of $\epsilon$ so that

$$y(x) \to y(x) + \epsilon\zeta(x), \tag{6.7}$$

where $\zeta(x)$ is an arbitrary function. The requirement of $I$ to be stationary means that

$$\delta I = 0, \tag{6.8}$$

or more accurately,

$$\frac{\delta I}{\delta \epsilon}|_{\epsilon=0} = 0, \qquad \text{for all } \zeta(x). \tag{6.9}$$

Using the multivariate Taylor series, we have

$$I(y,\epsilon) = \int_a^b \psi(x, y+\epsilon\zeta, y'+\epsilon\zeta')dx$$

$$= \int_a^b \psi(x,y,y')dx + \int_a^b [\epsilon(\zeta\frac{\partial\psi}{\partial y} + \zeta'\frac{\partial\psi}{\partial y'})]dx + O(\epsilon^2). \tag{6.10}$$

The first derivative of $I$ should be zero, and we have

$$\frac{\delta I}{\delta \epsilon} = \int_a^b [\frac{\partial \psi}{\partial y}\zeta + \frac{\partial \psi}{\partial y'}\zeta']dx = 0, \tag{6.11}$$

which is exactly what we mean that the change $\delta I$ (or the first-order variation) in the value of $I$ should be zero. Integrating this equation by parts, we have

$$\int_a^b [\frac{\partial \psi}{\partial y} - \frac{d}{dx}\frac{\partial \psi}{\partial y'}]\zeta dx = -\left[\zeta \frac{\partial \psi}{\partial y'}\right]_a^b. \tag{6.12}$$

If we require that $y(a)$ and $y(b)$ are fixed at the points $x = a$ and $x = b$, then these requirements naturally lead to $\zeta(a) = \zeta(b) = 0$. This means that the right-hand side of the equation is zero. That is,

$$\left[\zeta \frac{\partial \psi}{\partial y'}\right]_a^b = 0, \tag{6.13}$$

which gives

$$\int_a^b \left[\frac{\partial \psi}{\partial y} - \frac{d}{dx}\frac{\partial \psi}{\partial y'}\right]\zeta dx = 0. \tag{6.14}$$

As this equation holds for all $\zeta(x)$, the integrand must be zero. Therefore, we have the well-known Euler-Lagrange equation

$$\frac{\partial \psi}{\partial y} = \frac{d}{dx}(\frac{\partial \psi}{\partial y'}). \tag{6.15}$$

It is worth pointing out that this equation is very special in the sense that $\psi$ is known and the unknown is $y(x)$. It has many applications in mathematics, physics, engineering and earth sciences.

The simplest and classical example is to find the shortest path on a plane joining two points, say $(0, 0)$ and $(1, 1)$. We know that the total length along a curve $y(x)$ is

$$L = \int_0^1 \sqrt{1 + y'^2}dx. \tag{6.16}$$

Since $\psi = \sqrt{1 + y'^2}$ does not contain $y$, thus $\frac{\partial \psi}{\partial y} = 0$. From the Euler-Lagrange equation, we have

$$\frac{d}{dx}(\frac{\partial \psi}{\partial y'}) = 0, \tag{6.17}$$

its integral is

$$\frac{\partial \psi}{\partial y'} = \frac{y'}{\sqrt{1 + y'^2}} = A. \tag{6.18}$$

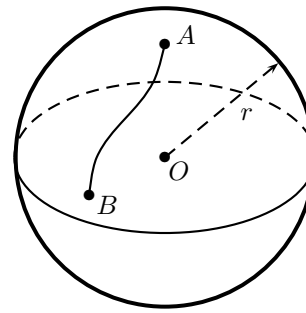



Figure 6.4: Geodesic path on the surface of a sphere.

Rearranging it as

$$y'^2 = \frac{A^2}{1-A^2}, \qquad \text{or} \qquad y' = \frac{A}{\sqrt{1-A^2}}, \tag{6.19}$$

and integrating again, we have

$$y = kx + c, \qquad k = \frac{A}{\sqrt{1-A^2}}. \tag{6.20}$$

This is a straight line. That is exactly what we expect from the plane geometry.

Well, you may say, this is trivial and there is nothing new about it. Let us now study a slightly more complicated example to find the shortest path on the surface of a sphere.

---

**Example 6.2:** For any two points $A$ and $B$ on the surface of a sphere with radius $r$ as shown in Fig. 6.4, we now use the calculus of variations to find the shortest path connecting $A$ and $B$ on the surface.

Since the sphere has a fixed radius, we need only two coordinates $(\theta, \phi)$ to uniquely determine the position on the sphere. The length element $ds$ can be written in terms of the two spherical coordinate angles

$$ds = r\sqrt{d\theta^2 + \sin^2\theta d\phi^2} = r\sqrt{(\frac{d\theta}{d\phi})^2 + \sin^2\theta}\ \ |d\phi|,$$

where in the second step we assume that $\theta = \theta(\phi)$ is a function of $\phi$ only, so that $\phi$ becomes the only independent variable. This is possible because $\theta(\phi)$ represents a curve on the surface of the sphere just as $y = y(x)$

represents a curve on a plane. Thus, we want to minimise the total length

$$L = \int_A^B ds = \int_{\phi_A}^{\phi_B} \sqrt{\theta'^2 + \sin^2\theta}\, d\phi,$$

where $\theta' = d\theta/d\phi$. Since the integrand

$$\psi = \sqrt{\theta'^2 + \sin^2\theta}$$

does not explicitly depend on $\phi$, we can use the simplified form of Euler-Lagrange equation (6.34) discussed later

$$\psi - \theta' \frac{\partial \psi}{\partial \theta'} = k,$$

where $k$ is a constant. We have

$$\sqrt{\theta'^2 + \sin^2\theta} - \theta' \frac{\theta'}{\sqrt{\theta'^2 + \sin^2\theta}} = k,$$

or

$$\theta'^2 = (\frac{d\theta}{d\phi})^2 = \frac{\sin^2\theta(\sin^2\theta - k^2)}{k^2}.$$

By taking the square roots and rearranging the above equation, we get

$$d\phi = \pm \frac{k d\theta}{\sin\theta\sqrt{\sin^2\theta - k^2}} = \pm \frac{\frac{1}{\sin^2\theta} d\theta}{\sqrt{(\frac{1}{k^2} - 1) - \cot^2\theta}}.$$

Its integration gives

$$\phi = \mp \sin^{-1}[\frac{\cot\theta}{\sqrt{\frac{1}{k^2} - 1}}] + \alpha.$$

where $\alpha$ is the integration constant. Taking $\sin$ of both sides, we have

$$\sin(\phi - \alpha) = \pm\beta\cot\theta = \sin\phi\cos\alpha - \sin\alpha\cos\phi,$$

where

$$\beta = \frac{1}{\sqrt{\frac{1}{k^2} - 1}}.$$

Multiplying both sides by $r\sin\theta$ and using $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$ and $z = r\cos\theta$, we have

$$y\cos\alpha - x\sin\alpha = \pm\beta z,$$

which corresponds to a plane passing through points $A$, $B$, and the origin $(x, y, z) = (0, 0, 0)$. Therefore, the intersection of the plane and sphere produces a great circle on the surface connecting $A$ and $B$, as the two signs correspond to two segments of the great circle, one of which is shorter than the other. However, in the special case when $A$ and $B$ are opposite points, the two segments will have the same length. Great circles are as important in geodesy as straight lines in plane geometry.

These examples are relatively simple. Let us now study a more complicated case so as to demonstrate the wide range of applications of the Euler-Lagrange equation. In mechanics, Hamilton's principle states that the configuration of a mechanical system is such that the action integral $I$ (the integral of the Lagrangian $\mathcal{L} = T - V$) is stationary with respect to variations in the path. That is to say that the configuration can be uniquely defined by its coordinates $q_i$ and time $t$, when moving from one configuration at time $t_0$ to another time $t = t^*$

$$I = \int_{t_0}^{t^*} \mathcal{L}(t, q_i, \dot{q}_i)dt, \qquad i = 1, 2, ..., N, \tag{6.21}$$

where $T$ is the total kinetic energy (usually, a function of $\dot{q}_i$), and $V$ is the potential energy (usually, a function of $q_i$). Here $\dot{q}_i$ means

$$\dot{q}_i = \frac{\partial q_i}{\partial t}. \tag{6.22}$$

In analytical mechanics, the Lagrangian $\mathcal{L}$ (=kinetic energy − potential energy) is often called the action, thus this principle is also called the principle of least action. The physical configuration or the path of movement follows a path that makes the action integral stationary.

In the special case, $x \to t$, the Euler-Lagrange equation becomes

$$\frac{\partial \mathcal{L}}{\partial q_i} = \frac{d}{dt}(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}), \tag{6.23}$$

which is the well-known Lagrange's equation. This seems too abstract so now let us look at a classic example.

**Example 6.3:** For a simple pendulum shown in Figure 6.5, we now try to derive its equation of oscillations. We know the kinetic energy $T$ and the potential energy $V$ are

$$T = \frac{1}{2}ml^2(\frac{d\theta}{dt})^2 = \frac{1}{2}ml^2\dot{\theta}^2, \qquad V = mgh = mgl(1 - \cos\theta).$$

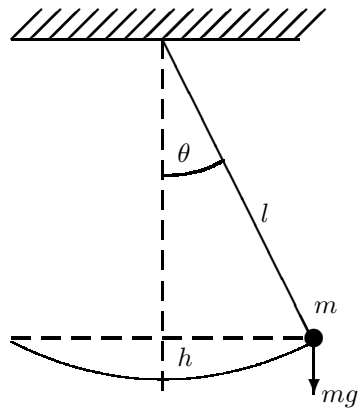


Figure 6.5: A simple pendulum.

Using $\mathcal{L} = T - V$, $q = \theta$ and $\dot{q} = \dot{\theta}$, we have

$$\frac{\partial \mathcal{L}}{\partial \theta} - \frac{d}{dt}(\frac{\partial \mathcal{L}}{\partial \dot{\theta}}) = 0,$$

which becomes

$$-mgl\sin\theta - \frac{d}{dt}(ml^2\dot{\theta}) = 0.$$

Therefore, we have the pendulum equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{l}\sin\theta = 0.$$

This is a nonlinear equation. If the angle is very small ($\theta \ll 1$), so $\sin\theta \approx \theta$, we then have the standard equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{l}\theta = 0.$$

for linear harmonic motion.

## 6.2 Variations with Constraints

Although the stationary requirement in the calculus of variations leads to the minimisation of the integral itself, there is no additional constraint. In this sense, the calculus of variations discussed up to now is unconstrained. However, sometimes these variations have certain additional constraints, for example, the sliding of a bead on a hanging

string. Now we want to make the integral $I$ stationary under another constraint integral $Q$ that is constant. We have

$$I = \int_a^b \psi(x, y, y')dx, \tag{6.24}$$

subjected to the constraint

$$Q = \int_a^b \phi(x, y, y')dx. \tag{6.25}$$

As for most optimisation problems under additional constraints, the method of Lagrange multipliers can transform the constrained problem into an unconstrained one by using a combined functional $J = I + \lambda Q$

$$J = \int_a^b [\psi + \lambda\phi]dx, \tag{6.26}$$

where $\lambda$ is the undetermined Lagrange multiplier. Replacing $\psi$ by $[\psi + \lambda\phi]$ in the Euler-Lagrange equation or following the same derivation, we have

$$[\frac{\partial\psi}{\partial y} - \frac{d}{dx}(\frac{\partial\psi}{\partial y'})] + \lambda[\frac{\partial\phi}{\partial y} - \frac{d}{dx}(\frac{\partial\phi}{\partial y'})] = 0. \tag{6.27}$$

Now we can come back to our example of the hanging rope problem with two fixed points. The total length of the rope is $L$, and it hangs from two fixed points $(-d, 0)$ and $(d, 0)$. From geometric consideration, we require that $2d < L$. In order to find the shape of the hanging rope under gravity, we now define its total gravitational potential energy $E_p$ as

$$E_p = \int_{x=-d}^{d} [\rho g y(x) ds] = \rho g \int_{-d}^{d} y\sqrt{1 + y'^2}dx, \tag{6.28}$$

where $\rho$ is the mass per unit length of the rope. The additional constraint is that the total length of the rope is a constant $(L)$. Thus,

$$Q = \int_{-d}^{d} \sqrt{1 + y'^2}dx = L. \tag{6.29}$$

By using the Lagrange multiplier $\lambda$, we have $J = E_p + \lambda Q$, or

$$J = \int_{-d}^{d} [\rho g y + \lambda]\sqrt{1 + y'^2}dx. \tag{6.30}$$

Since $\Psi = [\rho g y + \lambda]\sqrt{1 + y'^2}$ does not contain $x$ explicitly, or $\frac{\partial\Psi}{\partial x} = 0$, the Euler-Lagrange equation can be reduced into a simpler form in this special case. Using

$$\frac{d\Psi}{dx} = \frac{\partial\Psi}{\partial x} + \frac{\partial\Psi}{\partial y}\frac{dy}{dx} + \frac{\partial\Psi}{\partial y'}\frac{dy'}{dx} = 0 + y'\frac{\partial\Psi}{\partial y} + y''\frac{\partial\Psi}{\partial y'}, \tag{6.31}$$

and the Euler-Lagrange equation $\frac{\partial \Psi}{\partial y} = \frac{d}{dx}(\frac{\partial \Psi}{\partial y'})$, we have

$$\frac{d\Psi}{dx} = y'[\frac{d}{dx}(\frac{\partial \Psi}{\partial y'})] + y''\frac{\partial \Psi}{\partial y'} = \frac{d}{dx}[y'\frac{\partial \Psi}{\partial y'}], \tag{6.32}$$

which can again be written as

$$\frac{d}{dx}[\Psi - y'\frac{\partial \Psi}{\partial y'}] = 0. \tag{6.33}$$

The integration of this equation gives

$$\Psi - y'\frac{\partial \Psi}{\partial y'} = A = const. \tag{6.34}$$

Substituting the expression for $\Psi$ into the above equation, the stationary value of $J$ requires

$$\sqrt{1+y'^2} - \frac{y'^2}{\sqrt{1+y'^2}} = \frac{A}{\rho g y + \lambda}. \tag{6.35}$$

Multiplying both sides by $\sqrt{1+y'^2}$ and using the substitution $A\cosh\zeta = \rho g y + \lambda$, we have

$$y'^2 = \cosh^2\zeta - 1, \tag{6.36}$$

whose solution is

$$\cosh^{-1}[\frac{\rho g y + \lambda}{A}] = \pm\frac{x\rho g}{A} + K. \tag{6.37}$$

Using the boundary conditions $y = 0$ at $x = \pm d$ and the constraint $Q = L$, we have $K = 0$ and an implicit equation for $A$

$$\sinh(\frac{\rho g d}{A}) = \frac{\rho g L}{2A}. \tag{6.38}$$

Finally, the curve for the hanging rope becomes the following catenary

$$y(x) = \frac{A}{\rho g}[\cosh(\frac{\rho g x}{A}) - \cosh(\frac{\rho g d}{A})]. \tag{6.39}$$

---

**Example 6.4:** Dido's problem concerns the strategy to enclose a maximum area with a fixed length circumference. Legend says that Dido was promised a piece of land on the condition that it was enclosed by an oxhide. She had to cover as much land as possible using the given oxhide. She cut the oxhide into narrow strips with ends joined, and a whole region of a hill was enclosed.

Suppose the total length of the oxhide strip is $L$. The enclosed area $A$ to be maximised is

$$A = \int_{x_a}^{x_b} y(x)dx,$$

where $x_a$ and $x_b$ are two end points (of course they can be the same points). We also have the additional constraint

$$\int_{x_a}^{x_b} \sqrt{1+y'^2}dx = L = \text{const.}$$

This forms an isoperimetric variation problem. As $L$ is fixed, thus the maximisation of $A$ is equivalent to making $I = A + \lambda L$ stationary. That is

$$I = A + \lambda L = \int_{x_a}^{x_b} [y + \lambda\sqrt{1+y'^2}]dx.$$

Using the Euler-Lagrange equation, we have

$$\frac{\partial I}{\partial y} - \frac{d}{dx}\frac{\partial I}{\partial y'} = 0,$$

or

$$\frac{\partial}{\partial y}[y + \lambda\sqrt{1+y'^2}] - \frac{d}{dx}\frac{\partial}{\partial y'}[y + \lambda\sqrt{1+y'^2}] = 0,$$

which becomes

$$1 - \lambda\frac{d}{dx}\Big(\frac{y'}{\sqrt{1+y'^2}}\Big) = 0.$$

Integrating it once, we get

$$\frac{\lambda y'}{\sqrt{1+y'^2}} = x + K,$$

where $K$ is the integration constant. By rearranging, we have

$$y' = \pm\frac{x+K}{\sqrt{\lambda^2 - (x+K)^2}}.$$

Integrating this equation again, we get

$$y(x) = \mp\sqrt{\lambda^2 - (x+K)^2} + B,$$

where $B$ is another integration constant. This is equivalent to

$$(x+K)^2 + (y-B)^2 = \lambda^2,$$

which is essentially the standard equation for a circle with the centre at $(-K, B)$ and a radius $\lambda$. Therefore, the greatest area that can be enclosed by a fixed length is a circle.

---

A similar argument leads to the conclusion that the domain with the maximum volume enclosed by a surface with a fixed area is a sphere.

## 6.3 Variations for Multiple Variables

What we have discussed so far mainly concerns the variations in 2-D, and the variations are in terms of $y(x)$ or curves only. What happens if we want to study a surface in the full 3-D configuration? The principle in the previous sections can be extended to any dimensions with multiple variables, however, we will focus on the minimisation of a surface here. Suppose we want to study the shape of a soap bubble, the principle of least action leads to the minimal surface problem. The surface integral of a soap bubble should be stationary. A similar problem is the shape of the Earth under the influence of gravity. If we assume that the shape of the bubble is $u(x, y)$, then the total surface area is

$$A(u) = \iint_\Omega \Psi dxdy = \iint_\Omega \sqrt{1 + (\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2} dxdy, \tag{6.40}$$

where

$$\Psi = \sqrt{1 + (\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2} = \sqrt{1 + u_x^2 + u_y^2}. \tag{6.41}$$

In this case, the extended Euler-Lagrange equation for two variables $x$ and $y$ becomes

$$\frac{\partial \Psi}{\partial u} - \frac{\partial}{\partial x}(\frac{\partial \Psi}{\partial u_x}) - \frac{\partial}{\partial y}(\frac{\partial \Psi}{\partial u_y}) = 0. \tag{6.42}$$

Substituting $\Psi$ into the above equation and using $\frac{\partial \Psi}{\partial u} = \Psi_u = 0$ since $\Psi$ does not contain $u$ explicitly, we get

$$-\frac{\partial}{\partial x}[\frac{1}{\Psi}\frac{\partial u}{\partial x}] - \frac{\partial}{\partial y}[\frac{1}{\Psi}\frac{\partial u}{\partial y}] = 0, \tag{6.43}$$

or

$$(1 + u_y^2)u_{xx} - 2u_x u_y u_{xy} + (1 + u_x^2)u_{yy} = 0. \tag{6.44}$$

This is a nonlinear equation and its solution is beyond the scope of this book. This nonlinear equation has been one of the active research topics for more than a century. It has been proved that the fundamental solution to this equation is a sphere, and in fact we know that all free bubbles are spherical. Similarly, the shape of a planet without rotation should be spherical. However, it becomes a rotational ellipsoid if spin of the planet is included. For some problems, we can approximately assume that $u_x$ and $u_y$ are small, thus the above equation becomes Laplace's equation

$$u_{xx} + u_{yy} = 0. \tag{6.45}$$

The calculus of variations has many applications. The other classical examples include Fermat's principle in optics, the Sturm-Liouville problem, surface shape minimisation, the action principle, and of course finite element analysis.

## 6.4 Integral Equations

From the calculus of variations, we know that the unknown $y(x)$ to be optimised is the integrand of $I$. In a certain sense, this is an integral equation. In fact, many physical processes and laws of conservation are expressed in terms of integral forms rather than their differentiation counterparts. Naturally, one of the ways of constructing an integral equation is to integrate from a differential equation. Integral equations are much more complicated compared with differential equations. There is no universal solution technique for nonlinear equations; even the numerical simulations are usually not straightforward. Thus, we will mainly focus on the simplest types of integral equations.

### 6.4.1 Fredholm Integral Equations

A linear integral equation for $y(x)$ can be written in the following generic form

$$u(x) + \lambda \int_a^b K(x, \eta) y(\eta) d\eta = v(x) y(x), \tag{6.46}$$

where $K(x, \eta)$ is referred to as the kernel of the integral equation. The parameter $\lambda$ is a known constant. If the function $u(x) = 0$, the equation is then called homogeneous. If $u(x) \neq 0$, the equation is inhomogeneous.

If the function $v(x) = 0$, then the unknown $y(x)$ appears only once in the integral equation, and it is under the integral sign only. This is called the linear integral equation of the first kind

$$u(x) + \lambda \int_a^b K(x, \eta) y(\eta) d\eta = 0. \tag{6.47}$$

On the other hand, if $v(x) = 1$, equation (6.46) becomes the integral equation of the second kind

$$u(x) + \lambda \int_a^b K(x, \eta) y(\eta) d\eta = y(x). \tag{6.48}$$

An integral equation with fixed integration limits $a$ and $b$, is called a Fredholm equation. If the upper integration limit $b$ is not fixed, then the

equation becomes a Volterra equation. The integral equation becomes singular if at least one of its integration limits approaches infinite.

### 6.4.2 Volterra Integral Equation

In general, the Volterra integral equation can be written as

$$u(x) + \lambda \int_a^x K(x,\eta) y(\eta) d\eta = v(x) y(x). \tag{6.49}$$

The first kind [or $v(x) = 0$] and second kind [or $v(x) = 1$] are defined in a similar manner to the Fredholm equation.

The kernel is said to be separable or degenerate if it can be written in the finite sum form

$$K(x,\eta) = \sum_{i=1}^{N} f_i(x) g_i(\eta), \tag{6.50}$$

where $f_i(x)$ and $g_i(\eta)$ are functions of $x$ and $\eta$, respectively. A kernel is called a displacement kernel if it can be written as a function of the difference $(x - \eta)$ of its two arguments

$$K(x,\eta) = K(x - \eta). \tag{6.51}$$

## 6.5 Solution of Integral Equations

Most integral equations do not have closed-form solutions. For linear integral equations, closed-form solutions are only possible for the special cases of separable and displacement kernels.

### 6.5.1 Separable Kernels

For a Fredholm integral equation of the second kind with a separable kernel, we can substitute the kernel (6.50) into the equation and we have

$$u(x) + \lambda \int_a^b \sum_{i=1}^{N} f_i(x) g_i(\eta) y(\eta) d\eta = y(x), \tag{6.52}$$

which becomes

$$u(x) + \lambda \sum_{i=1}^{N} f_i(x) \int_a^b g_i(\eta) y(\eta) d\eta = y(x). \tag{6.53}$$

Because the integration limits are fixed, the integrals over $\eta$ are constants that are to be determined. By defining

$$\alpha_i = \int_a^b g_i(\eta) y(\eta) d\eta, \tag{6.54}$$

we now have the solution in the form

$$y(x) = u(x) + \lambda \sum_{i=1}^{N} \alpha_i f_i(x), \tag{6.55}$$

where the $N$ coefficients $\alpha_i$ are determined by

$$\alpha_i = \int_a^b g_i(\eta) u(\eta) d\eta + \lambda \sum_{i=1}^{N} \int_a^b [\alpha_i f_i(\eta) g_i(\eta)] d\eta, \tag{6.56}$$

for $i = 1, 2, ..., N$. Only for a few special cases can these coefficients be written as simple explicit expressions.

### 6.5.2 Volterra Equation

A Volterra equation with separable kernels may be solved by transforming into a differential equation via direct differentiation. In the case of a simple degenerate kernel

$$K(x, \eta) = f(x) g(\eta), \tag{6.57}$$

we have

$$y(x) = u(x) + \lambda \int_0^x f(x) g(\eta) y(\eta) d\eta, \tag{6.58}$$

which becomes

$$y(x) = u(x) + \lambda f(x) \int_0^x g(\eta) y(\eta) d\eta. \tag{6.59}$$

If $f(x) \neq 0$, it can be written as

$$\frac{y(x)}{f(x)} = \frac{u(x)}{f(x)} + \lambda \int_0^x g(\eta) y(\eta) d\eta. \tag{6.60}$$

Putting $\phi(x) = u(x)/f(x)$ and differentiating it, we have

$$[\frac{y(x)}{f(x)}]' = \phi'(x) + \lambda g(x) y(x). \tag{6.61}$$

By letting $\Psi(x) = y(x)/f(x)$, we have

$$\Psi'(x) - \lambda f(x) g(x) \Psi(x) = \phi'(x), \tag{6.62}$$

which is a first-order ordinary differential equation for $\Psi(x)$. This is equivalent to the standard form

$$\Psi' + P(x)\Psi = Q(x), \tag{6.63}$$

where

$$P(x) = -\lambda f(x)g(x), \qquad Q(x) = [\frac{u(x)}{f(x)}]'. \tag{6.64}$$

We can use the standard technique of multiplying by the integrating factor $\exp[\int P(x)dx]$ to obtain the solution. We get

$$y(x) = f(x)[e^{-\int P(x)dx}]\{\int [Q(x)e^{\int P(x)dx}]dx\}. \tag{6.65}$$

With appropriate boundary conditions, the exact form of the solution can be obtained.

---

**Example 6.5:** Let us try to solve an integral equation of Volterra type

$$y(x) = e^x + \int_0^x e^x \sin(\zeta)y(\zeta)d\zeta.$$

First, we divide both sides by $e^x$ to get

$$\frac{y(x)}{e^x} = 1 + \int_0^x \sin(\zeta)y(\zeta)d\zeta,$$

whose differentiation with respect to $x$ leads to

$$[\frac{y(x)}{e^x}]' = y(x)\sin(x), \qquad \text{or} \qquad \frac{1}{e^x}y'(x) - y(x)e^{-x} = y(x)\sin(x).$$

Dividing both sides by $y(x)$ and using $[\ln y(x)]' = y'(x)/y(x)$, we have

$$[\ln y(x)]' = e^x \sin x + 1.$$

By direct integration, we have

$$\ln y(x) = x - \frac{1}{2}e^x \cos x + \frac{1}{2}e^x \sin x.$$

Thus, we finally obtain

$$y(x) = \exp[x - \frac{e^x}{2}(\cos x - \sin x)].$$

---

There are other methods and techniques for solving integral equations such as the operator method, series method and the Fredholm theory. However, most integral equations do not have closed-form solutions. In this case, numerical methods are the best alternative.

## References

Carrier G. F. and Pearson C. E., *Partial Differential Equations: Theory and Technique*, 2nd Edition, Academic Press, (1988).

Courant R. and Hilbert D., *Methods of Mathematical Physics*, 2 volumes, Wiley-Interscience, New York, (1962).

Forsyth A. R., *Calculus of Variations*, Dover, New York, (1960).

Jeffrey A., *Advanced Engineering Mathematics*, Academic Press,(2002).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, Cambridge University Press, (2006).

Smith D. R., *Variational Methods in Optimization*, Dover, New York, (1998).

Wylie C. R., *Advanced Engineering Mathematics*, Tokyo, (1972).

Weinstock R., *Calculus of Variations: with applications to Physics and Engineering*, Dover, New York, (1974).

# Chapter 7

# Probability

All the mathematical models and differential equations we have discussed so far are deterministic in the sense that given accurate initial and boundary conditions, the solutions of the system can be determined (the only exception is to a certain degree the chaotic system). There is no intrinsic randomness in the differential equations. In reality, randomness occurs everywhere, and not all models are deterministic. In fact, it is necessary to use stochastic models and sometimes the only sensible models are stochastic descriptions. In these cases, we have to deal with probability and statistics.

## 7.1 Randomness and Probability

Randomness such as roulette-rolling and noise arises from the lack of information, or incomplete knowledge of reality. It can also come from the intrinsic complexity, diversity and perturbations of the system. The theory of probability is mainly the study of random phenomena so as to find non-random regularity.

For an experiment or trial such as rolling dice whose outcome depends on chance, the sample space $\Omega$ of the experiment is the set of all possible outcomes. The sample space can be either finite or infinite. For example, rolling a six-sided die will have six different outcomes, thus the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The elements of a sample space are the outcomes, and each subset of a sample space is called an event. For example, the event $S = \{2, 4, 6\}$ is a subset of $\Omega$. In a sample space $\Omega$, the outcomes of an experiment are represented as numbers (for example, 1 for heads and 0 for tails for tossing coins).

A real-valued variable that is defined for all the possible outcomes is referred to as a random variable, which is a function that associates

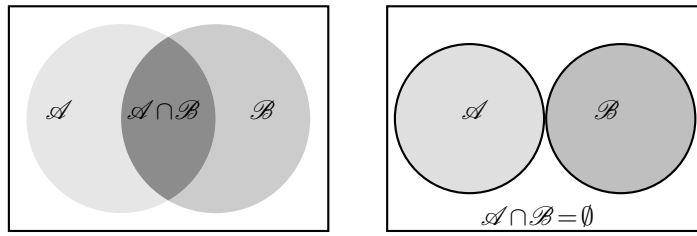



Figure 7.1: Venn diagrams: a) $P(\mathscr{A} \cap \mathscr{B}) = P(\mathscr{A})P(\mathscr{B})$ if $A$ and $B$ are independent, b) $P(\mathscr{A} \cup \mathscr{B}) = P(\mathscr{A}) + P(\mathscr{B})$ if $\mathscr{A} \cap \mathscr{B} = \emptyset$.

a unique numerical value with every outcome of an experiment, and its actual value varies from trial to trial as the experiment is repeated. The values of a random variable can be discrete (such as 1 to 6 in rolling a single die) or continuous (such as the level of noise). If a random variable only takes discrete values, it is called a discrete random variable. If its values are continuous, then it is called a continuous random variable.

Two events $A$ and $B$ can have various relationships and these can be represented by Venn diagrams as shown in Figure 7.1. The intersection $A \cap B$ of two events means the outcome of the random experiments belongs to both $A$ and $B$, and it is the case of '$A$ AND $B$'. If no event or outcome belongs to the intersection, that is $A \cap B = \emptyset$, we say these two events are mutually exclusive or disjoint.

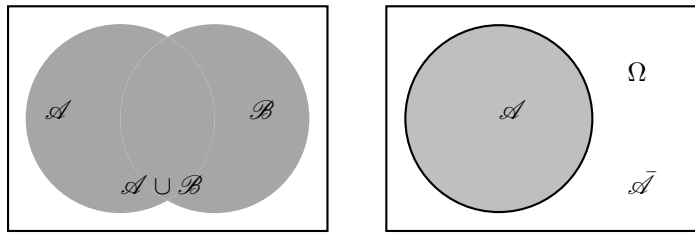
The union $A \cup B$ denotes the outcome belongs to either $A$ or $B$ or both, and this means the case of '$A$ OR $B$'. The complement $\bar{A} = \Omega - A$ (or not $A$) of the event $A$ is the set of outcomes that do not belong to $A$ but in the sample space $\Omega$ (see Figure 7.2). The $A - B$ means the outcomes in $A$ only.

Probability $P$ is a number or an expected frequency assigned to an event $A$ that indicates how likely it is that the event will occur when a random experiment is performed. This probability is often written as $P(A)$ to show that the probability $P$ is associated with event $A$. For a large number of fair trials, the probability can be calculated by

$$P(A) = \frac{N_A(\text{number of outcomes in the event A})}{N_\Omega(\text{total number of outcomes})}. \tag{7.1}$$

---

**Example 7.1:** If you tossed a coin for 1000 times, the head (H) occurred 511 times and the tail (T) occurred 489 times. The estimated probabilities




(a) (b)

Figure 7.2: Venn diagrams: a) $P(\mathscr{A} \cup \mathscr{B}) = P(\mathscr{A}) + P(\mathscr{B}) - P(\mathscr{A} \cap \mathscr{B})$, b) $P(\bar{\mathscr{A}}) = 1 - P(\mathscr{A})$.

$P(H)$ and $P(T)$ are

$$P(H) = \frac{511}{1000} = 0.511,$$

and

$$P(T) = \frac{489}{1000} = 0.489.$$

---

There are three axioms of probability, and they are:

Axiom I : $0 \leq P(A) \leq 1$.

Axiom II : $P(\Omega) = 1$.

Axiom III : $P(A \cup B) = P(A) + P(B)$, if $A \cap B = \emptyset$.

The first axiom says that the probability is a number between 0 and 1 inclusive. $P(A) = 0$ corresponds to impossibility while $P(A) = 1$ corresponds to absolute certainty. The second axiom simply means that an event must occur somewhere inside the sample space. The third axiom is often called the addition rule. Since $A$ and $\bar{A}$ are mutually exclusive ($A \cap \bar{A} = \emptyset$), we have

$$P(A) + P(\bar{A}) = P(A \cup \bar{A}) = P(\Omega) = 1, \tag{7.2}$$

or

$$P(A) = 1 - P(\bar{A}), \tag{7.3}$$

which is usually called the NOT rule. The third axiom can be further generalised to any two events $A$ and $B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{7.4}$$

In a special case when events $A_i (i = 1, 2, ..., n)$ exhaust the whole sample space such that $A = \cup_{i=1}^{n} A_i = A_1 \cup A_i \cup ... \cup A_n = \Omega$ and $A_i \cap A_j = \emptyset$ where $(i \neq j)$,

$$P(A \cap B) = \sum_{i=1}^{n} P(A_i \cap B). \tag{7.5}$$

Since $\Omega \cap B = B$, we also get $P(\Omega \cap B) = P(B) = \sum_{i=1}^{n} P(A_i \cap B)$, which are the useful properties of the total probability.

For example, if you randomly draw a card from a standard pack of 52 cards, what is the probability of it being a red king or a diamond with a face value being a prime number (if its face value is counted from 1 to 13). The prime numbers are 2, 3, 5, 7, 11, 13, therefore there are 6 cards that are primes. The possibility of event (A) of drawing a red king is $P(A) = \frac{2}{52} = \frac{1}{26}$. The probability of event (B) of drawing a prime number is $P(B) = \frac{6}{52} = \frac{3}{26}$. As a diamond king (13) is also a prime, this means $P(A \cap B) = \frac{1}{52}$. Therefore, the probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{26} + \frac{3}{26} - \frac{1}{52} = \frac{7}{52}. \tag{7.6}$$

Two events $A$ and $B$ are independent if the events have no influence on each other. That is to say, the occurrence of one of the events does not provide any information about whether or not the other event will occur. In this case, the probability of both occurring is equal to the product of the probabilities of the two individual events $P(A)$ and $P(B)$

$$P(A \cap B) = P(A) * P(B). \tag{7.7}$$

This can be easily extended to $n$ mutually independent events $A_i (i = 1, 2, ..., n)$. The probability of all these events happening is

$$P(\sum_{i=1}^{n} \cap A_i) = \prod_{i=1}^{n} P(A_i) = P(A_1)P(A_2) \cdots P(A_n). \tag{7.8}$$

---

**Example 7.2:** The probability of drawing a king from a pack of cards (Event A), and showing an even number of rolling a six-sided die (event B) is $P(A \cap B)$. We know $P(A) = 4/52$, and $P(B) = 3/6 = 1/2$. Since these two events are independent, the probability that both events occur is

$$P(A \cap B) = P(A)P(B) = \frac{4}{52} \cdot \frac{1}{2} = \frac{1}{26}.$$

---

If the two events are not independent, then one may affect the other event, in this case, we are dealing with the conditional probability which will be discussed later in the next section.

In calculating the probabilities, it is useful to know the possible combinations and permutations of certain objects. Suppose you have 5 pairs of shoes, 4 pairs of trousers, 7 shirts and 2 hats. This is equivalent to the lineup problem from your feet to your head. In this case, as the act of selecting each thing to wear is similar to putting it into slots in successive stages, the total number of all possible ways is simply the multiplication of all the possible choices for each stages. All possible outfits you can wear form a permutation problem, and the total number is $5 \times 4 \times 7 \times 5 = 700$. In order to line 5 objects marked $A, B, C, D, E$, in the first place, there are 5 possible choices, the second place has only 4 options, the third place 3 choices, the fourth place has 2 choices, and there is only one left for the last place. Thus the number of all possible permutations is $5 \times 4 \times 3 \times 2 \times 1 = 5!$. Following this line of reasoning, $n$ objects can in general be permutated in $n!$ ways.

Suppose there are $n = 20$ students in a class (named $S_1$, $S_2$, ...., $S_{20}$), we want to select 5 students at random to form a 5-student team to do some field work. This is different from the lineup problem because once you have selected any five students (say) $S_1$, $S_7$, $S_{10}$, $S_{15}$, $S_{19}$, it does not matter in what order you selected them, the final formed team is the same. There are 5! permutations within the same team. Order does not count in this case. This is a combination problem (also called a committee problem). As before, there are 5 places to line up the students, and the total number of all permutations for selecting 5 students is $20 * 19 * 18 * 17 * 16$. Therefore, the total number of combinations (of selecting 5 students) is

$$^{20}C_5 = \frac{20 * 19 * 18 * 17 * 15}{5!} = \frac{20!}{5!15!} = 15504. \tag{7.9}$$

In general, the total number of all possible combinations of selecting $k$ objects from $n$ is

$$^nC_k \equiv \binom{n}{k} \equiv \frac{n!}{k!(n-k)!}. \tag{7.10}$$

The consistency requires $0! = 1$.

---

**Example 7.3:** A research team of 5 members is chosen at random from 8 female students, 10 male students, and 7 professors. What is the probability of the team consisting of 2 female students, 2 male students, and 1 professor? The total number of possible teams is $^{25}C_5$. If two female students are selected, we have $^8C_2$. Similarly, $^{10}C_2$ for selecting 2 male

students, and ${}^7C_1$ for selecting one professor. Therefore, we have

$$N = \frac{{}^8C_2 \, {}^{10}C_2 \, {}^7C_1}{{}^{25}C_5} = \frac{42}{253} \approx 0.166.$$

---

There is an interesting 'birthday paradox' which is related to this context. The birthday paradox was first proposed in 1939 by Richard von Mises, which asks what is the probability of two people having the same birthday in a group of $n$ people. For a group of 367 people, it is certain that there must be at least two people having the same birthday as there are only 365 (or 366 if someone was born in a leap year) possible birthdays. Ignoring 29 February and the year of birth and assuming that the birthdays are evenly distributed throughout the year, we only have 365 different birthdays (days and months only). If the event $A$ denotes that all the $n$ people will have different birthdays (no birthday matching), the first person can have any date as his or her birthday, 365/365. The second person must be in other 364 dates, which is 364/365, and the $k$th person has $(365-k+1)/365$. Therefore, the probability of no two people having the same birthday is

$$P(A,n) = \frac{365}{365} \times \frac{364}{365} \times ... \times \frac{(365-n+1)}{365}$$

$$= \frac{365 * (364) * ... * (365-n+1)}{365^n} = \frac{365!}{(365-n)!365^n}. \tag{7.11}$$

Now the probability of two people with the same birthday is

$$P(\bar{A},n) = 1 - P(A,n) = 1 - \frac{365!}{(365-n)!365^n}. \tag{7.12}$$

The factorial 365! is a large number, but you do not have to deal with such large numbers. You can use a simple calculator to estimate it. For five people, the probability of two people with the same birthday is

$$P(\bar{A},5) = 1 - \frac{365 * 364 * 363 * 362 * 361}{365^5} \approx 0.027, \tag{7.13}$$

which seems insignificant. However, the interesting thing is that for $n = 23$, the probability becomes

$$P(\bar{A},23) = 1 - \frac{365!}{(365-23)!365^{23}} \approx 0.507. \tag{7.14}$$

This means that you have slightly more than a 50-50 chance of finding two people sharing the same birthday. If you increase $n$, you

get $P(\bar{A}, 30) \approx 0.706$ for $n = 30$, $P(\bar{A}, 40) \approx 0.891$ for $n = 40$, and $P(\bar{A}, 50) \approx 0.970$ and $P(\bar{A}, 70) \approx 0.9992$ (almost certainty) for $n = 70$.

Another issue concerning probability is that there is some difference in combinations when the member drawn is placed back or not. Suppose there are 10 red balls and 10 white balls in a bag. If we draw a ball (say a red, event A) from the bag and then put it back (with replacement), then we draw another ball (event B). $P(A) = 1/20$ and $P(B) = 1/20$. The probability of getting two red balls is $P(A \cap B) = P(A) * P(B) = 1/400$. We call this case I.

For a second case (Case II), if we do not put it back after we have drawn the first ball (without replacement), then the probability of event B is now different $P(B) = 1/19$ as there are now only 19 balls in the bag. The probability of getting two red balls now becomes $P(A \cap B) = \frac{1}{20} \times \frac{1}{19} = \frac{1}{380}$, which is different from $1/400$.

The reason here is that the two events are not independent in the case of no-replacement. If we use notation '$B|A$' which means that event B occurs given that event A has occurred, then we can use $P(B|A)$ to denote the probability of event B when there is no replacement in event A in the scenario described in Case II. Now $P(B)$ becomes $P(B|A)$. Hence, we have

$$P(A \cap B) = P(A)P(B|A), \tag{7.15}$$

which is often called the multiplication rule in probability theory. Similarly, we can get

$$P(A \cap B) = P(B)P(A|B). \tag{7.16}$$

This is essentially a conditional probability problem which forms the main topic of the next section.

## 7.2 Conditional Probability

In calculating the probabilities, we often assume that all possible outcomes of an experiment such as drawing a card are equally likely. Probabilities can change if additional information is known or some other event has already occurred and thus $P(B|A)$ denotes the probability that event B will occur given that event A has already occurred. The conditional probability can be calculated by

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{7.17}$$

Conversely, we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{7.18}$$

Using Eq.(7.15), we can write the above formulation as

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})}, \quad (7.19)$$

which is the well-known Bayes theorem. Here we have used $\bar{A} \cup A = \Omega$ and $P(\bar{A}) = 1 - P(A)$. As an example, we consider the drug test in sports.

---

**Example 7.4:** It is believed that the test is 99% accurate if athletes are taking drugs. For athletes not taking drugs, the positive test is only 0.5%. It is assumed that only one in 1000 athletes takes this kind of drug. Suppose an athlete is selected at random and the test shows positive for the drug. What is the probability that the athlete is really taking the drug? Event $A$ denotes an athlete is taking the drug, and $B$ denotes the event that the individual tests positive. Thus, $P(A) = 1/1000$, $P(B|A) = 0.99$ and $P(B|\bar{A}) = 0.005$. The probability that the athlete is actually taking the drug is

$$\begin{aligned} P(A|B) &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})} \\ &= \frac{0.001 * 0.99}{0.001 * 0.99 + 0.999 * 0.005} \approx 0.165. \end{aligned} \quad (7.20)$$

This is a surprisingly low probability.

---

## 7.3 Random Variables and Moments

### 7.3.1 Random Variables

For a discrete random variable $X$ with distinct values such as the number of cars passing through a junction or the number of major earthquakes in a year, each value $x_i$ may occur with a certain probability $p(x_i)$. In other words, the probability varies with the random variable. A probability function $p(x_i)$ is a function that defines probabilities to all the discrete values $x_i$ of the random variable $X$. As an event must occur inside a sample space, the requirement that all the probabilities must be summed to one leads to

$$\sum_{i=1}^{n} p(x_i) = 1. \quad (7.21)$$

The cumulative probability function of $X$ is defined by

$$P(X \leq x) = \sum_{x_i < x} p(x_i). \quad (7.22)$$

For a continuous random variable $X$ that takes a continuous range of values (such as the level of noise), its distribution is continuous and the probability density function $p(x)$ is defined for a range of values $x \in [a, b]$ for given limits $a$ and $b$ [or even over the whole real axis $x \in (-\infty, \infty)$]. In this case, we always use the interval $(x, x + dx]$ so that $p(x)$ is the probability that the random variable $X$ takes the value $x < X \leq x + dx$ is

$$\Phi(x) = P(x < X \leq x + dx) = p(x)dx. \tag{7.23}$$

As all the probabilities of the distribution shall be added to unity, we have

$$\int_a^b p(x)dx = 1. \tag{7.24}$$

The cumulative probability function becomes

$$\Phi(x) = P(X \leq x) = \int_a^x p(x)dx, \tag{7.25}$$

which is the definite integral of the probability density function between the lower limit $a$ up to the present value $X = x$.

### 7.3.2 Mean and Variance

Two main measures for a random variable $X$ with a given probability distribution $p(x)$ are its mean and variance. The mean $\mu$ or the expectation value of $E[X]$ is defined by

$$\mu \equiv E[X] \equiv <X> = \int xp(x)dx, \tag{7.26}$$

for a continuous distribution and the integration is within the integration limits. If the random variable is discrete, then the integration becomes the summation

$$E[X] = \sum_i x_i p(x_i). \tag{7.27}$$

The variance $\text{var}[X] = \sigma^2$ is the expectation value of the deviation squared $(X - \mu)^2$. That is

$$\sigma^2 \equiv \text{var}[X] = E[(X - \mu)^2] = \int (x - \mu)^2 p(x)dx. \tag{7.28}$$

The square root of the variance $\sigma = \sqrt{\text{var}[X]}$ is called the standard deviation, which is denoted by the symbol $\sigma$.

The variance becomes a simple sum

$$\sigma^2 = \sum_i (x-\mu)^2 p(x_i), \tag{7.29}$$

for a discrete distribution. In addition, any other formulae for a continuous distribution can be converted to their counterpart for a discrete distribution if the integration is replaced by the sum. Therefore, we will mainly focus on the continuous distribution in the rest of the section.

Other frequently used measures are the mode and median. The mode of a distribution is defined by the value at which the probability density function $p(x)$ is maximum. For an even number of data sets, the mode may have two values. The median $m$ of a distribution corresponds to the value at which the cumulative probability function $\Phi(m) = 1/2$. The upper and lower quartiles $Q_U$ and $Q_L$ are defined by $\Phi(Q_U) = 3/4$ and $\Phi(Q_L) = 1/4$.

### 7.3.3 Moments and Generating Functions

In fact, the mean is essentially the first moment if we define the $k$th moment of a random variable $X$ by

$$E[X^k] \equiv \mu_k = \int x^k p(x) dx, \qquad k = 1, 2, ..., N. \tag{7.30}$$

Similarly, the $k$th central moment is defined by

$$E[(X-\mu)^k] \equiv \nu_k = \int (x-\mu)^k p(x) dx, \qquad k = 1, 2, ..., N. \tag{7.31}$$

Obviously, the variance is the second central moment. From these definitions, it is straightforward to prove that

$$E[\alpha x + \beta] = \alpha E[X] + \beta, \qquad E[X^2] = \mu^2 + \sigma^2, \tag{7.32}$$

and

$$\text{var}[\alpha x + \beta] = \alpha^2 \text{var}[X]. \tag{7.33}$$

where $\alpha$ and $\beta$ are constants.

Most probability functions can be expressed in terms of moments and moment generating functions. The moment generating function is defined by

$$G_X(\nu) \equiv E[e^{\nu X}] = \int e^{\nu x} p(x) dx, \tag{7.34}$$

where $\nu \in \mathcal{R}$ is a real parameter. By expanding $\exp[\nu x]$ into power series and using the definition of various moments, it is straightforward to verify that

$$E[X^k] = \frac{d^k G_X(\nu)}{d\nu^k}\Big|_{\nu=0}, \tag{7.35}$$

and

$$\sigma^2 = \frac{d^2 G_X(0)}{d\nu^2} - [\frac{dG_X(0)}{d\nu}]^2. \tag{7.36}$$

## 7.4 Binomial and Poisson Distributions

### 7.4.1 Binomial Distribution

A discrete random variable is said to follow the binomial distribution $B(n,p)$ if its probability distribution is given by

$$B(n,p) = {}^nC_x p^x (1-p)^{n-x}, \qquad {}^nC_x = \frac{n!}{x!(n-x)!}, \tag{7.37}$$

where $x = 0, 1, 2, ..., n$ are the values that the random variable $X$ may take, $n$ is the number of trials. There are only two possible outcomes: success or failure. $p$ is the probability of a so-called 'success' of the outcome. Subsequently, the probability of the failure of a trial is $q = 1 - p$. Therefore, $B(n,p)$ represents the probability of $x$ successes and $n - x$ failures in $n$ trials. The coefficients come from the coefficients of the binomial expansions

$$(p+q)^n = \sum_{x=0}^{n} {}^nC_x p^x q^{n-x} = 1, \tag{7.38}$$

which is exactly the requirement that all the probabilities should be summed to unity.

---

**Example 7.5:** Tossing a coin 10 times, the probability of getting 7 heads is $B(n, 1/2)$. Since $p = 1/2$ and $x = 7$, then we have

$${}^{10}C_7(\frac{1}{2})^7(\frac{1}{2})^3 = \frac{15}{128},$$

which is about 0.117.

---

It is straightforward to prove that $\mu = E[X] = np$ and $\sigma^2 = npq = np(1-p)$ for a binomial distribution.

Another related distribution is the geometric distribution whose probability function is defined by

$$P(X = n) = pq^{n-1} = p(1-p)^{n-1}, \tag{7.39}$$

where $n \geq 1$. This distribution is used to calculate the first success, thus the first $n - 1$ trials must be a failure if $n$ trials are needed to observe the first success. The mean and variance of this distribution are $\mu = 1/p$ and $\sigma^2 = (1-p)/p^2$.

### 7.4.2 Poisson Distribution

The Poisson distribution can be thought of as the limit of the binomial distribution when the number of trial is very large $n \to \infty$ and the probability $p \to 0$ (small probability) with the constraint that $\lambda = np$ is finite. For this reason, it is often called the distribution for small-probability events. Typically, it is concerned with the number of events that occur in a certain time interval (e.g., number of telephone calls in an hour) or spatial area. The Poisson distribution is

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad \lambda > 0, \tag{7.40}$$

where $x = 0, 1, 2, ..., n$ and $\lambda$ is the mean of the distribution. Using the definition of mean and variance, it is straightforward to prove that $\mu = \lambda$ and $\sigma^2 = \lambda$ for the Poisson distribution. The parameter $\lambda$ is the location of the peak as shown in Figure 7.3.

---

**Example 7.6:** On average, there are about 2 major earthquakes ($\geq 7.0$ on the Richter scale) that occur somewhere in the world every year. This estimation is based on the fact that there were 235 major earthquakes between 1901 and 2007 recorded by US Geological Survey and others. What is the probability of no major earthquake occurring during a year? What is the probability of at least one major earthquake in a year? We know that $\lambda \approx 2$. The probability of no major earthquake is

$$P(X = 0) = \frac{2^0 e^{-2}}{0!} \approx 0.13533.$$

Thus, the probability of at least one major earthquake is $P(X > 0) \approx 1 - 0.13533 \approx 0.864$. In fact, the probability of having one earthquake is

$$P(X = 1) = \frac{2^1 e^{-2}}{1!} \approx 0.2707,$$

and the probability of two major earthquakes is

$$P(X = 2) = \frac{2^2 e^{-2}}{2!} \approx 0.2707.$$

However, the probability of 12 major earthquakes (occurring almost every month) is

$$P(X = 12) = \frac{2^{12} e^{-2}}{12!} \approx 0.00000115,$$

Which should be very rare in this case.

---

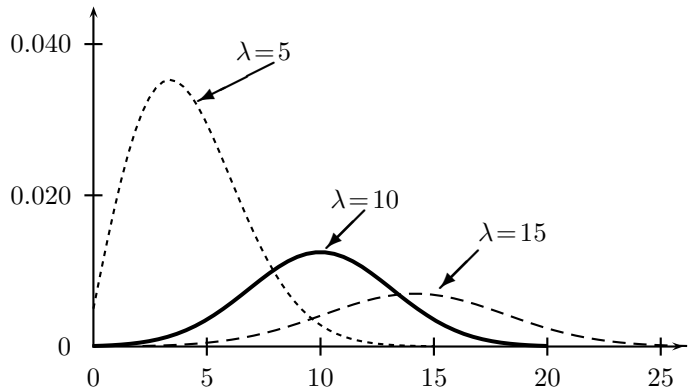


Figure 7.3: Poisson distributions for different values of $\lambda = 5, 10, 15$.

The moment generating function for the Poisson distribution is given by

$$G_X(\nu) = \sum_{x=0}^{\infty} \frac{e^{\nu x}\lambda^x e^{-\lambda}}{x!} = \exp[\lambda(e^{\nu} - 1)]. \tag{7.41}$$

## 7.5 Gaussian Distribution

The Gaussian distribution or normal distribution is the most important continuous distribution in probability and it has a wide range of applications. For a continuous random variable $X$, the probability density function (PDF) of a Gaussian distribution is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{7.42}$$

where $\sigma^2 = \text{var}[X]$ is the variance and $\mu = E[X]$ is the mean of the Gaussian distribution. From the Gaussian integral, it is easy to verify that

$$\int_{\infty}^{\infty} p(x)dx = 1, \tag{7.43}$$

and this is exactly the reason that the factor $1/\sqrt{2\pi}$ comes from the normalisation of all the probabilities. The probability function reaches a peak at $x = \mu$ and the variance $\sigma^2$ controls the width of the peak (see Figure 7.4).

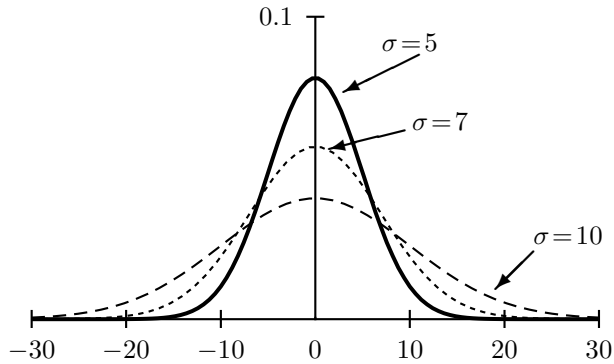


Figure 7.4: Gaussian distributions for $\sigma = 5, 7, 10$.

The cumulative probability function (CPF) for a normal distribution is the integral of $p(x)$, which is defined by

$$\Phi(x) = P(X < x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(\zeta-\mu)^2}{2\sigma^2}} d\zeta. \tag{7.44}$$

Using the error function defined by Chapter 1, we can write it as

$$\Phi(x) = \frac{1}{\sqrt{2}}[1 + \mathrm{erf}(\frac{x-\mu}{\sqrt{2}\sigma})]. \tag{7.45}$$

The moment generating function for the Gaussian distribution is given by

$$G_X(\nu) = e^{\mu\nu + \frac{1}{2}(\sigma\nu)^2}. \tag{7.46}$$

The Gaussian distribution can be considered as the limit of the Poisson distribution when $\lambda \gg 1$. Using the Sterling's approximation $x! \sim \sqrt{2\pi x}(x/e)^x$ for $x \gg 1$, and setting $\mu = \lambda$ and $\sigma^2 = \lambda$, it can be verified that the Poisson distribution can be written as a Gaussian distribution

$$P(x) \approx \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(x-\mu)^2}{2\lambda}}, \tag{7.47}$$

where $\mu = \lambda$. In statistical applications, the normal distribution is often written as $N(\mu, \sigma)$ to emphasise that the probability density function depends on two parameters $\mu$ and $\sigma$.

The standard normal distribution is a normal distribution $N(\mu, \sigma)$ with a mean of $\mu = 0$ and standard deviation $\sigma = 1$, that is $N(0, 1)$. This is useful to normalise or standardise data for statistical analysis. If we define a normalised variable

$$\xi = \frac{x - \mu}{\sigma}, \tag{7.48}$$

it is equivalent to give a score so as to place the data above or below the mean in the unit of standard deviation. In terms of the area under the probability density function, $\xi$ sorts where the data falls. It is worth pointing out that some books define $z = \xi = (x - \mu)/\sigma$ in this case, and call the standard normal distribution the Z distribution.

Table 7.1: Function $\phi$ defined by Eq.(7.50).

| $\xi$ | $\phi(\xi)$ | $\xi$ | $\phi$ |
|---|---|---|---|
| 0.0 | 0.500 | 1.0 | 0.841 |
| 0.1 | 0.540 | 1.1 | 0.864 |
| 0.2 | 0.579 | 1.2 | 0.885 |
| 0.3 | 0.618 | 1.3 | 0.903 |
| 0.4 | 0.655 | 1.4 | 0.919 |
| 0.5 | 0.692 | 1.5 | 0.933 |
| 0.6 | 0.726 | 1.6 | 0.945 |
| 0.7 | 0.758 | 1.7 | 0.955 |
| 0.8 | 0.788 | 1.8 | 0.964 |
| 0.9 | 0.816 | 1.9 | 0.971 |

Now the probability density function of standard normal distribution becomes

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}. \tag{7.49}$$

Its cumulative probability function is

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-\frac{\xi^2}{2}} d\xi = \frac{1}{2}[1 + \mathrm{erf}(\frac{\xi}{\sqrt{2}})]. \tag{7.50}$$

As the calculations of $\phi$ and the error function involve numerical integration, it is usual in practice to tabulate $\phi$ in a table (see Table 7.1) so that you do not have to calculate their values each time you use it.

## 7.6 Other Distributions

There are a number of other important distributions such as the exponential distribution, log-normal distribution, uniform distribution and the $\chi^2$-distribution. The uniform distribution has a probability density function

$$p = \frac{1}{\beta - \alpha}, \qquad x = [\alpha, \beta], \tag{7.51}$$

whose mean is $E[X] = (\alpha + \beta)/2$ and variance is $\sigma^2 = (\beta - \alpha)^2/12$.

The exponential distribution has the following probability density function

$$f(x) = \lambda e^{-\lambda x} \quad (x > 0), \tag{7.52}$$

and $f(x) = 0$ for $x \leq 0$. Its mean and variance are

$$\mu = 1/\lambda, \qquad \sigma^2 = 1/\lambda^2. \tag{7.53}$$

The log-normal distribution has a probability density function

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp[-\frac{(\ln x - \mu)^2}{2\sigma^2}], \tag{7.54}$$

whose mean and variance are

$$E[X] = e^{\mu+\sigma^2/2}, \qquad \text{var}[X] = e^{\sigma^2+2\mu}(e^{\sigma^2} - 1). \tag{7.55}$$

The $\chi^2$-distribution, called chi-square or chi-squared distribution, is very useful in statistical inference and the method of least squares. This distribution is for the quantity

$$\chi_n^2 = \sum_{i=1}^{n} (\frac{X_i - \mu_i}{\sigma_i})^2, \tag{7.56}$$

where the $n$-independent variables $X_i$ are normally distributed with means $\mu_i$ and variances $\sigma_i^2$. The probability density function for $\chi^2$-distribution is given by

$$p(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-x/2}, \tag{7.57}$$

where $x \geq 0$, and $n$ is called the degree of freedom.It can be verified that the mean of the distribution is $n$ and its variance is $2n$.

Here the $\Gamma$ function is given by

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx, \qquad (n > 0). \tag{7.58}$$

For $n$ is a positive integer, we have $\Gamma(n+1) = n!$ and $\Gamma(1) = 1$. However, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

For other distributions, readers can refer to more advanced books that are devoted to probability theory and statistical analysis.

## 7.7 The Central Limit Theorem

The most important theorem in probability is the central limit theorem which concerns a large number of trials and explains why the normal

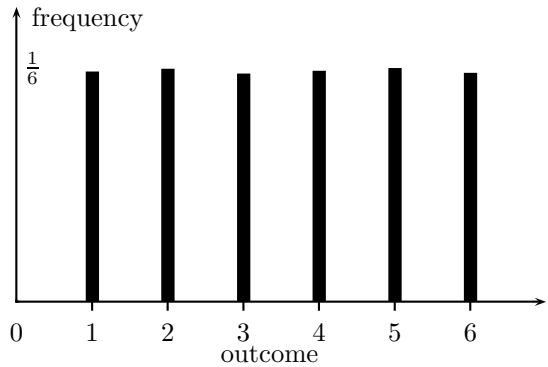



Figure 7.5: A uniform distribution.

distribution occurs so widely. This theorem is as follows: Let $X_i (i = 1, 2, ..., n)$ be $n$ independent random variables, each of which is defined by a probability density function $p_i(x)$ with a corresponding mean $\mu_i$ and a variance $\sigma_i^2$. The sum of all these random variables

$$\Theta = \sum_{i=1}^{n} X_i = X_1 + X_2 + ... + X_n, \tag{7.59}$$

is also a random variable whose distribution approaches the Gaussian distribution as $n \to \infty$. Its mean $E[\Theta]$ and variance $\text{var}[\Theta]$ are given by

$$E[\Theta] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu_i, \tag{7.60}$$

and

$$\text{var}[\Theta] = \sum_{i=1}^{n} \text{var}[\Theta] = \sum_{i=1}^{n} \sigma_i^2. \tag{7.61}$$

The proof of this theorem is beyond the scope of this book as it involves the moment generating functions, characteristics functions and other techniques. In geostatistics, we simply use these important results for statistical analysis.

In the special case when all the variables $X_i$ are described by the same probability density function with the same mean $\mu$ and variance $\sigma^2$, these results become

$$E[\Theta] = n\mu, \qquad \text{var}[\Theta] = n\sigma^2. \tag{7.62}$$

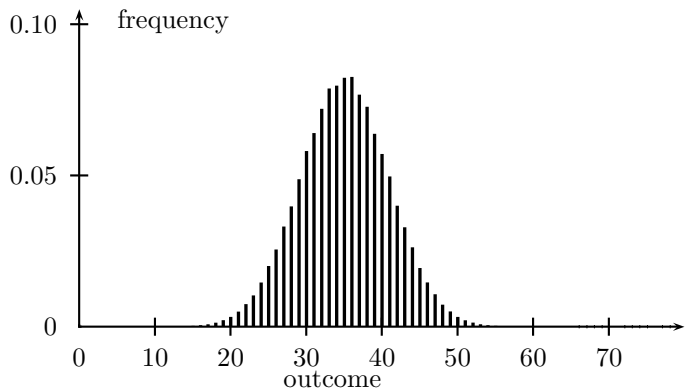



Figure 7.6: An approximate Gaussian distribution (the outcomes of the sum of face values in rolling 15 dice).

By defining a new variable

$$\xi_n = \frac{\Theta - n\mu}{\sigma\sqrt{n}}, \tag{7.63}$$

then the distribution of $\xi_n$ converges towards the standard normal distribution $N(0,1)$ as $n \to \infty$.

Let us see what the theorem means for a simple experiment of rolling a few dice. For a fair six-sided die, each side will appear equally likely with a probability of $1/6 \approx 0.1667$, thus the probability function after rolling it, say, 15,000 times approaches a uniform distribution as shown in Figure 7.5. If we roll $n = 15$ independent dice, the sums of the face values vary from 1 to 90. After rolling the 15 dice 10,000 times, the distribution is shown in Figure 7.6 and it approaches to a normal distribution as $n \to \infty$.

## 7.8 Weibull Distribution

Although the distribution functions of the real-world random processes are dominated by the Gaussian or normal distribution, however, there are some cases where other distributions can describe the related phenomena more accurately. Weibull's distribution is such a distribution with many applications in areas such as reliability analysis, engineering design and earth sciences. Therefore, it deserves a special introduction in detail. This distribution was originally developed by Swedish physicist, A. Weibull in 1939 to try to explain the fact, well-known but unexplained at that time, that the relative strength of a specimen decreases

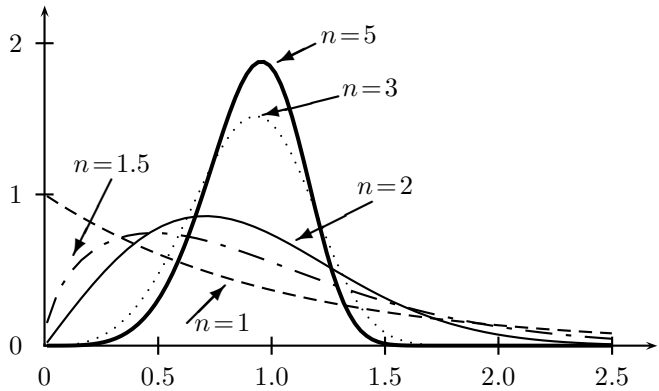



Figure 7.7: Weibull density function for various-shaped parameters.

with increasing dimension. Since then, it has been applied to study many real-world stochastic processes even including the distributions of wind speed, rainfall, energy resources and earthquakes.

Weibull's distribution is a three-parameter distribution given by

$$p(x,\lambda,\beta,n) = \begin{cases} \frac{n}{\lambda}(\frac{x-\beta}{\lambda})^{n-1}\exp[-(\frac{x-\beta}{\lambda})^n] & (x \geq \beta) \\ 0 & (x < \beta) \end{cases}, \tag{7.64}$$

where $\lambda$ is scaling parameter, and $n$ is the shape parameter, often referred to as the Weibull modulus. The parameter $\beta$ is the threshold of the distribution. By straightforward integration, we have the cumulative probability density distribution

$$\Phi(x,\lambda,\beta,n) = 1 - e^{-(\frac{x-\beta}{\lambda})^n}. \tag{7.65}$$

For the fixed values $\lambda = 1$ and $\beta = 0$, the variation of $n$ will give a broad range of shapes and can be used to approximate various distributions as shown in Fig. 7.7.

In reliability analysis, especially for a large infrastructure such as a dam or a tall building or an underground mining tunnel under stress, the survival probability is more conveniently represented as

$$P_s(V) = \exp[\int_V -(\frac{\sigma}{\sigma_0})^n \frac{dV}{V_0}], \tag{7.66}$$

where $V$ is the volume of the system. $\sigma_0$ is the failure stress (either tensile or shear) for the reference volume $V_0$. The failure probability is

$$P_f(V) = 1 - P_s(V). \tag{7.67}$$

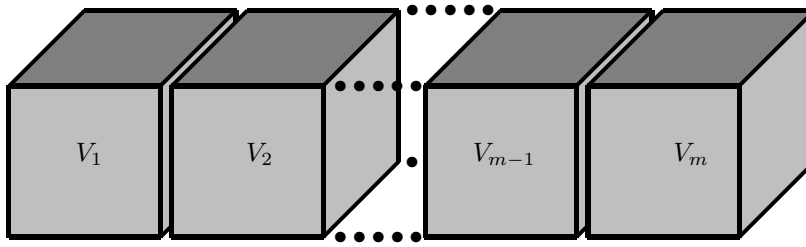



Figure 7.8: Weakest link analogy.

For constant stress $\sigma$ over the whole volume $V$, we simply have

$$P_s(V) = \exp[-(\frac{\sigma}{\sigma_0})^n \frac{V}{V_0}]. \tag{7.68}$$

At the reference point $\sigma = \sigma_0$ and $V = V_0$ often obtained using laboratory tests, we have $P_s(V_0) = e^{-1} \approx 0.3679$. As the stress becomes extreme, $\sigma \to \infty$, then $P_s \to 0$ and $P_f \to 1$.

The fundamental idea of this volume-related probability function is the weakest link theory. The larger the volume of a system, the more likely it is to have a critical flaw that causes potential failure. We can consider that the whole large volume $V$ consists of $m$ small volumes/blocks $V_1, V_2, ..., V_m$ and these small blocks are glued together (see Fig. 7.8), thus the probability of survival of the whole system is equivalent to the survival of all the subsystem blocks. If any one of these blocks fails, the system is considered flawed and thus failed. In the simplest case, $V_1 = V_2 = ... = V_m = V_0$ and $m = V/V_0$, the survival probability of the whole system under constant stress $\sigma$ is

$$\begin{aligned} P_s(V) = P_s(mV_0) &= \overbrace{P_s(V_0) \times P_s(V_0) \times ... \times P_s(V_0)}^{m} \\ &= [P_s(V_0)]^m = [e^{-(\frac{\sigma}{\sigma_0})^n}]^{\frac{V}{V_0}} \\ &= \exp[-\frac{V}{V_0}(\frac{\sigma}{\sigma_0})^n]. \end{aligned} \tag{7.69}$$

---

**Example 7.7:** Stalactites are tapering calcite needles hanging down from the roof of caves. If their length is continuously increasing, there is a danger that they will fall down due to failure under their own weight. In order to model such a stalactitic system, we can idealise it as a cone with the base diameter $d$ and length $h$ as shown in Fig. 7.9.

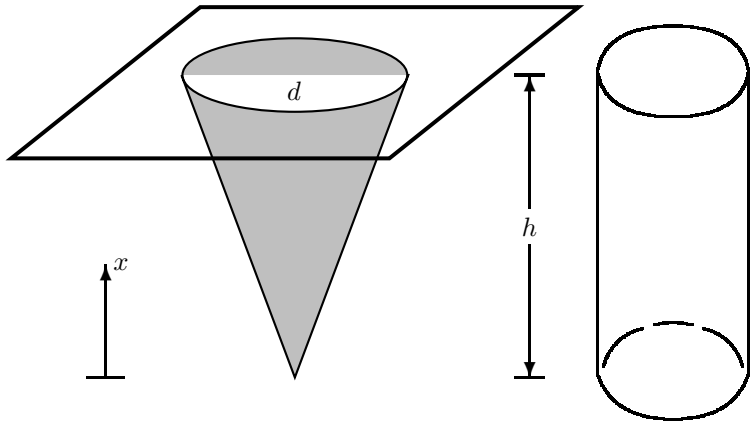



Figure 7.9: Idealisation of a stalactitic system as a cone with a base diameter of $d$ and height $h$.

**Model I:** First, let us use a cylinder model to approximate the shape. Then, the area of the cross section is uniform $A = \pi d^2/4$, and the weight under the cross section at $x$ is simply

$$W = \rho g A x = \frac{\rho g \pi d^2}{4} x,$$

where $g$ is the acceleration due to gravity, and $\rho$ is the average density. Thus the stress at $x$ is

$$\sigma = \frac{W}{A} = \rho g x.$$

Therefore, the survival probability of the stalactitic system is

$$P_s^{(I)}(V) = \exp[-\int_0^h (\frac{\sigma}{\sigma_0})^n \frac{A dx}{V_0}] = \exp[-(\frac{\rho g}{\sigma_0})^n \frac{\pi d^2}{4V_0} \int_0^h x^n dx]$$

$$= \exp[-(\frac{\rho g}{\sigma_0})^n \frac{\pi d^2 h^{n+1}}{4(n+1)V_0}],$$

which provides a lower limit of the probability as we overestimate the total volume (and thus the stress).

**Model II:** If we use a more realistic cone as the model, then the area of the cross section at any value of $x$ is $A = \pi(xd/2h)^2$. The weight under the cross section at $x$ is $W = \rho g \int_0^x A dx = \frac{\rho g \pi d^2}{12h^2} x^3$, so that the stress at $x$ is

$$\sigma = \frac{W}{A} = \frac{1}{3} \rho g x.$$

The survival probability of this system is

$$P_s^{(II)}(V) = e^{-(\frac{\rho g}{3\sigma_0})^n \frac{\pi d^2}{4h^2 V_0} \int_0^h x^{n+2} dx} = e^{-(\frac{\rho g}{3\sigma_0})^n \frac{\pi d^2 h^{n+1}}{4(n+3)V_0}},$$

which provides an upper limit of the probability. The shape of a real stalactitic tapering is more complicated. However, the probability decreases with increasing $h$ due to the fact that the volume increases with $h$, and a larger volume means a higher probability of containing critical flaws, and thus a higher probability of failure.

# References


Armstrong M., *Basic Linear Geostatistics*, Springer (1998).

Gardiner C. W., *Handbook of Stochastic Methods*, Springer, (2004).

Goodman R., *Teach Yourself Statistics*, London, (1957).

Kitanidis P. K., *Introduction to Geostatistics*, Cambridge University Press, (1997).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Papoulis A., *Probability and Statistics*, Englewood Cliffs, (1990).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, Cambridge University Press (2006).

Weibull W., A statistical distribution function of wide applicability, *J. Appl. Mech.*, **18**, 293-297 (1951).

Wylie C. R., *Advanced Engineering Mathematics*, Tokyo, (1972).

# Chapter 8

# Geostatistics

Statistics is the mathematics of data collection and interpretation, and the analysis and characterisation of numerical data by inference from sampling. Statistical methods involve reduction of data, estimates and significance tests, and relationship between two or more variables by analysis of variance, and the test of hypotheses. Geostatistics uses various statistical methods to analyse data in geosciences, though specialised tools such as kriging have been further developed to suitable specific applications.

## 8.1 Sample Mean and Variance

If a sample consists of $n$ independent observations $x_1, x_2, ..., x_n$ on a random variable $x$ such as variations in gold and oil prices, two important and commonly used parameters are sample mean and sample variance, which can easily be estimated from the sample. The sample mean is calculated by

$$\bar{x} \equiv <x> = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{8.1}$$

which is essentially the arithmetic average of the values $x_i$.

Generally speaking, if $u$ is a linear combination of $n$ independent random variables $y_1, y_2, ..., y_n$ and each random variable $y_i$ has an individual mean $\mu_i$ and a corresponding variance $\sigma_i^2$, we have the linear combination

$$u = \sum_{i=1}^{n} \alpha_i y_i = \alpha_1 y_1 + \alpha_2 y_2 + ... + \alpha_n y_n, \tag{8.2}$$

where the parameters $\alpha_i (i = 1, 2, ..., n)$ are the weighting coefficients. From the central limit theorem, we have the mean $\mu_u$ of the linear combination

$$\mu_u = E(u) = E(\sum_{i=1}^{n} \alpha_i y_i) = \sum_{i=1}^{n} \alpha E(y_i) = \sum \alpha_i \mu_i. \tag{8.3}$$

Then, the variance $\sigma_u^2$ of the combination is

$$\sigma_u^2 = E[(u - \mu_u)^2] = E\Big[\sum_{i=1}^{n} \alpha_i (y_i - \mu_i)^2\Big], \tag{8.4}$$

which can be expanded as

$$\sigma_u^2 = \sum_{i=1}^{n} \alpha_i^2 E[(y_i - \mu_i)^2] + \sum_{i,j=1;i\neq j}^{n} \alpha_i \alpha_j E[(y_i - \mu_i)(y_j - \mu_j)], \tag{8.5}$$

where $E[(y_i - \mu_i)^2] = \sigma_i^2$. Since $y_i$ and $y_j$ are independent, we have $E[(y_i - \mu_i)(y_j - \mu_j)] = E[(y_i - \mu_i)]E[(y_j - \mu_j)] = 0$. Therefore, we get

$$\sigma_u^2 = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2. \tag{8.6}$$

The sample mean defined in Eq.(8.1) can also be viewed as a linear combination of all the $x_i$ assuming each of which has the same mean $\mu_i = \mu$ and variance $\sigma_i^2 = \sigma^2$, and the same weighting coefficient $\alpha_i = 1/n$. Hence, the sample mean is an unbiased estimate of the sample due to the fact $\mu_{\bar{x}} = \sum_{i=1}^{n} \mu/n = \mu$. In this case, however, we have the variance

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{n} \frac{1}{n^2} \sigma^2 = \frac{\sigma^2}{n}, \tag{8.7}$$

which means the variance becomes smaller as the size $n$ of the sample increases by a factor of $1/n$.

The sample variance $S^2$ is defined by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{8.8}$$

It is worth pointing out that the factor is $1/(n-1)$ not $1/n$ because only $1/(n-1)$ will give the correct and unbiased estimate of the variance. From the probability theory in the earlier sections, we know that $E[x^2] = \mu^2 + \sigma^2$. The mean of the sample variance is

$$\mu_{S^2} = E[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2] = \frac{1}{n-1} \sum_{i=1}^{n} E[(x_i^2 - n\bar{x}^2)]. \tag{8.9}$$

Using $E[\bar{x}^2] = \mu^2 + \sigma^2/n$, we get

$$\mu_{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}\{E[x_i^2] - nE[\bar{x}^2]\}$$

$$= \frac{1}{n-1}\{n(\mu^2+\sigma^2) - n(\mu^2 + \frac{\sigma^2}{n})\} = \sigma^2. \tag{8.10}$$

Obviously, if we used the factor $1/n$ instead of $1/(n-1)$, we would get $\mu_{S^2} = \frac{n-1}{n}\sigma^2 < \sigma^2$, which would underestimate the sample variance. The other way to think about the factor $1/(n-1)$ is that we need at least one value to estimate the mean, we need at least 2 values to estimate the variance. Thus, for $n$ observations, only $n-1$ different values of variance can be obtained to estimate the total sample variance.

## 8.2 Method of Least Squares

### 8.2.1 Maximum Likelihood

For a sample of $n$ values $x_1, x_2, ..., x_n$ of a random variable $X$ whose probability density function $p(x)$ depends on a set of $k$ parameters $\beta_1, ..., \beta_k$, the joint probability is then

$$\Phi(\beta_1, ..., \beta_k) = \prod_{i=1}^{n} p(x_i, \beta_1, ..., \beta_k)$$

$$= p(x_1, \beta_1, ..., \beta_k)p(x_2, \beta_1, ..., \beta_k)\cdots p(x_n, \beta_1, ..., \beta_k). \tag{8.11}$$

The essence of the maximum likelihood is to maximise $\Phi$ by choosing the parameters $\beta_i$. As the sample can be considered as given values, the maximum likelihood requires that

$$\frac{\partial \Phi}{\partial \beta_i} = 0, \qquad (i = 1, 2, ..., k), \tag{8.12}$$

whose solutions for $\beta_i$ are the maximum likelihood estimates.

### 8.2.2 Linear Regression

For experiments and observations, we usually plot one variable such as pressure or price $y$ against another variable $x$ such as time or spatial coordinates. We try to present the data in such a way that we can see some trend in the data. For a set of $n$ data points $(x_i, y_i)$, the usual practice is to try to draw a straight line $y = a + bx$ so that it represents the major trend. Such a line is often called the regression line or the best fit line as shown in Figure 8.1.

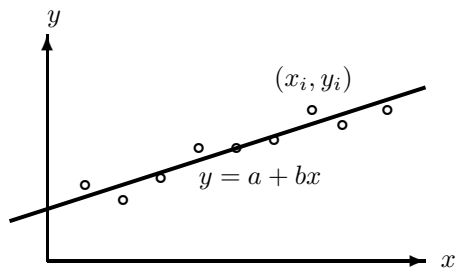



Figure 8.1: Least square and the best fit line.

The method of least squares is to try to determine the two parameters $a$ (intercept) and $b$ (slope) for the regression line from $n$ data points. Assuming that $x_i$ are known more precisely and $y_i$ values obey a normal distribution around the potentially best fit line with a variance $\sigma^2$, we have the probability

$$P = \prod_{i=1}^{n} p(y_i) = A \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - f(x_i)]^2\}, \tag{8.13}$$

where $A$ is a constant, and $f(x)$ is the function for the regression [$f(x) = a + bx$ for the linear regression]. It is worth pointing out that the exponent $\sum_{i=1}^{n} [y_i - f(x_i)]^2/\sigma^2$ is similar to the quantity $\chi^2_n$ defined in the $\chi^2$-distribution.

The essence of the method of least squares is to maximise the probability $P$ by choosing the appropriate $a$ and $b$. The maximisation of $P$ is equivalent to the minimisation of the exponent $\psi$

$$\psi = \sum_{i=1}^{n} [y_i - f(x_i)]^2. \tag{8.14}$$

We see that $\psi$ is the sum of the squares of the deviations $\epsilon_i^2 = (y_i - f(x_i))^2$ where $f(x_i) = a + bx_i$. The minimisation means the least sum of the squares, thus the name of the method of least squares.

In order to minimise $\psi$ as a function of $a$ and $b$, its derivatives should be zero. That is

$$\frac{\partial \psi}{\partial a} = -2 \sum_{i=1}^{n} [y - (a + bx_i)] = 0, \tag{8.15}$$

and

$$\frac{\partial \psi}{\partial b} = -2 \sum_{i=1}^{n} x_i [y_i - (a + bx_i)] = 0. \tag{8.16}$$

By expanding these equations, we have

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{8.17}$$

and

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i, \tag{8.18}$$

which is a system of linear equations for $a$ and $b$, and it is straightforward to obtain the solutions as

$$a = \frac{1}{n}[\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i] = \bar{y} - b\bar{x}, \tag{8.19}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \tag{8.20}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{8.21}$$

If we use the following notations

$$K_x = \sum_{i=1}^{n} x_i, \qquad K_y = \sum_{i=1}^{n} y_i, \tag{8.22}$$

and

$$K_{xx} = \sum_{i=1}^{n} x_i^2, \qquad K_{xy} = \sum_{i=1}^{n} x_i y_i, \tag{8.23}$$

then the above equations for $a$ and $b$ become

$$a = \frac{K_{xx}K_y - K_x K_{xy}}{nK_{xx} - (K_x)^2}, \qquad b = \frac{nK_{xy} - K_x K_y}{nK_{xx} - (K_x)^2}. \tag{8.24}$$

The residual error is defined by

$$\epsilon_i = y_i - (a + bx_i), \tag{8.25}$$

whose sample mean is given by

$$\begin{aligned}\mu_\epsilon &= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i = \frac{1}{n} y_i - a - b\frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \bar{y} - a - b\bar{x} = [\bar{y} - b\bar{x}] - a = 0.\end{aligned} \tag{8.26}$$

The sample variance $S^2$ is

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - (a + bx_i)]^2, \tag{8.27}$$

where the factor $1/(n-2)$ comes from the fact that two constraints are needed for the best fit, and the residuals therefore have $n-2$ degrees of freedom.

### 8.2.3 Correlation Coefficient

The correlation coefficient $r_{x,y}$ is a very useful parameter for finding any potential relationship between two sets of data $x_i$ and $y_i$ for two random variables $x$ and $y$, respectively. If $x$ has a mean $\mu_x$ and a sample variance $S_x^2$, and $y$ has a mean $\mu_y$ and a sample variance $S_y^2$, the correlation coefficient is defined by

$$r_{x,y} = \frac{\text{cov}(x,y)}{S_x S_y} = \frac{E[xy] - \mu_x\mu_y}{S_x S_y}, \tag{8.28}$$

where $\text{cov}(x,y) = E[(x-\mu_x)](y-\mu_y)$ is the covariance. If the two variables are independent, then $\text{cov}(x,y) = 0$, which means that there is no correlation between them ($r_{x,y} = 0$). If $r_{x,y}^2 = 1$, then there is a linear relationship between these two variables. $r_{x,y} = 1$ is an increasing linear relationship where the increase of one variable will lead to increase of another. $r_{x,y} = -1$ is a decreasing relationship when one increases while the other decreases.

For a set of $n$ data points $(x_i, y_i)$, the correlation coefficient can be calculated by

$$r_{x,y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{[n\sum x_i^2 - (\sum_{i=1}^{n} x_i)^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}}, \tag{8.29}$$

or

$$r_{x,y} = \frac{nK_{xy} - K_x K_y}{\sqrt{(nK_{xx} - K_x^2)(nK_{yy} - K_y^2)}}, \tag{8.30}$$

where $K_{yy} = \sum_{i=1}^{n} y_i^2$.

---

**Example 8.1:** At a strike-slip fault, both shear stress $\tau$ and the slip movement $s$ were recorded. We have
Shear stress (MPa) $\tau$: 16.2, 16.7, 16.8, 17.1, 17.4, 17.6, 18.3, 17.9;
Slip motion (cm/year) $s$: 0.55, 0.60, 0.75, 0.75, 0.85, 1.0, 1.1, 1.2.
The question is to find if there is any relationship between $\tau$ and $s$? From

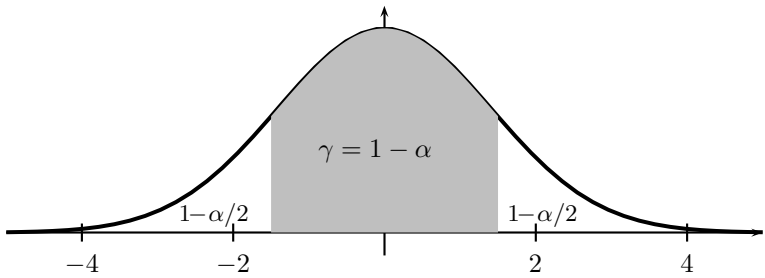



Figure 8.2: Confidence interval $\gamma = 1 - \alpha$.

these data, we know the sample mean $\mu_\tau = 17.25$, $\mu_s = 0.85$. The covariance $\text{cov}(h, s) = E[(h - \mu_h)(s - \mu_s)] = 0.132$. We also have the standard deviation of shear stress $S_h = 0.6422$ and the standard deviation of slip $S_s = 0.2179$. Therefore, the correlation coefficient $r$ is given by

$$r = \frac{\text{cov}(h, s)}{S_h S_s} \approx \frac{0.132}{0.6422 * 0.2179} \approx 0.94.$$

This is a relatively strong correlation. Indeed, there is a linear relationship between $\tau$ and $s$ in theory.

## 8.3 Hypothesis Testing

### 8.3.1 Confidence Interval

The confidence interval is defined as the interval $\theta_1 \leq X \leq \theta_2$ so that the probabilities at these two limits $\theta_1$ and $\theta_2$ are equal to a given probability $\gamma = 1 - \alpha$ (say, 95% or 99%). That is

$$P(\theta_1 \leq X \leq \theta_2) = \gamma = 1 - \alpha. \tag{8.31}$$

The predetermined parameter $\gamma$ is always near 1 so that it can be expressed as a small deviation $\alpha \ll 1$ from 1 (see Figure 8.2). If we choose $\gamma = 95\%$, it means that we can expect that about 95% of the sample will fall within the confidence interval while 5% of the data will not.

For the standard normal distribution, this means $P(-\theta \leq \xi \leq \theta) = 1 - \alpha$, so that

$$\phi(\xi \leq \theta) = 1 - \frac{\alpha}{2}. \tag{8.32}$$

If $\alpha = 0.05$, we have $\phi(\xi \leq \theta) = 0.975$ or $\theta = 1.960$. That is to say, $-\theta \leq \xi \leq \theta$ or $\mu - \theta\sigma \leq x \leq \mu + \theta\sigma$. We also know that if you repeat an

experiment $n$ times, the variance will decrease from $\sigma^2$ to $\sigma^2/n$, which is equivalent to saying that the standard deviation becomes $\sigma/\sqrt{n}$ for a sample size $n$. If $\alpha = 0.01$, then $\theta = 2.579$, we have

$$\mu - 2.579\frac{\sigma}{\sqrt{n}} \leq x \leq \mu + 2.579\frac{\sigma}{\sqrt{n}}. \tag{8.33}$$

On the other hand, for $\theta = 1$, we get $\mu - \sigma \leq x \leq \mu + \sigma$ and $\gamma = 0.682$. In other words, only 68.2% of the sample data will fall within the interval $[\mu - \sigma, \mu + \sigma]$ or

$$x = \mu \pm \sigma, \tag{8.34}$$

with a 68.2% confidence level.

It is conventional to use $\gamma = 0.95$ for probably significant, 0.99 for significant, and 0.999 for highly significant.

---

**Example 8.2:** The sample data of the time taken for a quick lunch at a restaurant are as follows: 19, 15, 30, 20, 15, 23, 28, 22, 23 minutes. Suppose you want to attend a lecture at 12:30, at what time should you start your order if you want to take 5% chance of being late? The sample mean is

$$\mu = \bar{x} = \frac{1}{9}(19 + 15 + 30 + 20 + 15 + 23 + 28 + 22 + 23) = 21.67.$$

The sample variance is

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 = 26.5,$$

which gives a standard deviation of $\sigma = 5.15$ minutes. If you are willing to take 5% chance, then $\phi(\xi) = 0.95$, it gives $\xi = 1.645$. So you should start

$$x = \mu + \xi\sigma = 30.15,$$

which is about 30 minutes earlier or at about 12:00.

---

### 8.3.2 Student's $t$-distribution

The Student's $t$-test is a very powerful method for testing the null hypothesis to see if the means of two normally distributed samples are equal. This method was designed by W. S. Gosset in 1908 and he had to use a pen name 'Student' because of his employer's policy (Guinness Brewery) in publishing research results at that time. This is a powerful method for hypothesis testing using small-size samples.

This test can also be used to test if the slope of the regression line is significantly different from 0. It has become one of the most popular methods for hypothesis testing. The theoretical basis of the $t$-test is the Student's $t$-distribution for a sample population with the unknown standard deviation $\sigma$, which of course can be estimated in terms of the sample variance $S^2$ from the sample data.

For $n$ independent measurements/data $x_1, x_2, ..., x_n$ with an estimated sample mean $\bar{x}$ and a sample variance $S^2$ as defined by Eq.(8.8), the $t$-variable is defined by

$$t = \frac{\bar{x} - \mu}{(S/\sqrt{n})}. \tag{8.35}$$

The Student's $t$-distribution with $k = n - 1$ degrees of freedom is the distribution for the random variable $t$, and the probability density function is

$$p(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(k/2)}[1 + \frac{t^2}{k}]^{-\frac{k+1}{2}}. \tag{8.36}$$

It can be verified that the mean is $E[t] = 0$. The variance is $\sigma^2 = k/(k-2)$ for $k > 2$ and infinite for $0 < k \leq 2$.

The corresponding cumulative probability function is

$$F(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(k/2)} \int_{\infty}^{t} [1 + \frac{\zeta^2}{k}]^{-\frac{k+1}{2}} d\zeta. \tag{8.37}$$

This integral leads to a hypergeometric function, which is not straightforward to calculate, which is why they are tabulated in many statistical tables. For a confidence level of $\gamma = 1 - \alpha$, the confidence interval is given by

$$F(\theta) = 1 - \frac{\alpha}{2}, \tag{8.38}$$

which is usually tabulated. For $\alpha = 0.05$ and 0.01 (or $1 - \alpha/2 = 0.975$ and 0.995), the values are tabulated in Table 8.1.

Suppose we are dealing with the 95% confidence interval, we have $p(-\theta \leq t \leq \theta) = 1 - \alpha = 0.95$ or $p(t \leq \theta) = 1 - \alpha/2 = 0.975$, we have $\theta = t_{\alpha,k} = 12.70(k = 1)$, $4.30(k = 2)$, $3.18(k = 3)$, ..., $2.228(k = 10)$, ..., 1.959 for $k \to \infty$. Hence,

$$\mu - \theta\frac{S}{\sqrt{n}} \leq t \leq \mu + \theta\frac{S}{\sqrt{n}}. \tag{8.39}$$

This is much more complicated than its counterpart, the standard normal distribution.

Table 8.1: Limits defined by $F(\theta) = 1 - \alpha/2$ in Eq.(8.38).

| $k$ | $F(\theta)_{0.975}$ | $F(\theta)_{0.995}$ |
|---|---|---|
| 1 | 12.7 | 63.7 |
| 2 | 4.30 | 9.93 |
| 3 | 3.18 | 5.84 |
| 4 | 2.78 | 4.60 |
| 5 | 2.57 | 4.03 |
| 6 | 2.45 | 3.71 |
| 7 | 2.37 | 3.50 |
| 8 | 2.31 | 3.36 |
| 9 | 2.26 | 3.25 |
| 10 | 2.23 | 3.17 |
| 20 | 2.09 | 2.85 |
| 50 | 2.01 | 2.68 |
| 100 | 1.98 | 2.63 |
| $\infty$ | 1.96 | 2.58 |

### 8.3.3 Student's $t$-test

There are quite a few variations of the Student's $t$-test, and most common $t$-tests are the one-sample $t$-test and the two-sample $t$-test. The one sample $t$-test is used for measurements that are randomly drawn from a population to compare the sample mean with a known number.

In order to do statistical testing, we first have to pose precise questions or form a hypothesis, which is conventionally called the null hypothesis. The basic steps of a $t$-test are as follows:

1. The null hypothesis: $H_0$: $\mu = \mu_0$ (often known value) for one sample, or $H_0$: $\mu_1 = \mu_2$ for two samples;

2. Calculate the $t$-test statistic $t$ and find the critical value $\theta$ for a given confidence level $\gamma = 1 - \alpha$ by using $F(t \leq \theta) = 1 - \alpha/2$;

3. If $|t| > \theta$, reject the hypothesis. Otherwise, accept the hypothesis.

---

**Example 8.3:** A study claims that a large region (with more than 250 subregions) has an averaged rate of soil erosion of 110 m/Ma (or $\mu_0 = 110$). Then, you randomly sampled 11 subregions to get the erosion rates and the results are: $x = 106, 112, 103, 108, 108, 109, 100, 106, 106, 99, 101$. Test the hypothesis:

$$\mathrm{H}_0 : \mu = \mu_0,$$

at a confidence level of $95\%$.

From the data, we know that $n = 11$, $\bar{x} = 105.273$, $S = 4.077$. Then, we have

$$t = \frac{(\bar{x} - \mu)}{(S/\sqrt{n})} = \frac{(105.273 - 110)}{4.077/\sqrt{11}} \approx -3.846.$$

We only use the positive value if we look at the statistical tables. We also know for $k = n - 1 = 10$ degrees of freedom at a $95\%$ confidence level, $\theta = 2.228$. At a $95\%$ confidence level, the probability of $t > \theta$ is $0.025$ (or $2.5\%$) and the probability $t < -\theta$ is also $0.025$. Thus, the hypothesis is not valid at a $95\%$ confidence level. At the same level of confidence, the true mean $\mu_0$ of erosion rate lies in the range of

$$\bar{x} - 2.228 * S/\sqrt{11} \leq \mu_0 \leq \bar{x} + 2.228 S/\sqrt{11}$$

or $102.53 \leq \mu_0 \leq 108.00$.

---

Another important $t$-test is the two-sample paired test. Assuming that two pairs of $n$ sample data sets $U_i$ and $V_i$ are independent and drawn from the same normal distribution, the paired $t$-test is used to determine whether they are significantly different from each other. The $t$-variable is defined by

$$t = \frac{(\bar{U} - \bar{V})}{S_d/\sqrt{n}} = (\bar{U} - \bar{V}) \sqrt{\frac{n(n-1)}{\sum_{i=1}^{n} (\tilde{U}_i - \tilde{V}_i)^2}}, \tag{8.40}$$

where $\tilde{U}_i = U_i - \bar{U}$ and $\tilde{V}_i = V_i - \bar{V}$. In addition,

$$S_d^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{U}_i - \tilde{V}_i)^2. \tag{8.41}$$

This is equivalent to apply the one-sample test to the difference $U_i - V_i$ data sequence.

---

**Example 8.4:** A novel pumping method of extracting oil was tried in a group of wells in an oilfield (denoted by $B$), while a standard method was used for the same wells before (denoted by $A$). At the end of the assessment, the flow rates (in units of 1000 gallons per day) of 8 wells were recorded as follows:
Standard method (A): $U_i$ = 76, 77, 76, 81, 77, 76, 75, 82;
Novel method (B): $V_i$ = 79, 81, 77, 86, 82, 81, 82, 80.
At a $95\%$ confidence level, can you say the new method is really better than the standard method?

If we suppose that the two methods do not produce any difference in results, *i.e.*, their means are the same. Thus the null hypothesis is:

$$H_0 : \mu_A = \mu_B.$$

We know that $\bar{U} = 77.5$, $\bar{V} = 81$. The combined sample variance $Sd = 2.828$. We now have

$$t = \frac{\bar{U} - \bar{V}}{S_d/\sqrt{n}} = \frac{77.5 - 81}{2.828/\sqrt{8}} = -3.5.$$

We know from the statistical table that the critical value $\theta = 2.37$ for $F(\theta) = 1 - \alpha/2$ and $k = n - 1 = 7$. As $t < -\theta$ or $t > \theta$, we can reject the null hypothesis. That is to say, the new pumping method does produce better results.

---

The variance analysis and hypothesis testing are important topics in applied statistics, and there are many excellent books on these topics. Readers can refer to the relevant books listed at the end of this book. It is worth pointing out that other important methods for hypothesis testing are Fisher's $F$-test, $\chi^2$-test, and non-parametric tests. What we have discussed in this chapter is just a tip of an iceberg, however, it forms a solid basis for further studies.

## 8.4 Data Interpolation

Before we can discuss the geostatistical interpolation methods such as kriging in detail, we have to briefly review the standard interpolation methods such as the spline interpolation and the Bézier curve.

### 8.4.1 Spline Interpolation

The spline interpolation is to construct a function, called spline function, of degree $m$ for given $n + 1$ known values $y_i$ at $n + 1$ data points $x_i (i = 0, 1, 2, ..., n)$. These given points are organised in an increasing order so that

$$x_0 < x_1 < ... < x_n. \tag{8.42}$$

The values $y_i$ at the given data points are often called knot values. The major requirement is that the constructed spline function $S(x)$ should be continuous and produce the exact value at the data points. The spline function can be constructed in a piecewise manner so that each spline function $S_i(x)$ is valid in each data interval $x \in [x_i, x_{i+1}]$.

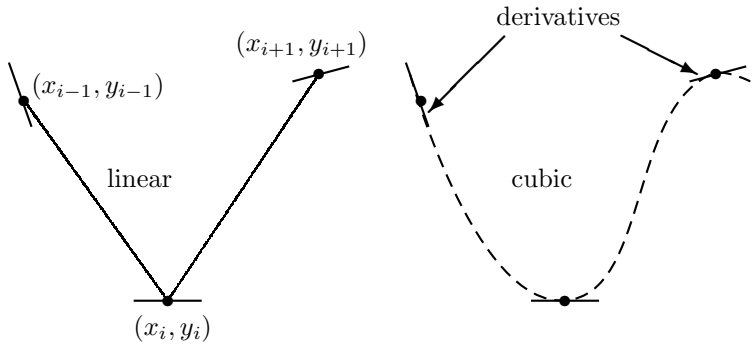



Figure 8.3: Spline construction for given values $(y_{i-1}, y_i, y_{i+1})$ and their derivatives $(y'_{i-1}, y'_i, y'_{i+1})$ at three points $x_{i-1}$, $x_i$ and $x_{i+1}$.

**Linear Spline Functions**

In the simplest case of three points with given values $(y_{i-1}, y_i, y_{i+1})$ at three distinct points $x_{i-1}$, $x_i$ and $x_{i+1}$, the simplest spline is the linear functions, which can be constructed in each interval $[x_i, x_{i+1}]$ so that we have

$$S_i(x) = y_i + \frac{(y_{i+1} - y_i)}{(x_{i+1} - x_i)}(x - x_i), \quad x \in [x_i, x_{i+1}], \tag{8.43}$$

which corresponds to the piecewise line segments in Fig. 8.3.

For two consecutive intervals, the functions $S_{i-1}$ and $S_i$ should be continuous, that is to say, $S_i(x_i) = S_{i-1}(x_i) = y_i$. As $i = 0, 2, ..., n-1$, there are $n$ such spline functions. In order to systematically construct the spline interpolation for $n+1$ points, we can write

$$y = f(x) = \sum_{i=0}^{n} y_i \theta(x), \tag{8.44}$$

where $\theta_i(x)$ are the linear elementary basis functions.

$$\theta_i(x) = \begin{cases} \frac{x - x_i}{h_i}, & x \in [x_i, x_{i+1}] \\ \\ \frac{x_i - x}{h_{i-1}}, & x \in [x_{i-1}, x_i], \end{cases} \tag{8.45}$$

where $h_i = (x_{i+1} - x_i)$ and $h_{i-1} = (x_i - x_{i-1})$. The linear spline function $S_i(x)$ in the interval $x \in [x_{i-1}, x_x$ is expressed as

$$S_i(x) = y_{i-1}\theta_{i-1} + y_i\theta_i, \tag{8.46}$$

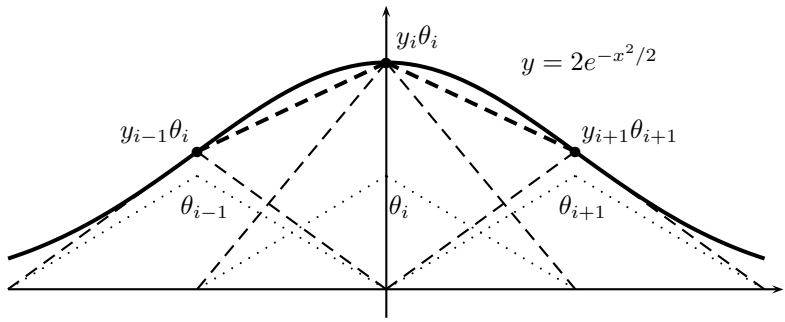



Figure 8.4: Linear spline construction from $\theta(x)$.

which can be represented geometrically as the dashed and dotted lines in Fig. 8.4. We can see that $S_i(x)$ is continuous, but not necessary smooth. The first derivative of the linear spline function are not continuous.

### Cubic Spline Functions

The linear spline interpolation is continuous, but not smooth as there is a discontinuity in the first derivatives. Now suppose the three derivatives $y'_{i-1}, y'_i, y'_{i+1}$ are also given at these three points $(x_{i-1}, y_{i-1})$, $(x_i, y_i)$ and $(x_{i+1}, y_{i+1})$, can we construct a class of better and smoother spline functions (such as the dashed curve shown in Fig. 8.3) so that their values and derivatives meet the given conditions? The answer is yes, that is the cubic spline function. You may wonder why cubic?

In the interval $[x_i, x_{i+1}]$, we now have four conditions: two function values $y_i, y_{i+1}$, and two derivatives $y'_i, y'_{i+1}$. Obviously, linear function is not enough, how about quadratic function $S(x) = ax^2 + bx + c$? The function is relatively smooth and the first derivative could meet these requirements, but the second derivative $S''(x) = a$ is constant, and if we require the second-derivative also continuous, this means that either the second is constant everywhere (thus $a$ is same everywhere) or there is a discontinuity. Furthermore, we have only three unknown $a, b, c$ with four conditions, so in general they are over-determined, and not all conditions will be met. Thus, we need a cubic function.

If we use the following generic cubic function

$$S_i(x) = \alpha_i(x - x_i)^3 + \beta_i(x - x_i)^2 + \gamma_i(x - x_i) + \delta_i, \tag{8.47}$$

where $i = 1, 2, ..., n$, then the spline function is twice continuous differ-

entiable. Its first derivative is

$$S_i'(x) = 3\alpha_i(x - x_i)^2 + 2\beta_i(x - x_i) + \gamma_i. \tag{8.48}$$

The four conditions become

$$S_i(x_i) = y_i, \qquad S_i(x_{i+1}) = y_{i+1}, \tag{8.49}$$

and

$$S_i'(x_i) = y_i', \qquad S_i'(x_{i+1}) = y_{i+1}'. \tag{8.50}$$

Thus, four equations and four unknown $\alpha_i, \beta_i, \gamma_i$ and $\delta_i$ which are uniquely determined. The general requirements among different intervals are $S_i(x), S_i'(x)$, and $S_i''(x)$ should be continuous. We have

$$S_i(x_i) = y_i, S_i(x_{i+1}) = y_{i+1}, \qquad i = 0, 2, ..., n-1, \tag{8.51}$$

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) = y'(x_{i+1}), \quad i = 0, 1, ..., n-2, \tag{8.52}$$

and

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}), \qquad i = 0, 1, ..., n-2. \tag{8.53}$$

In $n$ intervals, we have $4n$ unknown, but we have $4n-2$ conditions: $n+1$ from $y_i (i = 0, 1, ..., n-1)$; $2(n-1)$ from $y_i'$ and $S_i'$ $(i = 0, 1, ..., n-2)$; $n-1$ from $S_i'', (i = 0, 1, ..., n-2)$, so we need 2 more conditions.

The two extra conditions are at the two end points $i = 0$ and $i = n$. The clamped boundary conditions are to set

$$S_0'(x_0) = y_0', \qquad S_{n-1}'(x_n) = y_n', \tag{8.54}$$

which are usually given. If the derivatives at the end points are not given, we can use the natural or free boundary conditions:

$$S_0''(x_0) = 0, \qquad S_{n-1}''(x_n) = 0. \tag{8.55}$$

In order to find the spline functions $S_i(x)$, it is conventional to rewrite them in terms of the second derivative $\xi_i = S_i''(x)$, $(i = 0, 1, 2, ..., n-1)$, and we have

$$\xi_i = 6\alpha_i(x - x_i) + 2\beta_i. \tag{8.56}$$

At $x = x_i$, we have

$$\xi_i(x_i) = 2\beta_i, \tag{8.57}$$

or

$$\beta_i = \frac{\xi_i}{2}. \tag{8.58}$$

Using the continuity of $\xi_i$ at $x = x_{i+1}$, we have

$$\xi_i(x_{i+1}) = \xi_{i+1}(x_{i+1}) = 6\alpha_i h_i + 2\beta_i, \quad h_i = x_{i+1} - x_i. \tag{8.59}$$

Combining with Eq.(8.58), we have

$$\alpha_i = \frac{\xi_{i+1} - \xi_i}{6h_i}. \tag{8.60}$$

Since $S_i(x_i) = y_i$ at $x = x_{i+1}$, we have

$$\delta_i = y_i. \tag{8.61}$$

Substituting $\alpha_i, \beta_i$ and $\delta_i$ into $S_i(x_{i+1}) = y_{i+1}$ at $x = x_{i+1}$ and after some rearrangement, we have

$$\gamma_i = \frac{(y_{i+1} - y_i)}{h_i} - \frac{h_i(2\xi_i + \xi_{i+1})}{6}. \tag{8.62}$$

Now substituting these coefficients into Eq.(8.5), we can express $S_i$ in terms of the second derivatives $\xi_i$, and we have

$$S_i(x) = \frac{1}{6h_i}[\xi_{i+1}(x - x_i)^3 + \xi_i(x_{i+1} - x)^3]$$

$$+[\frac{y_{i+1}}{h_i} - \frac{h_i}{6}\xi_{i+1}](x - x_i) + [\frac{y_i}{h_i} - \frac{h_i}{6}\xi_i](x_{i+1} - x). \tag{8.63}$$

This is a cubic polynomial, and the only thing left is to find the coefficients $\xi_i$. Using the continuity conditions: $S_i(x_i) = S_{i-1}(x_i) = y_i$ and $S_i'(x_i) = S_{i-1}'(x_i)$ at $x = x_i$; $S_i(x_{i+1}) = S_{i+1}(x_{i+1}) = y_{i+1}$ and $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$ at $x = x_{i+1}$, we can rewrite the above equation as

$$h_{i-1}\xi_{i-1} + 2(h_{i-1} + h_i)\xi_i + h_i\xi_{i+1} = 6[\frac{(y_{i+1} - y_i)}{h_i} - \frac{(y_i - y_{i-1})}{h_{i-1}}], \tag{8.64}$$

where $i = 1, 2, ..., n-1$. Writing them in a matrix form, we have

$$\begin{pmatrix} 2(h_0 + h_1) & h_1 & ... & 0 \\ h_1 & 2(h_1 + h_2) & ... & 0 \\ \vdots & & \ddots & h_{n-2} \\ 0 & ... & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n-1} \end{pmatrix}$$

$$= \begin{pmatrix} 6(\frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0}) \\ \vdots \\ 6(\frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}}) \end{pmatrix}. \tag{8.65}$$

Since $\xi_0 = 0$ and $\xi_n = 0$ are given from the natural boundary conditions, this linear system will uniquely determine $\xi_1, ..., \xi_{n-1}$. For any

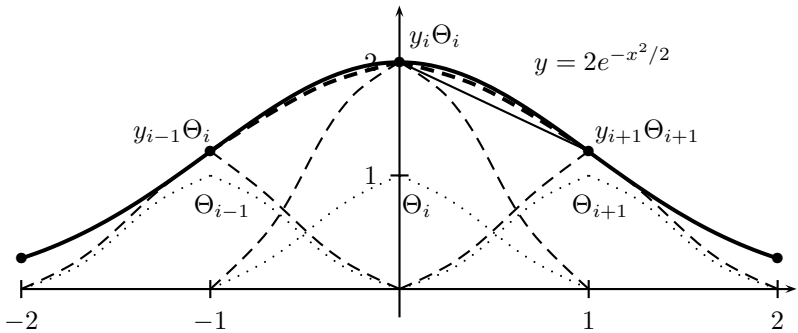


Figure 8.5: Cubic spline construction from $\Theta(x)$.

given set of data, we should solve the linear system to get $\xi_i$, then to compute $S_i(x)$.

In the case of equal spacing $h_0 = h_i = h_{n-1} = h$, the above equation becomes

$$\begin{pmatrix} 4 & 1 & ... & 0 \\ 1 & 4 & ... & 0 \\ \vdots & & \ddots & 1 \\ 0 & ... & 1 & 4 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n-1} \end{pmatrix} = \frac{6}{h^2} \begin{pmatrix} y_2 - 2y_1 + y_0 \\ y_3 - 2y_2 + y_1 \\ \vdots \\ y_n - 2y_{n-1} + y_{n-2} \end{pmatrix}. \tag{8.66}$$

---

**Example 8.5:** For the function

$$y = 2e^{-\frac{x^2}{2}},$$

we now try to approximate it using cubic spline functions constructed from five points $x_0 = -2, x_1 = -1, x_2 = 0,\ x_3 = 1, x_4 = 2$. We known that $y_0 = 0.2707, y_1 = 1.2130, y_2 = 2.0,\ y_3 = 1.2130, y_4 = 0.2707$. Since these points are equally-spaced with $h = 1$, equation (8.66) then becomes

$$\begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} -0.9327 \\ -9.443 \\ -0.9327 \end{pmatrix},$$

whose solution is

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 0.4080 \\ -2.565 \\ 0.4080 \end{pmatrix}.$$

At the two end points, we use the natural boundary condition $\xi_0 = 0$ and $\xi_4 = 0$. Now substituting $\xi_0, \xi_1, \xi_2, \xi_3, \xi_4$ into Eq.(8.63) for $i = 0, 1, 2, 3,$

we have

$$S_0 = \frac{0.4080}{6}(x+2)^3 + 0.4275(x+2),$$

$$S_1 = -0.4955(x+1)^3 + 1.282(x+1),$$

$$S_2 = 0.4955(x-1)^3 - 1.282(x-1),$$

and

$$S_3 = -\frac{0.4080}{6}(x-2)^3 - 0.4275(x-2).$$

These spline functions are plotted in Fig. 8.5 as the heavy dashed curve. We can see that cubic spline curves almost fall on the exact curve (solid) of the original function $y = 2\exp(-x^2/2)$.

---

Alternatively, similar to the elementary basis function $\theta_i(x)$ for linear spline, we can find the corresponding basis function $\Theta_i(x)$ for cubic function. In a given interval $[x_i, x_{i+1}]$, if we assume that $\Theta_i(x_i) = 0$, and $\Theta_i'(x_i) = 0$, we get

$$\Theta_i(x) = (x - x_i)^2[\alpha(x - x_i) + \beta]. \tag{8.67}$$

As we require that $\Theta_i(x)$ reaches the maximum $\Theta_i = 1$ at $x = x_{i+1}$, it leads to

$$\Theta_i'(x_{i+1}) = 3\alpha h_i^2 + 2\beta h_i = 0, \tag{8.68}$$

and

$$\Theta_i(x_{i+1}) = h_i^2(\alpha h_i + \beta) = 1. \tag{8.69}$$

The solutions are

$$\alpha = -\frac{2}{h_i^3}, \qquad \beta = \frac{3}{h_i^2}. \tag{8.70}$$

Therefore, we get

$$\Theta_i = (x - x_i)^2[\frac{3}{h_i^2} - \frac{2(x - x_i)}{h_i^3}]. \tag{8.71}$$

The cubic spline function can in general be written as

$$S(x) = \sum_{i=0}^{n-1} y_i \Theta_i(x), \tag{8.72}$$

which can be geometrically represented as the dashed and dotted curves in Fig. 8.5 where as an example, $y = 2\exp(-x^2/2)$ is approximated using the cubic spline functions.

### 8.4.2 Lagrange Interpolating Polynomials

We have seen that the construction of spline functions is tedious. Lagrange polynomials provide a systematic way to construct interpolation functions.

For any given $n$ points $(x_i, y_i), (i = 1, 2, ..., n)$, there is a Lagrange interpolating polynomial $P(x)$ of degree $k \leq (n-1)$ which passes through all $n$ points. That is

$$P(x) = \sum_{i=1}^{n} P_i(x) y_i, \tag{8.73}$$

where

$$P_i = \prod_{j=1, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)}. \tag{8.74}$$

For example, for $n = 4$, we have

$$P(x) = \frac{(x-x_2)(x-x_3)(x-x_4)y_1}{(x_1-x_2)(x_1-x_3)(x_1-x_4)} + \frac{(x-x_1)(x-x_3)(x-x_4)y_2}{(x_2-x_1)(x_2-x_3)(x_2-x_4)}$$
$$+ \frac{(x-x_1)(x-x_2)(x-x_4)y_3}{(x_3-x_1)(x_3-x_2)(x_3-x_4)} + \frac{(x-x_1)(x-x_2)(x-x_3)y_4}{(x_4-x_1)(x_4-x_2)(x_4-x_3)}. \tag{8.75}$$

---

**Example 8.6:** For four equally-spaced points $(x_i, y_i) = (1, 1)$, $(2, 2)$, $(3, -0.5)$, and $(4, 1)$, the Lagrange polynomial becomes

$$P(x) = -\frac{(x-2)(x-3)(x-4)}{6} y_1 + \frac{(x-1)(x-3)(x-4)}{2} y_2$$
$$- \frac{(x-1)(x-2)(x-4)}{2} y_3 + \frac{(x-1)(x-2)(x-3)}{6} y_4,$$

which is equivalent to

$$P(x) = 1.25x^3 - 9.25x^2 + 20x - 11.$$

The Lagrange polynomial and the four points are plotted in Fig. 8.6.

---

The disadvantage of Lagrange polynomials is that when $n$ increases, the order of the polynomials also increases, and this leads to greater oscillations between data points. For equally-spaced points, the Lagrange interpolation oscillates around the true function. However, the advantages of Lagrange polynomials are that they are unique and rigorous and thus they become handy in mathematical proofs. In addition, they form the basic formulation for shape functions in finite element analysis discussed in later chapters and they are also widely used in signal processing including audio-video analysis and seismic wave processing.

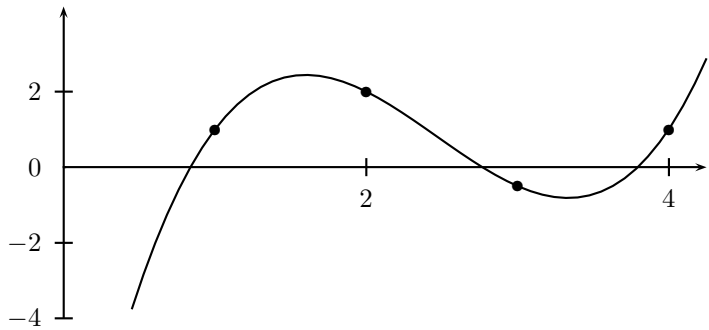



Figure 8.6: Lagrange polynomial of degree 3 for 4 given points.

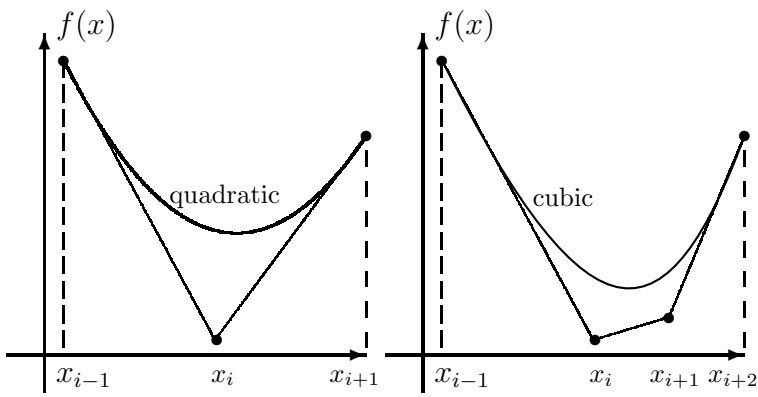



Figure 8.7: Bézier interpolation: quadratic and cubic.

### 8.4.3 Bézier Curve

We now know that linear and quadratic spline functions are not quite smooth, but splines of higher degrees are not straightforward to construct. There is an alternative way to construct smooth interpolation functions, that is to use Bézier curves (see Fig. 8.7). These interpolation curves are smooth and can be easily extended to higher dimensions to construct surfaces and volumes. Therefore, it is widely used in engineering, computer graphics, and earth sciences.

The quadratic Bézier curve for any three points: $P_0(x_{i-1},\ y_{i-1})$, $P_1(x_i, y_i)$, $P_2(x_{i+1}, y_{i+1})$ as shown in Fig. 8.7, can be constructed using a parameter $t \in [0, 1]$

$$P(t) = (1-t)^2 P_0 + 2t(1-t)P_1 + t^2 P_2, \tag{8.76}$$

which is equivalent to

$$x(t) = (1-t)^2 x_{i-1} + 2t(1-t)x_i + t^2 x_{i+1}, \tag{8.77}$$

and

$$y(t) = (1-t)^2 y_{i-1} + 2t(1-t)y_i + t^3 y_{i+1}. \tag{8.78}$$

Clearly, $t = 0$ corresponds to $(x_{i-1}, y_{i-1})$ (end point), while $t = 1$ gives $(x_{i+1}, y_{i+1})$ (another end point). The only unusual feature is that the curve does not go through the point $(x_i, y_i)$. This might be a disadvantage as it is not an exact interpolation, however, it becomes an advantage as the curve is very smooth and tangential to both end points. This characteristic is also true for Bézier curves of higher degrees.

For four points, the cubic Bézier curve leads to

$$x(t) = (1-t)^3 x_{i-1} + 3t(1-t)^2 x_i + 3t^2(1-t)x_{i+1} + t^3 x_{i+2}, \tag{8.79}$$

and

$$y(t) = (1-t)^3 y_{i-1} + 3t(1-t)^2 y_i + 3t^2(1-t)y_{i+1} + t^3 y_{i+2}, \tag{8.80}$$

where $t \in [0, 1]$. It is straightforward to extend any degree $n$ using coefficients of binomial expansion $[(1-t)+t]^n$.

## 8.5 Kriging

Kriging is a class of interpolation techniques for random fields used in geostatistics. It was named after a South-African mining engineer, D.G. Krige, who developed this method for grading gold mines. It is an exact interpolation method even for random observed data, and it is thus widely used in both deterministic and random simulations. Kriging methods are widely used in mining, remote sensing, environmental sciences, engineering and earth sciences.

The above Bézier and spline interpolation methods are referred to as deterministic interpolation methods because they are essentially based on the surrounding measured or observed values and the smoothness of the resulting surface is determined by specific mathematical functions. Kriging, on the other hand, is a family of statistical interpolation methods for constructing smooth interpolations based on the statistical autocorrelation among the measured points. Because the information of the measured data is used, kriging is thus capable of predicting with some statistical certainty (or accuracy). Therefore, it is widely used in earth sciences to make predictions for unmeasured locations based on the surrounding measured values.

Kriging is a geostatistical procedure to construct a linear estimate, with a minimum error variance, at a location where the true value is not known. Whether the data sets are mineral distribution in 3D, or observations of some quantities over a 2-D region of the Earth's surface, or the cross-section of a stratigraphic column in 1-D, or a time series, the formulations of kriging are the same if we formulate the method in terms of general location variable $\boldsymbol{x} = (x_1, x_2, ...)^T = (x, y, ...)^T$ in a domain $\Omega$.

The basic assumption of kriging is that the field value $Z(\boldsymbol{x})$ at the location $\boldsymbol{x}$ is a random variable, and each measurement is a realisation of the random variable. The difference in values is similar for a similar distance $h$, and the closer the observed spatial data are, the more positively correlated they are. Thus, the proximity of the data and clustering will give indication of the possible true values at unobserved locations. For given $n$ observations $(Z_i, i = 1, 2, ..., n)$, the general formula for predicting $Z(x, y)$ at location $(x, y)$ in kriging is given by

$$\hat{Z}(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i Z_i = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i), \tag{8.81}$$

which is the weighted sum of the measured data. $Z_i = Z(\boldsymbol{x}_i)$ is the observed value at the location $\boldsymbol{x}_i$ which is simply $\boldsymbol{x}_i = x_i$ one-dimensional case and $\boldsymbol{x}_i = (x_i, y_i)$ in two-dimensional case. $\lambda_i$ is the weighting coefficient, to be determined, associated with the $i$-th location. The essence of kriging is the choice of $\lambda_i$ in such a way that these coefficients depend not only on the distance from the prediction location to the measured points, but also on the spatial distribution of these measured points in terms of their spatial autocorrelation and/or variogram.

Let $\mu(Z_i)$ and $\mu(\hat{Z})$ be the sample mean of the measured data $Z_i$ and kriging mean, respectively. Assuming that the true value of $Z$ at $\boldsymbol{x}$ is $Z(\boldsymbol{x})$, the mean of the prediction or estimate should be unbiased, which requires that

$$E[\hat{Z} - Z] = \sum_{i=1}^{n} \lambda_i \mu(Z_i) - \mu(\hat{Z}) = 0. \tag{8.82}$$

There are many variations of kriging methods which depend on the assumption of the kriging mean. Three general kriging methods are: simple, ordinary, and universal kriging. In simple kriging, the kriging mean is assumed known and taken to be zero, or $\mu(\hat{Z}) = 0$. In the ordinary kriging, $\mu(\hat{Z})$ is assumed an unknown constant $\mu(\hat{Z}) = \mu$, which implies that

$$\sum_{i=1}^{n} \lambda_i = 1. \tag{8.83}$$

Universal kriging assumes a linear trend in the data, thus the kriging mean will not be a constant.

Ordinary kriging is the most widely used, while the universal kriging should only be used if there is a known trend in the observed data.

Let us define the residual of the data values as

$$\epsilon(\boldsymbol{x}_i) = Z(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i), \quad (i = 1, 2, ..., n), \tag{8.84}$$

where $\mu(\boldsymbol{x}_i)$ can be considered as constant, but unknown. The semivariogram $\gamma(h)$ is defined as

$$\gamma(h) \equiv \frac{1}{2} E\{[\epsilon(\boldsymbol{x}) - \epsilon(\boldsymbol{x} + h)]^2\} \tag{8.85}$$

where $h$ is the distance between all pairs of data from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$. The variogram is just $2\gamma(h)$. For the simple and ordinary kriging methods where $\mu(\boldsymbol{x}_i)$ is either zero or an unknown constant, we can use the measured values to calculate $\gamma(h)$, and we have

$$\gamma(h) = \frac{1}{2} E\{[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)]^2\}. \tag{8.86}$$

It is relatively straightforward to show that $\gamma(h) = \gamma(-h)$ and $\gamma(0) = 0$ due to the fact that the variance $[\epsilon(\boldsymbol{x}_i) - \epsilon(\boldsymbol{x}_j)]^2 = [\epsilon(\boldsymbol{x}_j) - \epsilon(\boldsymbol{x}_j)]^2$. However, the limit $\lim_{h \to 0} \gamma(h) = C_0 \geq 0$ is not necessarily zero. If $C_0 \neq 0$, we have the so-called nugget effect, which may be due either to some miscroscale variation or to measurement error. The term nugget effect comes from the spatial variation caused by small nuggets of ore, and this is of course an exaggeration. It is sometimes not easy to find a pattern or trend from a semivariogram calculated using the above formula, it is more feasible to divide the pair of data into many bins with regular intervals of $h$, which leads to

$$\gamma(h) = \frac{1}{2N(h)} \sum_{N(h)} [\epsilon(\boldsymbol{x}_i) - \epsilon(\boldsymbol{x}_j)] = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [\epsilon(\boldsymbol{x}_i) - \epsilon(\boldsymbol{x}_i + h)]^2, \tag{8.87}$$

where $N(h)$ is the set of data pair at $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ that have a similar $h$.

The covariance is defined as

$$C(h) = E[\epsilon(\boldsymbol{x})\epsilon(\boldsymbol{x} + h)]. \tag{8.88}$$

Now expand the expression of the variogram, and we have

$$\gamma(h) = \frac{1}{2}\{E[\epsilon^2(\boldsymbol{x})] + E[\epsilon(\boldsymbol{x} + h)^2] - 2E[\epsilon(x)\epsilon(\boldsymbol{x} + h)]\}$$

$$= \frac{1}{2}\{\mathrm{Var}[\epsilon(\boldsymbol{x})] + \mathrm{Var}[\epsilon(\boldsymbol{x} + h)] - 2C(h)\} = C(0) - C(h), \tag{8.89}$$

that is

$$\gamma(h) = C(0) - C(h), \tag{8.90}$$

where $C(0) = E[\epsilon^2(\boldsymbol{x})]$. Here we can see that the covariance $C(h)$ is a function of distance $h$ only, and $C(0)$ is the variance $\mathrm{Var}(\epsilon(\boldsymbol{x}))$ or $\sigma_Z^2$. In kriging, it is assumed that $C(h)$ a simple function of $h$. For example, $C(h) = \sigma_Z^2 \exp(-\theta h^2)$ where $\theta$ is a shape parameter. We will discuss this in more detail later.

The objective of the kriging is to reduce the error variance

$$\begin{aligned}\Pi &= E\{[(\hat{Z}(\boldsymbol{x}) - Z(\boldsymbol{x})]^2\} = E\{[\hat{\epsilon}(\boldsymbol{x}) - \epsilon(\boldsymbol{x})]^2\} \\ &= E[\hat{\epsilon}^2(\boldsymbol{x})] + E[\epsilon^2(\boldsymbol{x})] - 2E[\hat{\epsilon}(\boldsymbol{x})\epsilon(\boldsymbol{x})],\end{aligned} \tag{8.91}$$

where $\hat{\epsilon}(\boldsymbol{x}) = Z(\boldsymbol{x}) - \mu$. Substituting (8.81), we have

$$\Pi = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2\sum_{i=1}^{n} \lambda_i C(\boldsymbol{x}, \boldsymbol{x}_i) + C(0), \tag{8.92}$$

where $C(\boldsymbol{x}_i, \boldsymbol{x}_j) = E[\epsilon(\boldsymbol{x}_i)\epsilon(\boldsymbol{x}_j)]$ and $C(\boldsymbol{x}, \boldsymbol{x}_i) = E[\epsilon(\boldsymbol{x})\epsilon(\boldsymbol{x}_i)]$.

The minimisation of $\Pi$ requires the following stationary conditions

$$\frac{\partial \Pi}{\partial \lambda_i} = 2[\sum_{j=1}^{n} \lambda_j C(\boldsymbol{x}_i, \boldsymbol{x}_j) - C(\boldsymbol{x}, \boldsymbol{x}_i)] = 0, \tag{8.93}$$

where $i = 1, 2, ..., n$. This leads to

$$\sum_{j=1}^{n} \lambda_j C(\boldsymbol{x}_i, \boldsymbol{x}_j) = C(\boldsymbol{x}, \boldsymbol{x}_i), \tag{8.94}$$

which is a system of $n$ equations with $n$ unknown weighting coefficients $\lambda_i$. We can see that the weights $\lambda_i$ will depend on the data set, the location $\boldsymbol{x}$ of kriging estimate, the proximity of the data to the location being kriged, and the clustering of the data.

If we want to obtain the estimate at a location, say, $\boldsymbol{x} = \boldsymbol{x}_0$, the above system of equations can be rewritten as the following matrix form

$$\begin{pmatrix} C(\boldsymbol{x}_1, \boldsymbol{x}_1) & C(\boldsymbol{x}_1, \boldsymbol{x}_2) & ... & C(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ C(\boldsymbol{x}_2, \boldsymbol{x}_1) & C(\boldsymbol{x}_2, \boldsymbol{x}_2) & ... & C(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\boldsymbol{x}_n, \boldsymbol{x}_1) & C(\boldsymbol{x}_n, \boldsymbol{x}_2) & ... & C(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} C(\boldsymbol{x}_0, \boldsymbol{x}_1) \\ C(\boldsymbol{x}_0, \boldsymbol{x}_2) \\ \vdots \\ C(\boldsymbol{x}_0, \boldsymbol{x}_n) \end{pmatrix}$$

or

$$\boldsymbol{C}\boldsymbol{\lambda} = \boldsymbol{C_0}, \tag{8.95}$$

where $\boldsymbol{C}$ is a $n \times n$ covariance matrix which can be calculated from all possible pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of the measured data. $\boldsymbol{C_0}$ is the covariance between the measured data and the location being estimated. $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)^T$ is the $n$-vector weights.

As the value of the location being estimated is not known, it is not straightforward to calculate $C(\boldsymbol{x}_0, \boldsymbol{x}_i)$ as it involves $\epsilon(\boldsymbol{x}_0)$. A possible (but inefficient) way is to use the iterations starting from an initial guess. There is a better way to do this, that is to use the data to derive or construct some form of error covariance function. The covariances are typically taken from a parametric covariance function of spatial arguments, and this function should be positive definite over certain range or distance. To make it easier from estimations, all parameters of the covariance function are explicitly specified in kriging. In the case when covariance is stationary, the covariance is given as an explicit function of the spatial separation between pairs of locations.

In order to count the spatial proximity of the data, a second-order stationary covariance function is usually used. The second-order stationary means that the mean $\mu_Z$ and variance $\sigma_Z^2$ of the errors are constants, and the covariances depend only on the distance $h$ between input data. As the closer data are more likely correlated than distant data locations, the covariance function should decrease as $h$ increases.

For a given data set $Z_i, (i = 1, 2, ..., n)$, the computed semivariogram $\gamma(h)$ could be scattered around such as that shown in Fig. 8.8. From Eq.(8.90), we know that

$$C(h) = C(0) - \gamma(h) = C_0 + C_1 - \gamma(h), \tag{8.96}$$

which can determine $C(h)$ once we know the semivariogram $\gamma(h)$. For simplicity, a best fit model is usually used as the idealisation of the semivariogram derived from real experimental data. The value $C_0$ and and $C_1$ and the shape of the model can be fitted using measured data. In order to fit a semivariogram model, the number of data points or sample size should be reasonably big $n > 100$, in practice, however, $n = 50$ to 250 are acceptably reliable. If the sample size $n < 50$, we still can try to use kriging to get estimate predictions for unobserved locations if there is no better alternative.

There are many semivariogram models and we will give four most popular models: linear, spherical, exponential and Gaussian.

The simplest semivariogram model is the linear model

$$\gamma(h) = \begin{cases} 0 & (h = 0) \\ C_0 + C_1 \frac{h}{a} & (0 < h < a) \\ C_0 + C_1 & (h \geq a) \end{cases}, \tag{8.97}$$

where $a$ is the range and $C_0$ is the nugget effect due to sample errors. $C_1$ is the partial sill and $C_0 + C_1$ is the sill which is the asymptotic

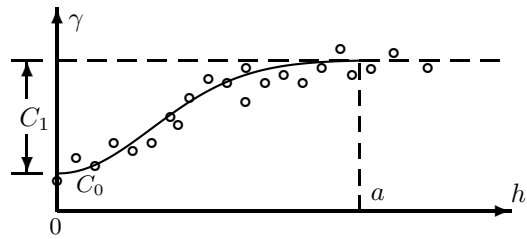



Figure 8.8: Semivariogram and idealisation.

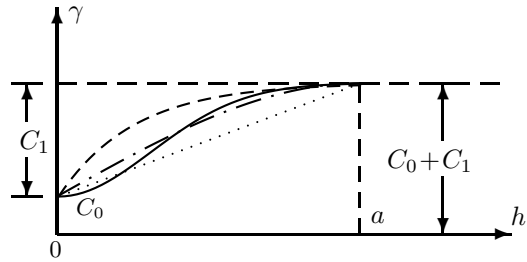



Figure 8.9: Semivariogram models: linear (dotted), spherical (dashed), exponential (dot-dashed), and Gaussian (solid).

value of $\gamma(h)$ as $h \gg a$, representing the variability in the absence of any spatial correlation. For the distance $h > a$, the sample values are no longer correlated.

A more elaborate model is the spherical model

$$\gamma(h) = \begin{cases} 0 & (h = 0) \\ C_0 + C_1\{\frac{1}{2}[\frac{3h}{a} - (\frac{h}{a})^3]\} & (0 < h < a) \\ C_0 + C_1 & (h \geq a) \end{cases} . \tag{8.98}$$

Exponential model is widely used and it has the following form

$$\gamma(h) = \begin{cases} 0 & (h = 0) \\ C_0 + C_1[1 - e^{-\beta|h|/a}] & (h > 0) \end{cases} , \tag{8.99}$$

where $\beta$ is a coefficient. In earth science literature, $\beta = 3$ is often used, while in engineering, $\beta = 1$ is used. A better and smoother model is the Gaussian model

$$\gamma(h) = \begin{cases} 0 & (h = 0) \\ C_0 + C_1[1 - e^{-\beta h^2/a^2}] & (h > 0) \end{cases} . \tag{8.100}$$

These four models are plotted in Fig. 8.9. As to which model should

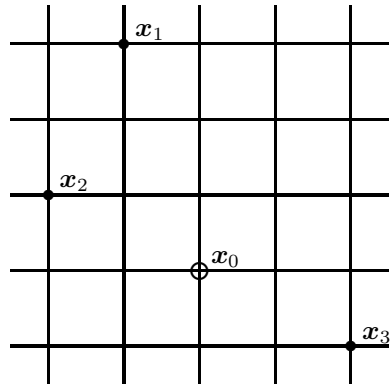



Figure 8.10: Kriging example: estimation of $Z(\boldsymbol{x}_0)$ at $\boldsymbol{x}_0 = (2, 1)$ from three observed data points: $Z_1$ at $\boldsymbol{x}_1$, $Z_2$ at $\boldsymbol{x}_2$, and $Z_3$ at $\boldsymbol{x}_3$.

be used, the choice depends on the problem and underlying random process.

---

**Example 8.7:** For any given three observed data points: $Z_1 = 15$ at $\boldsymbol{x}_1 = (1, 4)$, $Z_2 = 22$ at $\boldsymbol{x}_2 = (0, 2)$, and $Z_3 = 19$ at $\boldsymbol{x}_3 = (4, 0)$, let us now estimate the value $Z_0$ at the inaccessible location $\boldsymbol{x}_0(2, 1)$ as shown in Fig. 8.10. Since there are only three data points, it is not sufficiently accurate to fit the semivariogram model. Now we assume that the data points are drawn from a random process, and the semivariogram model is approximated as

$$\gamma(h) = 0.05 + 7[1 - e^{-3(h/5)^2}],$$

which corresponds to $C_0 = 0.05$, $C_1 = 7$, and $a = 5$. Then, we have

$$C_{ij} = C(\boldsymbol{x}_i, \boldsymbol{x}_j) = C_0 + C_1 - \gamma(h) = \begin{cases} C_0 + C_1 = 7.05 & (h = 0) \\ 7e^{-3(h/5)^2} & (h > 0) \end{cases}.$$

Using the notation $h_{ij} = |\boldsymbol{x}_i - \boldsymbol{x}_j|$, we have $h_{12} = |\boldsymbol{x}_1 - \boldsymbol{x}_2| = h_{21} = h_{03} = |\boldsymbol{x}_0 - \boldsymbol{x}_3| = h_{02} = \sqrt{5}$, and we get

$$C_{12} = C_{21} = C_{03} = C_{02} = 7e^{-3(\sqrt{5}/5)^2} = 3.84.$$

Similarly, $h_{13} = h_{31} = 5$ and $h_{23} = h_{32} = \sqrt{20}$, we have $C_{13} = C_{31} = 0.348$, $C_{23} = C_{32} = 0.635$. For $h_{ii} = 0$, we have $C_{11} = C_{22} = C_{33} = 9$. As $h_{01} = \sqrt{10}$, we have $C_{01} = 2.11$. Finally, we have

$$\begin{pmatrix} 7.05 & 3.84 & 0.348 \\ 3.84 & 7.05 & 0.635 \\ 0.348 & 0.635 & 7.05 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} 2.11 \\ 3.84 \\ 3.84 \end{pmatrix},$$

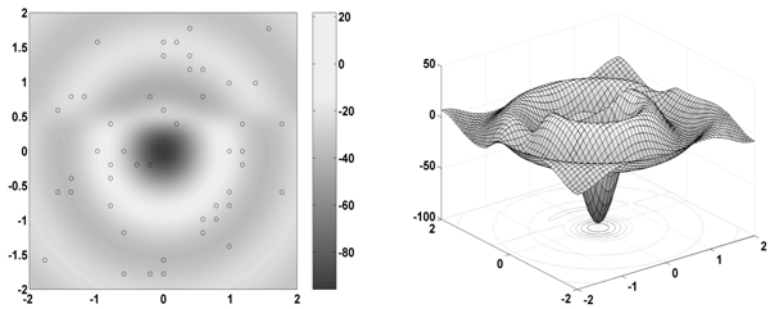



Figure 8.11: (a) Survey locations and contour predicted by ordinary kriging; (b) Gravity variations over the landscape.

whose solution is $\begin{pmatrix}\lambda_1 & \lambda_2 & \lambda_3\end{pmatrix} = \begin{pmatrix}0.003 & 0.498 & 0.499\end{pmatrix}$. We see that $\sum_{i=1}^{3} \lambda_i = 0.003 + 0.498 + 0.0499 = 1$. The estimated value of $Z_0$ is $Z_0 = \sum_{i=1}^{3} \lambda_i Z_i = 0.003 \times 15 + 0.498 \times 22 + 0.499 \times 19 = 20.5$.

Now let us look at a real application. A microgravity survey around an ancient underwater volcano was carried out so as to establish the topological variations of the volcano (there are better methods such as seismic profiling, but here we just want to demonstrate an application of kriging). There were 50 survey locations over an area of $2 \times 2$ km$^2$. The survey points are shown in Fig. 8.11 (marked with circles), and the gravity variations predicted by kriging are also shown. Here the unit of the relative change of gravity is milligal (or mGal=$10^{-5}$ m/s$^2$). The corresponding topological variations can be derived from the gravity data using appropriate inverse methods.

# References

Armstrong M., *Basic Linear Geostatistics*, Springer (1998).

Krige D. G., A statistical approach to some basic mine valuation problems on the Witwatersrand, *J. of Chem., Metal. Mining Soc. of South Africa*, **52**, 119-139 (1951).

Goodman R., *Teach Yourself Statistics*, London, (1957).

Kitanidis P. K., *Introduction to Geostatistics*, Cambridge University Press, (1997).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, Cambridge University Press, (2006).

# Part II

# Numerical Algorithms

# Chapter 9

# Numerical Integration

The beauty of mathematical modelling is that a closed-form solution or a simple formula will give tremendous insight into the mechanism and behaviour of the underlying real-world process. However, such closed-form solutions are rarely possible. In most cases, only approximate solutions are possible. Such approximations can be obtained by using simplified models or solving the mathematical models using approximation techniques. Alternatively, we can use numerical methods to get the approximate solutions. In this chapter, we will introduce the fundamentals of the numerical techniques, and in the following chapters we will study various commonly used methods in detail.

## 9.1 Root-Finding Algorithms

The essence of root-finding algorithms is to use iteration procedure to obtain the approximate (though sometimes quite accurate) solutions, starting from some initial guess solution. In fact, even ancient Babylonians knew how to find the square root of 2 using the iterative methods. From the numerical technique we learnt at school, we know we can numerically compute the square root of any real number $k$ ( so that $x = \sqrt{k}$) using the equation

$$x = \frac{1}{2}(x + \frac{k}{x}), \tag{9.1}$$

starting with a random guess, say, $x = 1$. The reason is that the above equation can be rearranged to get $x = \sqrt{k}$. In order to carry out the iteration, we use the notation $x_n$ for the value of $x$ at $n$-th iteration. Thus, equation (9.1) provides a way of calculating the estimate of $x$ at

$n+1$ (denoted as $x_{n+1}$). We have

$$x_{n+1} = \frac{1}{2}(x_n + \frac{k}{x_n}). \tag{9.2}$$

If we start from an initial value, say, $x_0 = 1$ at $n = 0$, we can do the iterations to meet the accuracy we want.

---

**Example 9.1:** To find $\sqrt{5}$, we have $k = 5$ with an initial guess $x_0 = 1$, and the first five iterations are as follows:

$$x_1 = \frac{1}{2}(x_0 + \frac{5}{x_0}) = 3,$$

$$x_2 = \frac{1}{2}(x_1 + \frac{5}{x_1}) \approx 2.333333333, \qquad x_3 \approx 2.238095238,$$

$$x_4 \approx 2.236068895, \;\; x_5 \approx 2.236067977.$$

We can see that $x_5$ after 5 iterations is very close to its true value $\sqrt{5} = 2.23606797749979...$, which shows that the iteration method is quite efficient.

---

The reason that this iterative process works is that the series $x_1$, $x_2$, ..., $x_n$ converges towards the true value $\sqrt{k}$ due to the fact that $x_{n+1}/x_n = \frac{1}{2}(1 + k/x_n^2) \to 1$ as $x_n \to \sqrt{k}$. However, a good choice of the initial value $x_0$ will speed up the convergence. Wrong choice of $x_0$ could make the iteration fail, for example, we cannot use $x_0 = 0$ as the initial guess. If we use $x_0 < 0$ as the initial guess, it will converge at $-\sqrt{k}$ which is another root. So a sensible choice should be an educated guess. At the initial step, if $x_0^2 < k$, $x_0$ is the lower bound and $k/x_0$ is upper bound. If $x_0^2 > k$, then $x_0$ is the upper bound and $k/x_0$ is the lower bound. For other iterations, the new bounds will be $x_n$ and $k/x_n$. In fact, the value $x_{n+1}$ is always between these two bounds $x_n$ and $k/x_n$, and the new estimate $x_{n+1}$ is thus the mean or average of the two bounds. This guarantees that the series converges towards the true value of $\sqrt{k}$. This method is similar to the bisection method below.

### 9.1.1 Bisection Method

The above-mentioned iteration method to find $x = \sqrt{k}$ is in fact equivalent to finding the solution or the root of the function $f(x) = x^2 - k = 0$. For any function $f(x)$ in the interval $[a, b]$, the root-finding bisection method works in the following way as shown in Fig. 9.1.

The iteration procedure starts with two initial guessed bounds $x_a$ (lower bound), and $x_b$ (upper bound) so that the true root $x = x_*$ lies

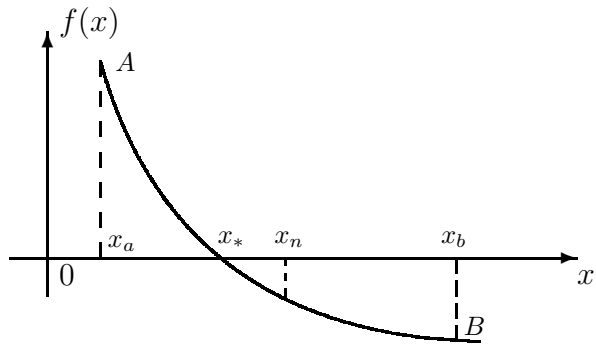



Figure 9.1: Bisection method for finding the root $x_*$ of $f(x_*) = 0$ between two bounds $x_a$ and $x_b$ in the domain $x \in [a, b]$.

between these two bounds. This requires that $f(x_a)$ and $f(x_b)$ have different signs. In our case shown in Fig. 9.1, $f(x_a) > 0$ and $f(x_b) < 0$, but $f(x_a)f(x_b) < 0$. The obvious choice is $x_a = a$ and $x_b = b$. The next estimate is just the midpoint of $A$ and $B$, and we have

$$x_n = \frac{1}{2}(x_a + x_b). \tag{9.3}$$

We then have to test the sign of $f(x_n)$. If $f(x_n) < 0$ (having the same sign as $f(x_b)$), we then update the new upper bound as $x_b = x_n$. If $f(x_n) > 0$ (having the same sign as $f(x_a)$), we update the new lower bound as $x_a = x_n$. In a special case when $f(x_n) = 0$, you have found the true root. The iterations continue in the same manner until a given accuracy is achieved or the prescribed number of iterations is reached.

---

**Example 9.2:** If we want to find $\sqrt{\pi}$, we have

$$f(x) = x^2 - \pi = 0.$$

We can use $x_a = 1$ and $x_b = 2$ since $\pi < 4$ (thus $\sqrt{\pi} < 2$). The first bisection point is

$$x_1 = \frac{1}{2}(x_a + x_b) = \frac{1}{2}(1 + 2) = 1.5.$$

Since $f(x_a) < 0$, $f(x_b) > 0$ and $f(x_1) = -0.8916 < 0$, we update the new lower bound $x_a = x_1 = 1.5$. The second bisection point is

$$x_2 = \frac{1}{2}(1.5 + 2) = 1.75,$$

and $f(x_2) = -0.0791 < 0$, so we update lower bound again $x_a = 1.75$. The third bisection point is

$$x_3 = \frac{1}{2}(1.75 + 2) = 1.875.$$

Since $f(x_3) = 0.374 > 0$, we now update the new upper bound $x_b = 1.875$. The fourth bisection point is

$$x_4 = \frac{1}{2}(1.75 + 1.875) = 1.8125.$$

It is within 2.5% of the true value of $\sqrt{\pi} \approx 1.7724538509$.

In general, the convergence of the bisection method is very slow, and Newton's method is a much better choice in most cases.

### 9.1.2 Newton's Method

Newton's method is a widely-used classic method for finding the zeros of a nonlinear univariate function of $f(x)$ on the interval $[a, b]$. It is also referred to as the Newton-Raphson method. At any given point $x_n$ shown in Fig. 9.2, we can approximate the function by a Taylor series

$$f(x_{n+1}) = f(x_n + \Delta x) \approx f(x_n) + f'(x_n)\Delta x, \tag{9.4}$$

where

$$\Delta x = x_{n+1} - x_n, \tag{9.5}$$

which leads to

$$x_{n+1} - x_n = \Delta x \approx \frac{f(x_{n+1}) - f(x_n)}{f'(x_n)}, \tag{9.6}$$

or

$$x_{n+1} \approx x_n + \frac{f(x_{n+1}) - f(x_n)}{f'(x_n)}. \tag{9.7}$$

Since we try to find an approximation to $f(x) = 0$ with $f(x_{n+1})$, we can use the approximation $f(x_{n+1}) \approx 0$ in the above expression. Thus we have the standard Newton iterative formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{9.8}$$

The iteration procedure starts from an initial guess value $x_0$ and continues until certain criteria are met. A good initial guess will use fewer steps, however, if there is no obvious initial good starting point, you

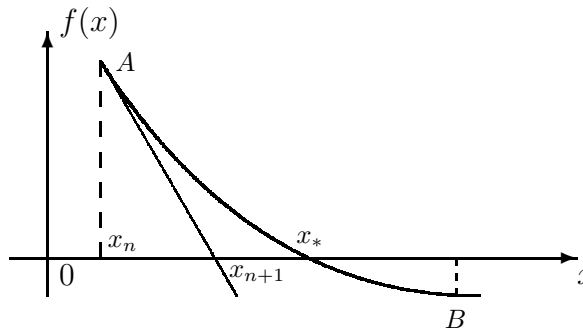



Figure 9.2: Newton's method of approximating the root $x_*$ by $x_{n+1}$ from the previous value $x_n$.

can start at any point on the interval $[a, b]$. But if the initial value is too far from the true zero, the iteration process may fail. So it is a good idea to limit the number of iterations.

---

**Example 9.3:** To find the root of

$$f(x) = x - e^{-x} = 0,$$

we use Newton's method starting from $x_0 = 1$. We know that

$$f'(x) = 1 + e^{-x},$$

and thus the iteration formula becomes

$$x_{n+1} = x_n - \frac{x_n - e^{-x_n}}{1 + e^{-x_n}}.$$

Since $x_0 = 1$, we have

$$x_1 = 1 - \frac{1 - e^{-1}}{1 + e^{-1}} \approx 0.5378828427,$$

and

$$x_2 \approx 0.5669869914, \qquad x_3 \approx 0.5671432859.$$

We can see that $x_3$ (only three iterations) is very close to the true root is $x_* \approx 0.5671432904$.

---

We have seen that Newton's method is very efficient and that is why it is so widely used. Therefore, we will implement this method

using both Matlab and Octave programs which are listed in Appendix B. Newton's method can be modified for solving optimisation problems because it is equivalent to finding the root of the first derivative $f'(\boldsymbol{x}) = 0$ once the objective function $f(\boldsymbol{x})$ is given.

### 9.1.3 Iteration Method

Sometimes we have to find roots of functions of multiple variables, and Newton's method can be extended to carry out such a task. For nonlinear multivariate functions

$$\boldsymbol{F}(\boldsymbol{x}) = [F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), ..., F_N(\boldsymbol{x})]^T, \tag{9.9}$$

where $\boldsymbol{x} = (x, y, ..., z)^T = (x_1, x_2, ..., x_p)^T$, an iteration method is usually needed to find the roots

$$\boldsymbol{F}(\boldsymbol{x}) = 0. \tag{9.10}$$

Newton-Raphson iteration procedure is widely used. We first approximate $\boldsymbol{F}(\boldsymbol{x})$ by a linear residual function $\mathbf{R}(\boldsymbol{x}; \boldsymbol{x}^n)$ in the neighbourhood of an existing approximation $\boldsymbol{x}^n$ to $\boldsymbol{x}$, and we have

$$\boldsymbol{R}(\boldsymbol{x}, \boldsymbol{x}^n) = \boldsymbol{F}(\boldsymbol{x}^n) + \mathbf{J}(\boldsymbol{x}^n)(\boldsymbol{x} - \boldsymbol{x}^n), \tag{9.11}$$

and

$$\mathbf{J}(\boldsymbol{x}) = \nabla \boldsymbol{F}, \tag{9.12}$$

where $\mathbf{J}$ is the Jacobian of $\boldsymbol{F}$. That is

$$\mathbf{J}_{ij} = \frac{\partial F_i}{\partial x_j}. \tag{9.13}$$

Here we have used the notation $\boldsymbol{x}^n$ for the vector $\boldsymbol{x}$ at the $n$-th iteration, which should not be confused with the power $\boldsymbol{u}^n$ or a vector $\boldsymbol{u}$. This might be confusing, but such notations are widely used in the literature of numerical analysis. An alternative (and better) notation is to denote $\boldsymbol{x}^n$ by $\boldsymbol{x}^{(n)}$, which shows the vector value at $n$-th iteration using a bracket. However, we will use both notations if no confusion arises.

To find the next approximation $\boldsymbol{x}^{n+1}$ from the current estimate $\boldsymbol{x}^n$, we have to try to satisfy $\boldsymbol{R}(\boldsymbol{x}^{n+1}, \mathbf{u}^n) = 0$, which is equivalent to solving a linear system with $\mathbf{J}$ as the coefficient matrix

$$\boldsymbol{x}^{n+1} = \boldsymbol{x}^n - \mathbf{J}^{-1}\mathbf{F}(\boldsymbol{x}^n), \tag{9.14}$$

under a given termination criterion $\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\| \leq \epsilon$. Iterations require an initial starting vector $\boldsymbol{x}^0$, which is often set to $\boldsymbol{x}^0 = 0$.

**Example 9.4:** To find the root of the system

$$x - e^{-y} = 0, \qquad x^2 - y = 0,$$

we first write it as

$$\boldsymbol{F}(\boldsymbol{x}) = \begin{pmatrix} x_1 - e^{-x_2} \\ x_1^2 - x_2 \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

The Newton-Raphson iteration formula becomes

$$\boldsymbol{x}^{n+1} = \boldsymbol{x}^n - \mathbf{J}^{-1}\boldsymbol{F}(\boldsymbol{x}^n),$$

where the Jacobian $\boldsymbol{J}$ is

$$\mathbf{J} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 1 & e^{-x_2} \\ 2x_1 & -1 \end{pmatrix},$$

whose inverse is

$$\boldsymbol{A} = \mathbf{J}^{-1} = \frac{1}{-1 - 2x_1 e^{-x_2}} \begin{pmatrix} -1 & -e^{-x_2} \\ -2x_1 & 1 \end{pmatrix}$$

$$= \frac{1}{1 + 2x_1 e^{-x_2}} \begin{pmatrix} 1 & e^{-x_2} \\ 2x_1 & -1 \end{pmatrix}.$$

Therefore, the iteration equation becomes

$$\boldsymbol{x}^{n+1} = \boldsymbol{x}^n - \boldsymbol{u}^n$$

where

$$\boldsymbol{u}^n = \boldsymbol{J}^{-1}\boldsymbol{F}(\boldsymbol{x}^n) = \frac{1}{1 + 2x_2 e^{-x_2}} \begin{pmatrix} 1 & e^{-x_2} \\ 2x_1 & -1 \end{pmatrix} \begin{pmatrix} x_1 - e^{-x_2} \\ x_1^2 - x_2 \end{pmatrix}$$

$$= \frac{1}{1 + 2x_1 e^{-x_2}} \begin{pmatrix} x_1 + (x_1^2 - 1 - x_2)e^{-x_2} \\ x_1^2 + x_2 - 2x_1 e^{-x_2} \end{pmatrix}.$$

If we start with the initial guess $\boldsymbol{x}^0 = (0, 0)^T$, we have the first estimate

$$\boldsymbol{x}^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

and the second iteration gives

$$\boldsymbol{x}^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0.33333 \\ -0.33333 \end{pmatrix} = \begin{pmatrix} 0.66667 \\ 0.3333 \end{pmatrix}.$$

If we continue this way, the third iteration gives

$$\boldsymbol{x}^3 = \boldsymbol{x}^2 - \begin{pmatrix} 0.01520796 \\ -0.09082847 \end{pmatrix} = \begin{pmatrix} 0.6514462 \\ 0.42415551 \end{pmatrix}.$$

Finally, the fourth iteration gives

$$\boldsymbol{x}^4 = \boldsymbol{x}^3 - \begin{pmatrix} -0.001472389 \\ -0.002145006 \end{pmatrix} = \begin{pmatrix} 0.65291859 \\ 0.4263005 \end{pmatrix}.$$

The true root occurs at $(0.6529186405, 0.4263027510)$, and we can see that even after four iterations, the estimates are very close to the true values.

## 9.2 Numerical Integration

An interesting feature of differentiations and integrations is that you can get the explicit expressions of derivatives of most functions and complicated expressions if they exist, while it is very difficult and sometimes impossible to express an integral in an explicit form, even for seemingly simple integrands. For example, the error function, widely used in engineering and sciences, is defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{9.15}$$

The integration of this simple integrand $\exp(-t^2)$ does not lead to any simple explicit expression, which is why it is often written as erf(), referred to as the error function. If we pick up a mathematical handbook, we know that $\operatorname{erf}(0) = 0$, and $\operatorname{erf}(\infty) = 1$, while

$$\operatorname{erf}(0.5) \approx 0.52049, \qquad \operatorname{erf}(1) \approx 0.84270. \tag{9.16}$$

If we want to calculate such integrals, numerical integration is the best alternative.

### 9.2.1 Trapezium Rule

Now if we want to numerically evaluate the following integral

$$\mathcal{I} = \int_a^b f(x)dx, \tag{9.17}$$

where $a$ and $b$ are fixed and finite. We know that the value of the integral is exactly the total area under the curve $y = f(x)$ between $a$

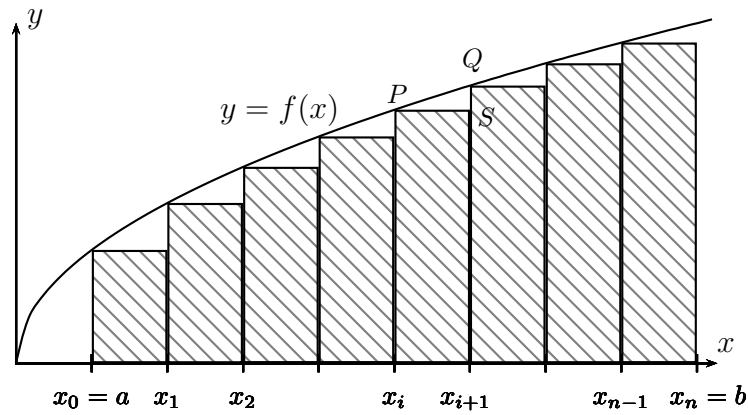



Figure 9.3: Numerical integration: $n$ thin strips to approximate the integral of a function $f(x)$.

and $b$. As both the integral and the area can be considered as the sum of the values over many small intervals, the simplest way of evaluating such numerical integration is to divide up the integral interval into $n$ equal small sections and split the area into $n$ thin strips so that $h \equiv \Delta x = (b-a)/n$, $x_0 = a$ and $x_i = ih+a (i = 1, 2, ..., n)$. The values of the functions at the dividing points $x_i$ are denoted as $y_i = f(x_i)$, and the value at the midpoint between $x_i$ and $x_{i+1}$ is labelled as $y_{i+1/2} = f_{i+1/2}$

$$y_{i+1/2} = f(x_{i+1/2}) = f_{i+1/2}, \qquad x_{i+1/2} = \frac{x_i + x_{i+1}}{2}. \tag{9.18}$$

The accuracy of such approximations depends on the number $n$ and the way to approximate the curve in each interval. Figure 9.3 shows such an interval $[x_i, x_{i+1}]$ which is exaggerated in the figure for clarity. The curve segment between $P$ and $Q$ is approximated by a straight line with a slope

$$\frac{\Delta y}{\Delta x} = \frac{f(x_{i+1}) - f(x_i)}{h}, \tag{9.19}$$

which approaches $f'(x_{i+1/2})$ at the midpoint point when $h \to 0$.

The trapezium (formed by $P$, $Q$, $x_{i+1}$, and $x_i$) is a better approximation than the rectangle ($P$, $S$, $x_{i+1}$ and $x_i$) because the former has an area

$$A_i = \frac{f(x_i) + f(x_{i+1})}{2} h, \tag{9.20}$$

which is close to the area

$$\mathcal{I}_i = \int_{x_i}^{x_{i+1}} f(x)dx, \tag{9.21}$$

under the curve in the small interval $x_i$ and $x_{i+1}$. If we use the area $A_i$ to approximate $\mathcal{I}_i$, we have the trapezium rule of numerical integration. Thus, the integral is simply the sum of all these small trapeziums, and we have

$$\begin{aligned}\mathcal{I} &\approx \frac{h}{2}[f_0 + 2(f_1 + f_2 + ... + f_{n-1}) + f_n] \\ &= h[f_1 + f_2 + ... + f_{n-1} + \frac{(f_0 + f_n)}{2}].\end{aligned} \tag{9.22}$$

From the Taylor series (2.37), we know that

$$\begin{aligned}\frac{f(x_i) + f(x_{i+1})}{2} &\approx \frac{1}{2}\Big\{[f(x_{i+1/2}) - \frac{h}{2}f'(x_{i+1/2}) + \frac{1}{2!}(\frac{h}{2})^2 f''(x_{i+1/2})] \\ &+[f(x_{i+1/2}) + \frac{h}{2}f'(x_{i+1/2}) + \frac{1}{2!}(\frac{h}{2})^2]f''(x_{i+1/2})\Big\} \\ &= f(x_{i+1/2}) + \frac{h^2}{8}f''(x_{i+1/2}).\end{aligned} \tag{9.23}$$

where $O(h^2 f'')$ means that the value is the order of $h^2 f''$, or $O(h^2) = Kh^2 f''$ where $K$ is a constant. Therefore, the error of the estimate of $\mathcal{I}$ is $h \times O(h^2 f'') = O(h^3 f'')$.

### 9.2.2 Order Notation

Now let us briefly introduce the order notations. Loosely speaking, for two functions $f(x)$ and $g(x)$, if

$$\frac{f(x)}{g(x)} \to K, \qquad x \to x_0, \tag{9.24}$$

where $K$ is a finite, non-zero limit, we write

$$f = O(g). \tag{9.25}$$

The big $O$ notation means that $f$ is asymptotically equivalent to the order of $g(x)$. If the limit is unity or $K = 1$, we say $f(x)$ is order of $g(x)$. In this special case, we write

$$f \sim g, \tag{9.26}$$

which is equivalent to $f/g \to 1$ and $g/f \to 1$ as $x \to x_0$. Obviously, $x_0$ can be any value, including 0 and $\infty$. The notation $\sim$ does not

necessarily mean $\approx$ in general, though they might give the same results, especially in the case when $x \to 0$ [for example, $\sin x \sim x$ and $\sin x \approx x$ if $x \to 0$].

When we say $f$ is order of 100 (or $f \sim 100$), this does not mean $f \approx 100$, but it can mean that $f$ is between about 50 and 150. The small $o$ notation is used if the limit tends to 0. That is

$$\frac{f}{g} \to 0, \qquad x \to x_0, \tag{9.27}$$

or

$$f = o(g). \tag{9.28}$$

If $g > 0$, $f = o(g)$ is equivalent to $f \ll g$. For example, for $\forall x \in \mathcal{R}$, we have $e^x \approx 1 + x + O(x^2) \approx 1 + x + \frac{x^2}{2} + o(x)$.

---

**Example 9.5:** Another classical example is Stirling's asymptotic series for factorials

$$n! \sim \sqrt{2\pi n}(\frac{n}{e})^n(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51480n^3} - ...). \tag{9.29}$$

This is a good example of asymptotic series. For standard power expansions, the error $R_k(h^k) \to 0$, but for an asymptotic series, the error of the truncated series $R_k$ decreases compared with the leading term [here $\sqrt{2\pi n}(n/e)^n$]. However, $R_n$ does not necessarily tend to zero. In fact,

$$R_2 = \frac{1}{12n} \cdot \sqrt{2\pi n}(n/e)^n,$$

is still very large as $R_2 \to \infty$ if $n \gg 1$. For example, for $n = 100$, we have $n! = 9.3326 \times 10^{157}$, while the leading approximation is $\sqrt{2\pi n}(n/e)^n = 9.3248 \times 10^{157}$. The difference between these two values is $7.7740 \times 10^{154}$, which is still very large, though three orders smaller than the leading approximation.

---

### 9.2.3 Simpson's Rule

The trapezium rule introduced earlier is just one of the simple and popular schemes for numerical integration with the error of $O(h^3 f'')$. If we want higher accuracy, we can either reduce $h$ or use a better approximation for $f(x)$). A small $h$ means a large $n$, which implies that we have to do the sum of many small sections, and it may increase the computational time.

On the other hand, we can use higher-order approximations for the curve. Instead of using straight lines or linear approximations for curve

segments, we can use parabolas or quadratic approximations. For any consecutive three points $x_{i-1}$, $x_i$ and $x_{i+1}$, we can construct a parabola in the form

$$f(x_i + t) = f_i + \alpha t + \beta t^2, \qquad t \in [-h, h]. \tag{9.30}$$

As this parabola must go through the three known points $(x_{i-1},\ f_{i-1})$ at $t = -h$, $(x_i, f_i)$ at $t = 0$ and $x_{i+1}, f_{i+1}$ at $t = h$, we have the following equations for $\alpha$ and $\beta$

$$f_{i-1} = f_i - \alpha h + \beta h^2, \tag{9.31}$$

and

$$f_{i+1} = f_i + \alpha h + \beta h^2, \tag{9.32}$$

which lead to

$$\alpha = \frac{f_{i+1} - f_{i-1}}{2h}, \qquad \beta = \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}. \tag{9.33}$$

We will see in later chapters that $\alpha$ is the centred approximation for the first derivative $f_i'$ and $\beta$ is the central difference scheme for the second derivative $f_i''$. Therefore, the integral from $x_{i-1}$ to $x_{i+1}$ can be approximated by

$$\mathcal{I}_i = \int_{x_{i-1}}^{x_{i+1}} f(x)dx \approx \int_{-h}^{h} [f_i + \alpha t + \beta t^2]dt = \frac{h}{3}[f_{i-1} + 4f_i + f_{i+1}], \tag{9.34}$$

where we have substituted the expressions for $\alpha$ and $\beta$. To ensure the whole interval $[a, b]$ can be divided up to form three-point approximations without any point left out, $n$ must be even. Therefore, the estimate of the integral becomes

$$\mathcal{I} \approx \frac{h}{3}[f_0 + 4(f_1 + f_3 + ... + f_{n-1}) + 2(f_2 + f_4 + ... + f_{n-2}) + f_n], \tag{9.35}$$

which is the standard Simpson's rule.

As the approximation for the function $f(x)$ is quadratic, an order higher than the linear form, the error estimate of Simpson's rule is thus $O(h^4)$ or $O(h^4 f'''')$ to be more specific. There are many variations of Simpson's rule with higher order accuracies such as $O(h^5 f^{(4)})$ and $O(h^7 f^{(6)})$.

---

**Example 9.6:** We know the exact value of the integral

$$I = \int_0^{\pi/2} \sin^2(x)dx = \frac{\pi}{4}.$$

Let us now estimate it using the Simpson rule with $n = 8$ and $h = (\pi/2 - 0)/8 = \pi/16$. We have

$$I \approx \frac{h}{3}[f_0 + 4(f_1 + f_3 + f_5 + f_7) + 2(f_2 + f_4 + f_6) + f_8].$$

Since $f_i = \sin^2(x_i) = \sin^2(i * h)$, we have $f_0 = 0$, $f_1 = 0.03806$, $f_2 = 0.14644$, $f_3 = 0.308658$, $f_4 = 0.5$, $f_5 = 0.69134$, $f_6 = 0.85355$, $f_7 = 0.86193$, and $f_8 = 1$. Now the integral estimate is

$$I \approx \frac{\pi}{48}[0 + 4 \times 2.00 + 2 \times 1.50 + 1] \approx 0.71994.$$

The true value is $\pi/4 = 0.78539$, so the error is about 9%. The order of the estimate is $O(h^4 f'''')$. Since $f'''' = -8\cos^2(x) + 8\sin^2(x)$ or $|f''''(0)| = 8$, so the error is $O(\pi^4/16^4 * 8) = O(0.01189)$. Thus, we can expect the estimate to be accurate only to the first decimal place.

---

From the example, we have seen that the accuracy of Simpson's rule is only $O(h^4 f'''')$, and such estimate of integral usually requires very small $h$ (or large $n$). This means the evaluations of the integrand at many points. Is there any way to go around this tedious slow process and evaluate the integral more accurately using fewer points of evaluation? The answer is yes, and the numerical technique is called the Gaussian integration or Gaussian quadrature.

## 9.3 Gaussian Integration

To get higher-order accuracy, we can use polynomials to construct various integration schemes. However, there is an easier way to do this. That is to use the Gauss-Legendre integration or simply Gaussian integration. Since any integral $\mathcal{I}$ with integration limits $a$ and $b$ can be transformed to an integral with limits $-1$ and $+1$ by using

$$\zeta = \frac{2(x-a)}{(b-a)} - 1, \tag{9.36}$$

so that

$$\mathcal{I} = \int_a^b g(x)dx = \frac{(b-a)}{2}\int_{-1}^{1} f(\zeta)d\zeta, \tag{9.37}$$

where we have used $dx = (b-a)d\zeta/2$. Therefore, we only have to study the integral

$$J = \int_{-1}^{1} f(\zeta)d\zeta. \tag{9.38}$$

The $n$ values of the function or $n$ integration points are given by a polynomial of $n-1$ degree. For equal spacing $h$, this numerical integration technique is often referred to as the Newton-Cotes quadrature

$$J = \int_{-1}^{1} f(d\zeta)d\zeta = \sum_{i=1}^{n} w_i f(\zeta_i), \tag{9.39}$$

where $w_i$ is the weighting coefficient attached to $f(\zeta_i)$. Such integral will have an error of $O(h^n)$. For example, $n = 2$ with equal weighting corresponds to the trapezium rule because

$$J = f_{-1} + f_1. \tag{9.40}$$

For the case of $n = 3$, we have

$$J = \frac{1}{3}[f_{-1} + 4f_0 + f_1], \tag{9.41}$$

which corresponds to Simpson's rule.

The numerical integration we use so far is carried out at equally-spaced points $x_0, x_1, ..., x_i, ..., x_n$, and these points are fixed *a priori*. There is no particular reason why we should use the equally-spaced points apart from the fact that it is easy and simple. In fact, we can use any sampling points or integration points as we wish to improve the accuracy of the estimate to the integral. If we use $n$ integration points $(\zeta_i, i = 1, 2, ..., n)$ with a polynomial of $2n-1$ degrees or Legendre polynomial $P_n(x)$, we now have $2n$ unknowns $f_i$ and $\zeta_i$. This means that we can easily construct quadrature formula, often called Gauss quadrature or Gaussian integration.

Mathematically, we have the Gauss quadrature

$$J = \int_{-1}^{1} f(\zeta)d\zeta = \sum_{i=1}^{n} w_i f(\zeta_i), \tag{9.42}$$

where $\zeta_i$ is determined by the zeros of the Legendre polynomial $P_n(\zeta_i) = 0$ and the weighting coefficient is given by

$$w_i = \frac{2}{(1-\zeta_i^2)[P_n'(\zeta_i)]^2}. \tag{9.43}$$

The error of this quadrature is of the order $O(h^{2n})$. The proof of this formulation is beyond the scope of this book. Readers can find the proof in more advanced mathematical books.

Briefly speaking, Legendre polynomials are obtained by the following generating function or Rodrigue's formula

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n(x^2-1)^n}{dx^n}. \tag{9.44}$$

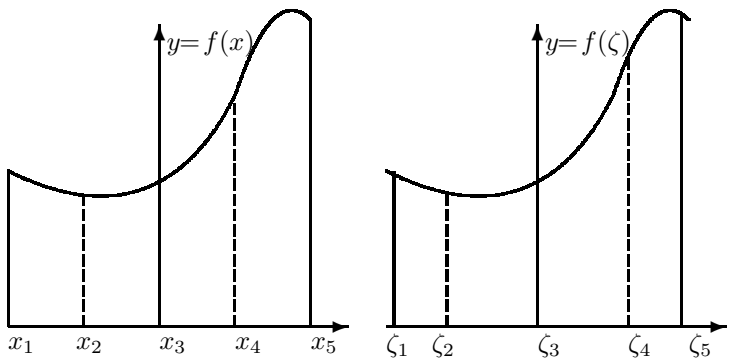



Figure 9.4: Five integration points: a) Equally spaced with $h = 1/4$; and b) Gauss points with $|\zeta_1 - \zeta_3| \approx 0.90617$ and $|\zeta_2 - \zeta_3| \approx 0.53847$.

For example,

$$P_0(x) = 0, \;\; P_1(x) = x, \;\; P_2(x) = \frac{1}{2}(3x^2 - 1), \tag{9.45}$$

and

$$P_3(x) = \frac{1}{2}(5x^3 - 3x), \;\; P_4 = \frac{1}{8}(3 - 30x^2 + 35x^4). \tag{9.46}$$

For both Newton-Cotes quadrature and Gauss quadrature, Figure 9.4 shows their difference and similarity.

The computation of locations $\zeta_i$ of the Gaussian integration points and weighting coefficients $w_i$ is complicated, though straightforward once we know the Legendre polynomials. For example, $n = 2$, we have

$$P_2(\zeta) = \frac{1}{2}(3\zeta^2 - 1) = 0, \tag{9.47}$$

which has two solutions

$$\zeta_{\pm 1} = \pm\sqrt{\frac{1}{3}} \approx \pm 0.5774. \tag{9.48}$$

Since $P_2'(\zeta) = 3\zeta$, we have

$$w_1 = \frac{2}{(1 - \zeta_1^2)(3\zeta_1)^2} = \frac{2}{(1 - (\sqrt{1/3})^2) * (3\sqrt{1/3})^2} = 1 = w_{-1}. \tag{9.49}$$

The coefficients $w_i$ and the integration points are usually listed in tables for various values of $n$.

For multiple integrals

$$J = \int_{-1}^{1}\int_{-1}^{1} f(\zeta,\eta)d\zeta d\eta, \tag{9.50}$$

these Gaussian quadrature can easily be extended by evaluating the integral with $\eta$ being kept constant first, then evaluating the outer integral. We have

$$J = \int_{-1}^{1}[\int_{-1}^{n} f(\zeta,\eta)d\zeta]d\eta = \int_{-1}^{1}\sum_{i=1}^{n} w_i f(\zeta_i,\eta)d\eta$$

$$= \sum_{i=1}^{n} w_i \int_{-1}^{1} f(\zeta_i,\eta)d\eta = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j f(\zeta_i,\eta_j), \tag{9.51}$$

where we have used

$$\int_{-1}^{1} f(\zeta_i,\eta)d\eta = \sum_{j=1}^{n} w_j f(\zeta_i,\eta_j). \tag{9.52}$$

---

**Example 9.7:** To evaluate the integral

$$I = \frac{2}{\sqrt{\pi}}\int_{-1}^{1} e^{-x^2}dx = \int_{-1}^{1} f(x)dx, \quad f(x) = \frac{2}{\sqrt{\pi}}e^{-x^2}.$$

We know that its exact value from (9.15) is

$$I = 2\,\mathrm{erf}(1) = 1.685401585899...$$

Let us now estimate it using Simpson's rule for three point integration at $x_{-1} = -1$, $x_0 = 0$ and $x_1 = 1$, and we have

$$I \approx \frac{1}{3}(f_{-1} + 4f_0 + f_1)$$

$$\approx \frac{2}{3\sqrt{\pi}}[e^{-(-1)^2} + 4\times 1 + e^{-(1)^2}] \approx 1.7812,$$

which differs from its exact value by about $5.6\%$.

If we use the 3-point Gauss quadrature at $x_{\pm 1} = \pm\sqrt{\frac{3}{5}}$ and $x_0 = 0$ with weighting coefficients $w_{\pm 1} = \frac{5}{9}$ and $w_0 = \frac{8}{9}$, we have

$$I \approx \sum_{i=-1}^{1} w_i f(x_i)$$

$$\approx \frac{2}{\sqrt{\pi}}[\frac{5}{9}e^{-(-\sqrt{3/5})^2} + \frac{8}{9} \times 1 + \frac{5}{9}e^{-(\sqrt{3/5})^2}] \approx 1.6911,$$

which is a better approximation than 1.7812. In fact, the error of Gauss quadrature is just $(1.6911 - 1.6854)/1.6911 \approx 0.3\%$. This higher accuracy is why the Gaussian quadrature is so widely used.

We will implement the 7-point Gaussian integration using both Matlab and Octave and the programs will be provided in Appendix B. For triple integrals and other integrals, the Gauss quadrature can be constructed in a similar way. We will see more numerical techniques in the rest of the book.

## 9.4 Optimisation

Optimisation is everywhere. Almost routinely we have to minimise the cost, and maximise the efficiency whether it is business, or planning your holidays or anything else. Mathematical programming is the study of such planning procedure using mathematical tools. Nowadays, computer simulations become an indispensable tool for implementing various optimisation techniques. Here we will only briefly touch on some of the basic ideas used in optimisation.

### 9.4.1 Unconstrained Optimisation

The simplest optimisation without any constraints is probably the search of the maxima or minima of a function $f(x)$. This requires to find the root of the first derivatives or the stationary condition

$$f'(x) = 0. \tag{9.53}$$

However, the stationary condition $f'(x) = 0$ is just a necessary condition, but it is not a sufficient condition. If $f'(x_*) = 0$ and $f''(x_*) > 0$, it is a local minimum. Conversely, if $f'(x_*) = 0$ and $f''(x_*) < 0$, then it is a local maximum. However, if $f'(x_*) = 0$ but $f''(x)$ is indefinite (both positive and negative) when $x \to x_*$, then $x_*$ corresponds to a saddle point. For example, $f(x) = x^3$ has a saddle point $x_* = 0$ because $f'(0) = 0$ but $f''$ changes sign from $f''(0+) > 0$ to $f''(0-) < 0$.

**Example 9.8:** For example, in order to find the maximum or minimum of a univariate function $f(x)$

$$f(x) = xe^{-x^2}, \qquad -\infty < x < \infty, \tag{9.54}$$

we have to find first the stationary point $x_*$ when the first derivative $f'(x)$ is zero. That is

$$\frac{df(x_*)}{dx_*} = e^{-x_*^2} - 2x_*^2 e^{-x_*^2} = 0. \tag{9.55}$$

Since $\exp(-x_*^2) \neq 0$, we have

$$x_* = \pm\frac{\sqrt{2}}{2}. \tag{9.56}$$

From the basic calculus we know that the maximum requires $f''(x_*) \leq 0$ while minimum requires $f''(x_*) \geq 0$. At $x_* = \sqrt{2}/2$, we have

$$f''(x_*) = (4x_*^2 - 6)x_* e^{-x_*^2} = -2\sqrt{2}e^{-1/2} < 0, \tag{9.57}$$

so it corresponds to a maximum $f(x_*) = \frac{1}{2}e^{-1/2}$. Similarly, at $x_* = -\sqrt{2}/2$, $f''(x_*) = 2\sqrt{2}e^{-1/2} > 0$, we have a minimum $f(x_*) = -\frac{1}{2}e^{-1/2}$.

Since a maximum of a function $f(x)$ can be converted into a minimum of $A - f(x)$ where $A$ is a large positive number, most optimisation problems are more conveniently expressed in terms of minima. For example, we know the maximum of $f(x) = e^{-x^2}, x \in (-\infty, \infty)$ is 1 at $x_* = 0$. This problem can be converted to a minimum problem $1 - f(x)$ or $10 - f(x)$ or any $A - f(x)$ where $A > 1$.

### 9.4.2 Newton's Method

The Newton's method can be modified for solving optimisation problems because it is equivalent to finding the root of the first derivative $f'(\boldsymbol{x})$ based on the stationary conditions once the objective function $f(\boldsymbol{x})$ is given. For a given function $f(\boldsymbol{x})$ which is continuously differentiable, we have the Taylor expansion about a known point $\boldsymbol{x} = \boldsymbol{x}_n$ and $\Delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}_n$

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_n) + (\nabla f(\boldsymbol{x}_n))^T \Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^T \nabla^2 f(\boldsymbol{x}_n)\Delta\boldsymbol{x} + ..., \tag{9.58}$$

which is written in terms of quadratic forms. This is minimised near a critical point when $\Delta\boldsymbol{x}$ is the solution of the following linear equation

$$\nabla f(\boldsymbol{x}_n) + \nabla^2 f(\boldsymbol{x}_n)\Delta\boldsymbol{x} = 0. \tag{9.59}$$

This leads to

$$\boldsymbol{x} = \boldsymbol{x}_n - G^{-1}\nabla f(\boldsymbol{x}_n), \tag{9.60}$$

where $G = \nabla^2 f(\boldsymbol{x}_n)$ is the Hessian matrix which is given by

$$G(\boldsymbol{x}) \equiv \nabla^2 f(\boldsymbol{x}) \equiv \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{pmatrix}, \tag{9.61}$$

which is symmetric due to the fact that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}. \tag{9.62}$$

If the iteration procedure starts from the initial vector $\boldsymbol{x}^{(0)}$ (usually a guessed point in the domain), Newton's iteration formula for the $n$th iteration is

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - G^{-1}(\boldsymbol{x}^{(n)}) f(\boldsymbol{x}^{(n)}). \tag{9.63}$$

It is worth pointing out that if $f(\boldsymbol{x})$ is quadratic, then the solution can be found exactly in a single step. However, this method is not efficient for non-quadratic functions.

In order to speed up the convergence, we can use a smaller step size $\alpha \in (0, 1]$ so that we have modified Newton's method

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \alpha G^{-1}(\boldsymbol{x}^{(n)}) f(\boldsymbol{x}^{(n)}). \tag{9.64}$$

It might sometimes be time-consuming to calculate the Hessian matrix for second derivatives. A good alternative is to use an identity matrix $G = \boldsymbol{I}$ so that $G^{-1} = \boldsymbol{I}$, and we have the quasi-Newton method

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \alpha \boldsymbol{I} \nabla f(\boldsymbol{x}^{(n)}), \tag{9.65}$$

which is essentially the steepest descent method.

### 9.4.3 Steepest Descent Method

The essence of this method is to find the lowest possible objective function $f(\boldsymbol{x})$ from the current point $\boldsymbol{x}^{(n)}$. From the Taylor expansion of $f(\boldsymbol{x})$ about $\boldsymbol{x}^{(n)}$, we have

$$f(\boldsymbol{x}^{(n+1)}) = f(\boldsymbol{x}^{(n)} + \Delta \boldsymbol{s}) \approx f(\boldsymbol{x}^{(n)} + (\nabla f(\boldsymbol{x}^{(n)}))^T \Delta \boldsymbol{s}, \tag{9.66}$$

where $\Delta \boldsymbol{s} = \boldsymbol{x}^{(n+1)} - \boldsymbol{x}^{(n)}$ is the increment vector. Since we try to find a lower (better) approximation to the objective function, it requires that the second term on the right hand is negative. So

$$f(\boldsymbol{x}^{(n)} + \Delta \boldsymbol{s}) - f(\boldsymbol{x}^{(n)}) = (\nabla f)^T \Delta \boldsymbol{s} < 0. \tag{9.67}$$

From vector analysis, we know the inner product $\boldsymbol{u}^T\boldsymbol{v}$ of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is largest when they are parallel but in opposite directions. Therefore, $(\nabla f)^T\Delta\boldsymbol{s}$ becomes the smallest when

$$\Delta\boldsymbol{s} = -\alpha\nabla f(\boldsymbol{x}^{(n)}), \tag{9.68}$$

where $\alpha > 0$ is the step size. This is the case when the direction $\Delta\boldsymbol{s}$ is along the steepest descent in the negative gradient direction. This method is often referred to as the 'hill-climbing' in the case of finding maxima.

The choice of the step size $\alpha$ is very important. A very small step size means slow movement towards the local minimum, while a large step may overshoot and subsequently makes it move far away from the local minimum. Therefore, the step size $\alpha = \alpha^{(n)}$ should be different at each iteration step and should be chosen so that it minimises the objective function $f(\boldsymbol{x}^{(n+1)}) = f(\boldsymbol{x}^{(n)}, \alpha^{(n)})$. Therefore, the steepest descent method can be written as

$$f(\boldsymbol{x}^{(n+1)}) = f(\boldsymbol{x}^{(n)}) - \alpha^{(n)}(\nabla f(\boldsymbol{x}^{(n)}))^T\nabla f(\boldsymbol{x}^{(n)}). \tag{9.69}$$

In each iteration, the gradient and step size will be calculated. Again, a good initial guess of both the starting point and the step size is useful.

---

**Example 9.9:** Let us minimise the function

$$f(x_1, x_2) = 10x_1^2 + 5x_1x_2 + 10(x_2 - 3)^2,$$

where

$$(x_1, x_2) = [-10, 10] \times [-15, 15],$$

using the steepest descent method starting with the initial $\boldsymbol{x}^{(0)} = (10, 15)^T$. We know that the gradient

$$\nabla f = (20x_1 + 5x_2, 5x_1 + 20x_2 - 60)^T,$$

therefore

$$\nabla f(\boldsymbol{x}^{(0)}) = (275, 290)^T.$$

In the first iteration, we have

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \alpha_0\begin{pmatrix}275\\290\end{pmatrix}.$$

The step size $\alpha_0$ should be chosen such that $f(\boldsymbol{x}^{(1)})$ is at the minimum, which means that

$$f(\alpha_0) = 10(10-275\alpha_0)^2 + 5(10-275\alpha_0)(15-290\alpha_0) + 10(12-290\alpha_0)^2,$$

should be minimised. This becomes an optimisation problem for a single independent variable $\alpha_0$. All the techniques for univariate optimisation problems such as Newton's method can be used to find $\alpha_0$. We can also obtain the solution by setting

$$\frac{df}{d\alpha_0} = -159725 + 3992000\alpha_0 = 0,$$

whose solution is $\alpha_0 \approx 0.04001$.

At the second step, we have

$$\nabla f(\boldsymbol{x}^{(1)}) = (-3.078, 2.919)^T, \;\; \boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \alpha_1 \begin{pmatrix} -3.078 \\ 2.919 \end{pmatrix}.$$

The minimisation of $f(\alpha_1)$ gives $\alpha_1 \approx 0.066$, and the new location is

$$\boldsymbol{x}^{(2)} \approx (-0.797, 3.202)^T.$$

At the third iteration, we have

$$\nabla f(\boldsymbol{x}^{(2)}) = (0.060, 0.064)^T, \;\; \boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} - \alpha_2 \begin{pmatrix} 0.060 \\ 0.064 \end{pmatrix}.$$

The minimisation of $f(\alpha_2)$ leads to $\alpha_2 \approx 0.040$, and thus

$$\boldsymbol{x}^{(3)} \approx (-0.8000299, 3.20029)^T.$$

Then, the iterations continue until a prescribed tolerance is met.

From calculus, we can set the first partial derivatives equal to zero

$$\frac{\partial f}{\partial x_1} = 20x_1 + 5x_2 = 0, \;\; \frac{\partial f}{\partial x_2} = 5x_1 + 20x_2 - 60 = 0,$$

we know that the minimum occurs exactly at

$$\boldsymbol{x}_* = (-4/5, 16/5)^T = (-0.8, 3.2)^T.$$

We see that the steepest descent method gives almost the exact solution after only 3 iterations.

---

In finding the step size $\alpha_n$ in the above steepest descent method, we have used the stationary condition $df(\alpha_n)/d\alpha_n = 0$. Well, you may say that if we use this stationary condition for $f(\alpha_0)$, why not use the same method to get the minimum point of $f(\boldsymbol{x})$ in the first place. There are two reasons here. The first reason is that this is a simple example for demonstrating how the steepest descent method works. The second reason is that even for complicated multiple variables $f(x_1, ..., x_p)$ (say

$p = 500$), $f(\alpha_n)$ at any step $n$ is still a univariate function, and the optimisation of such $f(\alpha_n)$ is much simpler compared with the original multivariate problem.

It is worth pointing out that in our example, the convergence from the second iteration to the third iteration is slow. In fact, the steepest descent is typically slow once the local minimisation is near. This is because near the local minimisation the gradient is nearly zero, and thus the rate of descent is also slow. If high accuracy is needed near the local minimum, other local search methods should be used.

### 9.4.4 Constrained Optimisation

All optimisation problems can in general be expressed as nonlinearly constrained optimisation problems, often written in the following generic form

$$\begin{aligned} &\underset{\boldsymbol{x}\in\Re^n}{\text{maximise/minimise}} f(\boldsymbol{x}), \\ &\text{subject to } \phi_j(\boldsymbol{x}) = 0, \quad (j = 1, 2, ..., M), \\ &\psi_k(\boldsymbol{x}) \geq 0, \quad (k = 1, ..., N), \end{aligned} \tag{9.70}$$

where $f(\boldsymbol{x})$ is the objective function or cost function. $\phi_i(\boldsymbol{x})$ are constraints in terms of $M$ equalities, and $\psi_j(\boldsymbol{x})$ are constraints written as $N$ inequalities. If the constraints are all linear, it becomes a linearly constrained problem.

If we want to minimise a function $f(\boldsymbol{x})$

$$\underset{\boldsymbol{x}\in\Re^n}{\text{minimise}} f(\boldsymbol{x}), \tag{9.71}$$

subject to the following nonlinear equality constraint

$$g(\boldsymbol{x}) = 0, \tag{9.72}$$

then we can combine the objective function $f(x)$ with the equality to form a new function, called the Lagrangian

$$\Pi = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}), \tag{9.73}$$

where $\lambda$ is the Lagrange multiplier, which is an unknown scalar to be determined. This essentially converts the constrained problem into an unconstrained problem for $\Pi(\boldsymbol{x})$. If we have $m$ equalities,

$$g_j(\boldsymbol{x}) = 0, \qquad (j = 1, ..., m), \tag{9.74}$$

then we need $M$ Lagrange multipliers $\lambda_j (j = 1, ..., m)$. We have

$$\Pi(x, \lambda_j) = f(\boldsymbol{x}) + \sum_{j=1}^{m} \lambda_j g_j(\boldsymbol{x}). \tag{9.75}$$

The requirement of stationary conditions leads to

$$\frac{\partial \Pi}{\partial x_i} = \frac{\partial f}{\partial x_i} + \sum_{j=1}^{m} \lambda_j \frac{\partial g_j}{\partial x_i}, \quad (i = 1, ..., p), \tag{9.76}$$

and

$$\frac{\partial \Pi}{\partial \lambda_j} = g_j = 0, \quad (j = 1, ..., m). \tag{9.77}$$

These $m + p$ equations will determine the $p$-component of $\boldsymbol{x}$ and $m$ Lagrange multipliers. As

$$\frac{\partial \Pi}{\partial g_j} = \lambda_j,$$

we can consider $\lambda_j$ as the rate of the change of the quantity $\Pi$ as a functional of $g_j$.

---

**Example 9.10:** To solve the optimisation problem

$$\underset{(x,y)\in\Re^2}{\text{maximise}} f(x, y) = xy^2,$$

subject to the condition $g(x, y) = x^2 + y^2 - 1 = 0$. We define

$$\Pi = f(x, y) + \lambda g(x, y) = xy^2 + \lambda(x^2 + y^2 - 1).$$

The stationary conditions become

$$\frac{\partial \Pi}{\partial x} = y^2 + 2\lambda x = 0, \qquad \frac{\partial \Pi}{\partial y} = 2xy + 2\lambda y = 0,$$

and

$$\frac{\partial \Pi}{\partial \lambda} = x^2 + y^2 - 1 = 0.$$

The condition $xy + \lambda y = 0$ implies that $y = 0$ or $\lambda = -x$. The case of $y = 0$ can be eliminated as it leads to $x = 0$ from $y^2 + 2\lambda x = 0$, which does not satisfy the last condition $x^2 + y^2 = 1$. Therefore, the only valid solution is $\lambda = -x$. From the first stationary condition, we have

$$y^2 - 2x^2 = 0, \text{ or } y^2 = 2x^2.$$

Substituting this into the third stationary condition, we have

$$x^2 - 2x^2 - 1 = 0,$$

which gives

$$x = \pm 1.$$

So we have four stationary points

$$P_1(1, \sqrt{2}),\ P_2(1, -\sqrt{2}),\ P_3(-1, \sqrt{2}),\ P_4(-1, -\sqrt{(2)}).$$

The values of function $f(x, y)$ at these four points are

$$f(P_1) = 2, f(P_2) = 2, f(P_3) = -2, f(P_4) = -2.$$

Thus, the function reaches its maxima at $(1, \sqrt{2})$ and $(1, -\sqrt{2})$. The Lagrange multiplier for this case is $\lambda = -1$.

Mathematical optimisation itself has vast literature. There are other important methods such as linear programming, simplex method, Hooke-Jeeves pattern search, metaheuristic methods, bioinspired algorithms, and nonlinear programming.

# References

Gill P. E., Murray W. and Wright M. H., *Practical optimisation*, Academic Press, (1981).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Kuhn H. W. and Tucker A. W., Nonlinear programming, *Proc. 2nd Berkeley Symposium*, pp. 481-492, University of California Press, (1951).

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, 3rd Edition, Cambridge University Press, (2006).

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

Sawaragi Y., Nakayama H. and Tanino T., *Theory of Multiobjective Optimisation*, Academic Press, (1985).

# Chapter 10

# Finite Difference Method

The finite difference method is one of the most popular methods that are used commonly in computer simulations. It has the advantage of simplicity and clarity, especially in 1-D configuration and other cases with regular geometry. The finite difference method essentially transforms a differential equation into a coupled set of algebraic equations by replacing the continuous derivatives with finite difference approximations on a grid of mesh or node points that spans the domain of interest based on the Taylor series expansions. In general, the boundary conditions and boundary nodes need special treatment.

## 10.1 Integration of ODEs

The second-order or higher-order ordinary differential equations can be written as a first-order system of ODEs. Since the technique for solving a system is essentially the same as that for solving a single equation

$$\frac{dy}{dx} = f(x, y), \tag{10.1}$$

we shall focus on the first-order equation in the rest of this section. In principle, the solution can be obtained by direct integration,

$$y(x) = y_0 + \int_{x_0}^{x} f(x, y(x))dx, \tag{10.2}$$

but in practice it is usually impossible to do the integration analytically as it requires the solution of $y(x)$ to evaluate the right-hand side. Thus, some approximations shall be utilised. Numerical integration

is the most common technique for obtaining approximate solutions. There are various integration schemes with different orders of accuracy and convergent rates. These schemes include the simple Euler scheme, Runge-Kutta method, Relaxation method, and many others.

### 10.1.1 Euler Scheme

Using the notation $h = \Delta x = x_{n+1} - x_n$, $y_n = y(x_n)$, $x_n = x_0 + n\Delta x$ $(n = 0, 1, 2, ..., N)$, and $' = d/dx$ for convenience, the explicit Euler scheme can simply be written as

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y)dx \approx y_n + hf(x_n, y_n). \tag{10.3}$$

This is a forward difference method as it is equivalent to the approximation of the first derivative

$$y'_n = \frac{y_{n+1} - y_n}{\Delta x}. \tag{10.4}$$

The order of accuracy can be estimated using the Taylor expansion

$$y_{n+1} = y_n + hy'|_n + \frac{h^2}{2}y''|_n + ... \approx y_n + hf(x_n, y_n) + O(h^2). \tag{10.5}$$

Thus, the Euler method is first-order accurate.

For any numerical algorithms, the algorithm must be stable in order to reach convergent solutions. Thus, stability is an important issue in numerical analysis. Defining $\delta y$ as the discrepancy between the actual numerical solution and the true solution of the Euler finite difference equation, we have

$$\delta y_{n+1} = [1 + hf'(y)] = \xi \delta y_n. \tag{10.6}$$

In order to avoid the discrepancy to grow, it requires the following stability condition $|\xi| \leq 1$.

---

**Example 10.1:** For the ordinary differential equation

$$y'(x) + 2xy(x) + y^2(x) = 0,$$

with the initial condition $y(0) = 1$. Its analytical solution (see Example. 4.1) is

$$y(x) = \frac{2e^{-x^2}}{(\sqrt{\pi}\text{erf}(x) + 2)}.$$

Now let us solve it by the Euler scheme. Now we have

$$f(x) = -2xy(x) - y^2(x),$$

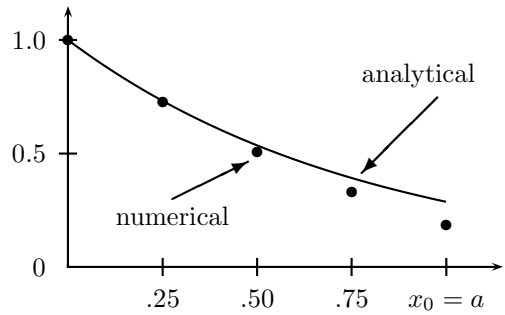



Figure 10.1: Comparison of numerical results with analytical solution.

and

$$y_{n+1} = y_n + hf(x_n, y_n).$$

Using $h = \Delta x = 0.25$, $x_0 = 0$ and $y_0(x_0) = 1$, we have

$$y_1(0.25) = 1 + hf(x_0, y_0) \approx 1 + 0.25 \times (-1) = 0.75,$$

$$y_2(0.5) = 1 + hf(x_1, y_1) \approx 0.75 + 0.25 \times (-0.75) = 0.5156.$$

Similarly,

$$y_3(0.75) \approx 0.3203, \qquad y_4(1.0) \approx 0.1745.$$

The analytical value at $x = 1$ is

$$y_* = \frac{2e^{-1^2}}{(\sqrt{\pi}\mathrm{erf}(1) + 2)} \approx 0.210599.$$

We can see that $y_4$ underestimates the true value by about 17% (see Fig. 10.1). As $h = 0.25$, so the error will be in the order of $O(h)$, so about 25%. Indeed, this simple integration scheme is not so accurate.

The stability restricts the size of interval $h$, which is usually small. One alternative that can use larger $h$ is the implicit Euler scheme, and this scheme approximates the derivative by a backward difference $y'_n = (y_n - y_{n-1})/h$ and the right-hand side of Eq.(10.2) is evaluated at the new $y_{n+1}$ location. Now the scheme can be written as

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}). \tag{10.7}$$

The stability condition becomes

$$\delta y_{n+1} = \xi \delta y_n = \frac{\delta y_n}{1 - hf'(y)}, \tag{10.8}$$

which is always stable if $f'(y) = \frac{\partial f}{\partial y} \leq 0$. This means that any step size is acceptable. However, the step size cannot be too large as the accuracy reduces as the step size increases. Another practical issue is that, for most problems such as nonlinear ODEs, the evaluation of $y'$ and $f'(y)$ requires the value of $y_{n+1}$ which is unknown. Thus, an iteration procedure is needed to march to a new value $y_{n+1}$, and the iteration starts with a guess value which is usually taken to be zero for most cases. The implicit scheme generally gives better stability.

### 10.1.2 Leap-Frog Method

The Leap-frog scheme is the central difference

$$y'_n = \frac{y_{n+1} - y_{n-1}}{2\Delta x}, \tag{10.9}$$

which leads to

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n). \tag{10.10}$$

The central difference method is second-order accurate. In a similar way as Eq.(10.6), the leap-frog method becomes

$$\delta y_{n+1} = \delta y_{n-1} + 2hf'(y)\delta y_n, \tag{10.11}$$

or

$$\delta y_{n+1} = \xi^2 \delta y_{n-1}, \tag{10.12}$$

where $\xi^2 = 1 + 2hf'(y)\xi$. This scheme is stable only if $|\xi| \leq 1$, and a special case is $|\xi| = 1$ when $f'(y)$ is purely imaginary. Therefore, the central scheme is not necessarily a better scheme than the forward scheme.

### 10.1.3 Runge-Kutta Method

We have so far seen that stability of the Euler method and the central difference method is limited. The Runge-Kutta method uses a trial step to the midpoint of the interval by central difference and combines with the forward difference at two steps

$$\hat{y}_{n+1/2} = y_n + \frac{h}{2} f(x_n, y_n), \tag{10.13}$$

$$y_{n+1} = y_n + hf(x_{n+1/2}, \hat{y}_{n+1/2}). \tag{10.14}$$

This scheme is second-order accurate with higher stability compared with previous simple schemes. One can view this scheme as a predictor-corrector method. In fact, we can use multisteps to devise higher-order

methods if the right combinations are used to eliminate the error terms order by order. The popular classical Runge-Kutta method can be written as

$$a = hf(x_n, y_n),\ b = hf(x_n + h/2, y_n + a/2),\ c = hf(x_n + h, y_n + b/2),$$

$$d = hf(x_n + h, y_n + c), \qquad y_{n+1} = y_n + \frac{a + 2(b + c) + d}{6}, \quad (10.15)$$

which is fourth-order accurate. Generally speaking, the higher-order scheme is better than the lower scheme, but not always.

## 10.2 Hyperbolic Equations

The numerical solution of partial differential equations is more complicated than that of ODEs because it involves time and space variables and the geometry of the domain of interest. Usually, boundary conditions are more complex. In addition, nonlinear problems are very common in engineering applications. We start with the simplest first-order equations and then move onto more complicated cases.

### 10.2.1 First-Order Hyperbolic Equation

For simplicity, we start with the one-dimensional scalar equation of hyperbolic type,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \quad (10.16)$$

where $c$ is a constant or the velocity of advection. By using the forward Euler scheme for time and centred-spaced scheme, we have

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c[\frac{u_{j+1}^n - u_{j-1}^n}{2h}] = 0, \quad (10.17)$$

where $t = n\Delta t, n = 0, 1, 2, ...,\ x = x_0 + jh, j = 0, 1, 2, ...,$ and $h = \Delta x$. In order to see how this method behaves numerically, we use the von Neumann stability analysis by assuming that the solutions are the spatial eigenmodes or Fourier modes. Let look at it as an example.

---

**Example 10.2:** Assuming the independent solutions or eigenmodes or Fourier modes in spatial coordinate $x$ in the form of $u_j^n = \xi^n e^{ikhj}$, and substituting into Eq.(10.17), we have

$$\frac{\xi^n(\xi - 1)e^{ikhj}}{\Delta t} + c\frac{\xi^n[e^{ikh} - e^{-ikh}]e^{ikhj}}{2h} = 0.$$

Dividing both sides by $\xi^n e^{ikhj}$ and using $\sin x = (e^{ix} - e^{-ix})/2$, we have

$$\xi = 1 - i\frac{c\Delta t}{h}\sin(kh). \tag{10.18}$$

The stability criteria $|\xi| \leq 1$ require

$$|\xi| = \sqrt{1 + (\frac{c\Delta t}{h})^2 \sin^2 kh} \leq 1,$$

or

$$(\frac{c\Delta t}{h})^2 \sin^2 kh \leq 0.$$

However, this inequality is impossible to satisfy and this scheme is thus unconditionally unstable.

---

To avoid the difficulty of instability, we can use other schemes such as the upwind scheme and Lax scheme. For the upwind scheme, the equation becomes

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c[\frac{u_j^n - u_{j-1}^n}{h}] = 0, \tag{10.19}$$

whose stability condition is

$$|\xi| = |1 - \frac{c\Delta t}{h}[1 - \cos(kh) + i\sin(kh)]| \leq 1, \tag{10.20}$$

which is equivalent to

$$0 < \frac{c\Delta t}{h} \leq 1. \tag{10.21}$$

This the well-known Courant-Friedrichs-Lewy stability condition, often referred to as the Courant stability condition. Thus, the upwind scheme is conditionally stable.

### 10.2.2 Second-Order Wave Equation

Higher-order equations such as second-order wave equation can be written as a system of hyperbolic equations and then be solved using numerical integration. They can also be solved by direct discretisation using finite difference scheme. The wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \tag{10.22}$$

consists of second derivatives. If we approximate the first derivatives at each time step $n$ using

$$u_i' = \frac{u_{i+1}^n - u_i^n}{\Delta x}, \qquad u_{i-1}' = \frac{u_i^n - u_{i-1}^n}{\Delta x}, \tag{10.23}$$

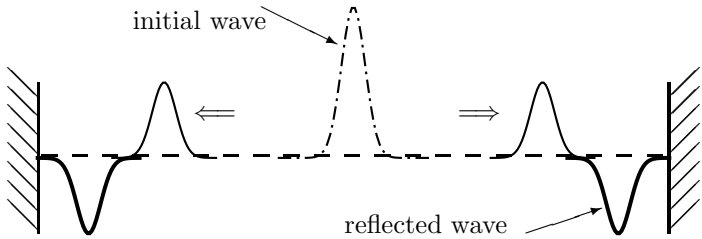



Figure 10.2: Travelling wave solution of $u_{tt} - c^2 u_{xx} = 0$.

then we can use the following approximation for the second derivative

$$u_i'' = \frac{u_i' - u_{i-1}'}{\Delta x} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2}. \tag{10.24}$$

This is in fact a central difference scheme of second-order accuracy. If we use the similar scheme for time-stepping, then we get a central difference scheme in both time and space.

Thus, the numerical scheme for this equation becomes

$$\frac{u_i^{n+1} - 2u_i^n + u_i^{n-1}}{(\Delta t)^2} = c^2 \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2}. \tag{10.25}$$

This is a two-level scheme with a second-order accuracy. The idea of solving this difference equation is to express (or to solve) $u_i^{n+1}$ at time step $t = n + 1$ in terms of the known values or data $u_i^n$ and $u_i^{n-1}$ at two previous time steps $t = n$ and $t = n - 1$.

Solving the wave equation (10.22) with the initial condition

$$u(x, 0) = qe^{-[\frac{20}{L}(x - \frac{L}{2}]^2}, \tag{10.26}$$

and wave reflection boundary conditions at both ends $u(0, t) = u(L, t) = 0$, we have the solution shown in Figure 10.2. We can see that the initial profile is split into two travelling waves: one travels to the left and one to the right.

## 10.3 Parabolic Equation

For the parabolic equation such as the diffusion or heat conduction equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}(D\frac{\partial u}{\partial x}), \tag{10.27}$$

a simple Euler method for the time derivative and centred second-order approximations for space derivatives lead to

$$u_j^{n+1} = u_j^n + \frac{D\Delta t}{h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \tag{10.28}$$

Let us look at its stability.

---

**Example 10.3:** From the application of von Neumann stability analysis by substituting $u_j^n = \xi^n e^{ikhj}$ into this equation, we have

$$\xi^{n+1} e^{ikhj} = \xi^n e^{ikhj} + \frac{D\Delta t \xi^n}{h^2}[e^{ikh(j+1)} - 2e^{ikhj} + e^{ikh(j-1)}].$$

Dividing both sides by $\xi^n e^{ikhj}$, we have

$$\xi = 1 + \frac{D\Delta t}{h^2}[e^{ikh} - 2 + e^{-ikh}].$$

Using $\cos x = (e^{ix} + e^{-ix})/2$ and $\sin^2(x/2) = (1 - \cos x)/2$, we have

$$\xi = 1 - \frac{4D\Delta t}{h^2}\sin^2(\frac{kh}{2}). \tag{10.29}$$

The stability requirement $\xi \leq 1$ leads to the constraint on the timestep,

$$\Delta t \leq \frac{h^2}{2D}. \tag{10.30}$$

This scheme is conditionally stable.

---

For simplicity, we consider a 1-D heat conduction equation $u_t = \kappa u_{xx}$ with an initial condition

$$u(x, 0) = q[H(x - 5L/8) - H(x - 3L/8)]$$

where $H(x)$ is a Heaviside function: $H(x) = 1$, if $(x \geq 0)$, and $H(x) = 0$ if $(x < 0)$.

The evolution of the temperature profile is shown in Figure 10.3 where the initial profile is plotted as a dashed curve. We can see that the profile is gradually smoothed out as time increases and this is the typical behaviour of the diffusive system. The time-stepping scheme we used limits the step size of time as larger time steps will make the scheme unstable. There are many ways to improve this, and one of most widely used schemes is the implicit scheme.

To avoid the difficulty caused by very small timesteps, we now use an implicit scheme for time derivative differencing, and thus we have

$$u_j^{n+1} - u_j^n = \frac{D\Delta t}{h^2}(u_{j+1}^{n+1} + 2u_j^{n+1} + u_{j-1}^{n+1}). \tag{10.31}$$

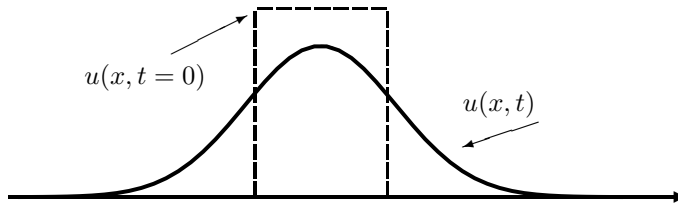



Figure 10.3: The 1-D time-dependent diffusion equation: $u_t - \kappa u_{xx} = 0$.

Applying the stability analysis, we have

$$\xi = \frac{1}{1 + \frac{4D\Delta t}{h^2} \sin^2 \frac{kh}{2}}, \tag{10.32}$$

whose norm is always less than unity ($|\xi| \leq 1$). This means the implicit scheme is unconditionally stable for any size of time steps. That is why implicit methods are more desirable in simulations. However, there is one disadvantage of this method, which requires more programming skills because the inverse of a large matrix is usually needed in implicit schemes.

## 10.4 Elliptical Equation

In the parabolic equation, if the time derivative is zero or $u$ does not change with time $u_t = 0$, then we reach a steady-state problem that is governed by the elliptic equation. For the steady state heat conduction problem, we generally have the Poisson problem,

$$\nabla \cdot [\kappa(u, x, y, t) \nabla u] = f, \tag{10.33}$$

If $\kappa$ is a constant, this becomes

$$\nabla^2 u = q, \qquad q = \frac{f}{\kappa}. \tag{10.34}$$

There are many methods available to solve this problem, such as the boundary integral method, the relaxation method, and the multigrid method. Two major ones are the long-time approximation of the transient parabolic diffusion equations, the other includes the iteration method. Long time approximation method is essentially based on the fact that the parabolic equation

$$\frac{\partial u}{\partial t} + \kappa \nabla^2 u = f, \tag{10.35}$$

evolves with a typical scale of $\sqrt{\kappa t}$. If $\sqrt{\kappa t} \gg 1$, the system is approaching its steady state. Assuming $t \to \infty$ and $\kappa \gg 1$, we then have

$$\nabla^2 u = \frac{f}{\kappa} - \frac{1}{\kappa} u_t \to 0. \tag{10.36}$$

Thus, the usual numerical methods for solving parabolic equations are valid. However, other methods may obtain the results more quickly.

The iteration method uses the second-order scheme for space derivatives, and equation (10.34) in the 2-D case becomes

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta y)^2} = q. \tag{10.37}$$

If we use $\Delta x = \Delta y = h$, then the above equation simply becomes

$$(u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j}) - 4u_{i,j} = h^2 q, \tag{10.38}$$

which can be written as

$$\mathbf{A}\mathbf{u} = \mathbf{b}. \tag{10.39}$$

In principle, one can solve this equation using matrix inverse techniques such as the Gauss-Seidel iteration.

# References

Langtangen H. P., *Computational Partial Differential Equations: Numerical Methods and Diffpack Programming*, Springer, (1999).

Moler C. B., *Numerical Computing with MATLAB*, SIAM, (2004).

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

Smith G. D., *Numerical Solutions of Partial Differential Equations: Finite Differerence Methods*, 3rd ed., Clarendon Press, Oxford, (1985).

# Chapter 11

# Finite Volume Method

## 11.1 Introduction

The finite difference method discussed in the previous chapter approximates the ordinary differential equations and partial differential equations using Taylor series expansions, resulting in a system of algebraic equations. The finite volume method resembles the finite difference method in certain ways but the starting point is the integral formulation of the problem. It uses the integral form of the partial differential equations in terms of conservation laws, then approximates the surface and boundary integrals in the control volumes. This becomes convenient for problems involving flow or flux boundaries.

For a hyperbolic equation that is valid in the domain $\Omega$ with boundary $\partial\Omega$,

$$\frac{\partial u}{\partial t} - \nabla \cdot (\kappa \nabla u) = q, \tag{11.1}$$

or written in terms of flux function $\mathbf{F} = \mathbf{F}(u) = -\kappa \nabla u$, we have

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{F} = q. \tag{11.2}$$

The integral form of this equation becomes

$$\int_\Omega \frac{\partial u}{\partial t} d\Omega + \int_\Omega \nabla \cdot \mathbf{F} = \int_\Omega q d\Omega. \tag{11.3}$$

If the integral form is decomposed into many small control volumes, or finite volumes, $\Omega = \bigcup_{i=1}^{N} \Omega_i$ and $\Omega_i \bigcap \Omega_j = \emptyset$. By defining the control volume cell average or mean value

$$u_i = \frac{1}{V_i} \int_{\Omega_i} u d\Omega_i, \qquad q_i = \frac{1}{V_i} \int_{\Omega_i} q d\Omega_i, \tag{11.4}$$

where $V_i = |\Omega_i|$ is the volume of the small control volume $\Omega_i$, the above equation can be written as

$$\frac{\partial u_i}{\partial t} + \sum_{i=1}^{N} \frac{1}{V_i} \int_{\Omega_i} \nabla \cdot \mathbf{F}(u_i) d\Omega_i = q_i, \tag{11.5}$$

By using the divergence theorem

$$\int_V \nabla \cdot \mathbf{F} = \int_\Gamma \mathbf{F} \cdot \mathbf{n} dA, \tag{11.6}$$

we have

$$\frac{\partial u_i}{\partial t} + \sum_{i=1}^{N} \frac{1}{V_i} \int_{\Gamma_i} \mathbf{F} \cdot \mathbf{dS} = q_i, \tag{11.7}$$

where $\mathbf{dS} = \mathbf{n} dA$ is the surface element and $\mathbf{n}$ is the outward pointing unit vector on the surface $\Gamma_i$ enclosing the finite volume $\Omega_i$. The integration can be approximated using various numerical integration schemes. In the simplest 1-D case with $h = \Delta x$, the integration

$$u_i = \frac{1}{h} \int_{(i-1/2)h}^{(i+1/2)h} u dx, \tag{11.8}$$

is a vertex-centred finite volume scheme. In the following sections, we will discuss the three major types of partial differential equations (elliptic, parabolic and hyperbolic) and their finite volume discretisations.

## 11.2 Elliptic Equations

Laplace's equation is one of the most studied elliptic equations

$$\nabla^2 u(x, y) = 0, \qquad (x, y) \in \Omega, \tag{11.9}$$

its integral form becomes

$$\int_\Omega \nabla^2 u d\Omega = \int_\Gamma \frac{\partial u}{\partial \mathbf{n}} \cdot \mathbf{dS} = 0. \tag{11.10}$$

The end equations after discretisation will depend on the details of the problem such as boundary conditions and also the geometry of the domain. Let us look at an example.

---

**Example 11.1:** For the simple regular 2-D grid points $(i\Delta x, j\Delta y)$, the control volume in this case is a cell centred at $(i\Delta x, j\Delta y)$ with a size of $\Delta x$ (along $x-$axis) and $\Delta y$ (along $y-$axis), and the boundary integral on

any cell consists of four parts integrated on each of the four sides. By using the simple approximation $\frac{\partial u}{\partial n}$ with $\frac{\partial u}{\partial x} = (u_{i+1,j} - u_{i,j})/\Delta x$ and $\frac{\partial u}{\partial y} = (u_{i,j+1} - u_{i,j})/\Delta y$, we have

$$\int_{\Omega_{i,j}} \frac{\partial u}{\partial n} d\Omega = \frac{\Delta y}{\Delta x}(u_{i+1,j} + u_{i-1,j} - 2u_{i,j}) + \frac{\Delta x}{\Delta y}(u_{i,j+1} + u_{i,j-1} - 2u_{i,j}) = 0.$$

Dividing both sides with $\Delta x \Delta y$, and letting $\Delta x = \Delta y = h$, we obtain

$$(u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1}) - 4u_{i,j} = 0, \tag{11.11}$$

which resembles finite difference methods in many ways. In fact, this is exactly the Laplace operator for a 5-point differencing scheme.

## 11.3 Hyperbolic Equations

For the hyperbolic equation of the conservation law in the 1-D case

$$\frac{\partial u}{\partial t} + \frac{\partial \Psi(u)}{\partial x} = 0, \tag{11.12}$$

we have its integral form in the fixed domain

$$\int_{x_a}^{x_b} \frac{\partial u}{\partial t} dx = \frac{\partial}{\partial t} \int_{x_a}^{x_b} u dx = -\{\Psi[u(x_b)] - \Psi[u(x_a)]\} = 0. \tag{11.13}$$

If we use the mid-point $u^*$ to approximate the integral, we have

$$(x_b - x_a)\frac{\partial u^*}{\partial t} = -\{\Psi[u(x_b)] - \Psi[u(x_a)]\}. \tag{11.14}$$

If we choose the control volume $[(i - 1/2)\Delta x, (i + 1/2)\Delta x]$ centred at the mesh point $x_i = i\Delta x = ih$ with the approximation $u_i \approx u_i^*$ in each interval, and using the forward differencing scheme for the time derivative, we have

$$u_i^{n+1} - u_i^n = -\frac{\Delta t}{h}[\Psi(x_{i+1/2}) - \Psi(x_{i-1/2})]. \tag{11.15}$$

By further approximation of the flux $\Psi(x_{i+1/2}) \approx \Psi(x_i)$, we have the upward scheme

$$u_i^{n+1} - u_i^n = -\frac{\Delta t}{h}[\Psi(u_i) - \Psi(u_{i-1})], \tag{11.16}$$

which is conditionally stable as we know this from the finite difference method. For the simplest flux $\Psi(u) = cu$, we have

$$u^{n+1} = u_i^n - \frac{c\Delta t}{h}(u_i^n - u_{i-1}^n), \tag{11.17}$$

and its stability requires that

$$0 < \frac{c\Delta t}{h} \leq 1. \tag{11.18}$$

## 11.4 Parabolic Equations

For the case of heat conduction

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} + q(u, x, t), \tag{11.19}$$

we have its integral form

$$\int_t \int_\Omega (\frac{\partial u}{\partial t} - k\frac{\partial^2 u}{\partial x^2} - q)dxdt = 0. \tag{11.20}$$

If we use the control volume from $(i-1/2)h$ to $(i+1/2)h$ where $h = \Delta x$, and with time from step $n$ to $n+1$, we have

$$\int_{n\Delta t}^{(n+1)\Delta t} \int_{(i-1/2)h}^{(i+1/2)h} (\frac{\partial u}{\partial t} - k\frac{\partial^2 u}{\partial x^2} - q)dxdt = 0. \tag{11.21}$$

By using the mid-point approximation

$$\int_a^b \psi(x)dx = \psi[\frac{(a+b)}{2}](b-a), \tag{11.22}$$

and the DuFort-Frankel scheme where we first approximate the gradient

$$\frac{\partial^2 u}{\partial x^2} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}, \tag{11.23}$$

then replace $-2u_i^n$ with $-(u_j^{n+1} + u_j^{n-1})$, we have

$$\frac{u_i^{n+1} - u_i^{n-1}}{2\Delta t} = \frac{[(u_{i+1}^n - (u_i^{n+1} + u_i^{n-1}) + u_{i-1}^n)]}{h^2} + q_i^n, \tag{11.24}$$

where we have used the central scheme for time as well. This is exactly the DuFort-Frankel explicit scheme in the finite difference method; however, the starting point is different. In addition, the finite volume scheme is more versatile in dealing with irregular geometry and more

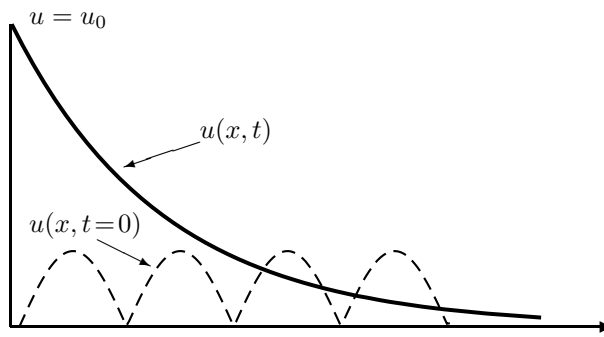



Figure 11.1: Heat conduction simulations via finite volume method.

natural in applying boundary conditions. Following the stability analysis, we get $|\xi| < 1$ is always true and thus the Dufort-Frankel scheme is unconditionally stable for all $\Delta t$ and $\Delta x$.

---

**Example 11.2:** In order to demonstrate how the finite volume method works and also show the similarity and difference between the finite volume method and the finite difference method, we start with a simple heat conduction problem

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

with boundary and initial conditions

$$u(x,0) = 0.1|\sin(4\pi x/L)|, \qquad u(0,t) = 1, \;\; u(0,t) = 0.$$

Using the finite volume method (11.24) and solving for $u_i^{n+1}$, we have

$$u_i^{n+1} = \frac{1-\beta}{1+\beta}u_i^{n-1} + \frac{\beta}{1+\beta}[u_{i+1}^n + u_{i-1}^n],$$

where $\beta \equiv 2k\Delta t/h^2$. The numerical solution is shown in Fig. 11.1 where the dashed curve corresponds to the initial profile while the solid curve corresponds to the profile at $t = 0.03$ or after 150 time steps.

---

# References

Chung T. J., *Computational Fluid Dynamics*, Cambridge University Press, (2002).

Fletcher C. A. J. and Fletcher C. A., *Computational Techniques for Fluid Dynamics*, Vol. I, Springer-Verlag, GmbH, (1997).

LeVeque R. J., *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, (2002).

Langtangen H. P., *Computational Partial Differential Equations: Numerical Methods and Diffpack Programming*, Springer, (1999).

Kant T., *Finite Elements in Computational Mechanics*, Vols. I/II, Pergamon Press, Oxford, (1985).

Mitchell A. R. and Griffiths D. F., *Finite Difference Method in Partial Differential Equations*, Wiley & Sons, New York, (1980).

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

Puckett E. G. and Colella, P., *Finite Difference Methods for Computational Fluid Dynamics*, Cambridge University Press, (2005).

Versteeg H. K, Malalasekra W. and Malalasekra W., *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*, Prentice Hall, (1995).

# Chapter 12

# Finite Element Method

In the finite difference methods, we approximate the equations at a finite number of discrete points, and there are many limitations in finite difference methods. One such disadvantage is that it is not straightforward to deal with irregular geometry. More versatile and efficient methods are greatly needed. In fact, the finite element method is one of the most successful methods in scientific computing and has a wide range of applications.

The basic idea of finite element analysis is to divide the domain into many small blocks or elements as shown in Fig. 12.1. This is equivalent to imaginarily cutting a solid structure such as a building or geological region into many pieces or elements. These small blocks are characterised by nodes, edges and surfaces, and the whole domain can be considered as if these blocks or elements are glued together at these nodes and along the element boundaries. In this way, we essentially transform a continuum system with infinite degrees of freedom into a discrete finite system with finite degrees of freedom. In fact, this is the origin of the name 'finite elements'. Since most continuum systems are governed by differential equations, the major advantage of this transformation is that the differential equation for a continuum system is transformed into a set of simultaneous algebraic equations for the discrete system with a finite number of elements. The approximations to any field quantities such as displacements and stresses over these finite elements use the piecewise polynomial interpolation techniques.

The fundamental aim of finite element analysis is to formulate the numerical method in such a way that the partial differential equation, in combination with the appropriate boundary conditions and loads, will be transformed into algebraic equations in terms of matrices. For time-dependent problems involving partial differential equations, the equations will be transformed into an ordinary differential equation in

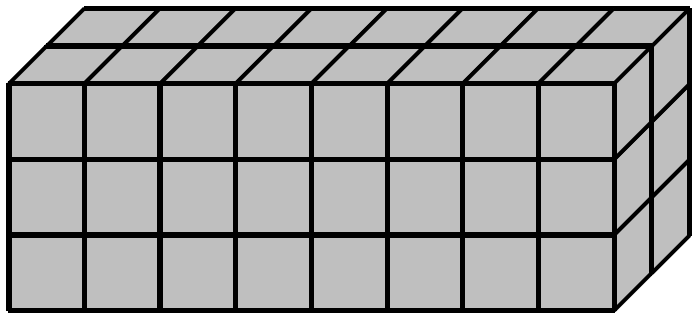


Figure 12.1: Concept of finite elements.

terms of matrices, which will in turn be discretised and converted into algebraic equations by time-stepping or some iteration techniques. For example, a linear elastic problem can be formulated in such a way that it is equivalent to an equation of the following type

$$\mathbf{Ku} = \mathbf{f}, \tag{12.1}$$

where $\mathbf{K}$ is the stiffness matrix, and $\mathbf{f}$ is a vector corresponding to nodal forces and some contribution from boundary conditions. $\mathbf{u}$ is the unknown vector to be solved and it corresponds to a nodal degree of freedom such as the displacement.

## 12.1 Concept of Elements

### 12.1.1 Simple Spring Systems

The basic idea of finite element analysis is to divide a model into many pieces or elements with discrete nodes. These elements form an approximate system to the whole structure in the domain of interest, so that the physical quantities such as displacements can be evaluated at these discrete nodes. Other quantities such as stresses and strains can then be evaluated at certain points (usually Gaussian integration points) inside elements. The simplest elements are those with two nodes in 1-D, the triangular element with three nodes in 2-D, and tetrahedral elements with four nodes in 3-D.

In order to show the basic concept, we now focus on the simplest 1-D spring element with two nodes (see Figure 12.2). The spring has a stiffness constant $k$ (N/m) with two nodes $i$ and $j$. At nodes $i$ and

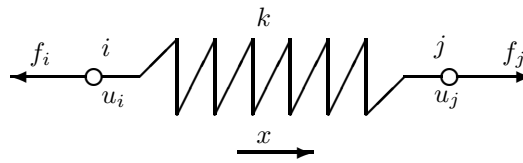



Figure 12.2: A spring element.

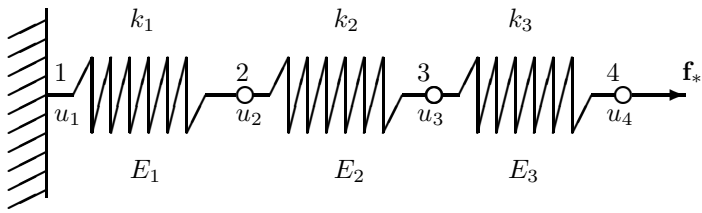



Figure 12.3: A simple spring system.

$j$, the displacements (in metres) are $u_i$ and $u_j$, respectively. $f_i$ and $f_j$ are nodal forces.

From Hooke's law, we know the displacement $\Delta u = u_j - u_i$ is related to $f$, or

$$f = k(\Delta u). \tag{12.2}$$

At node $i$, we have

$$f_i = -f = -k(u_j - u_i) = ku_i - ku_j, \tag{12.3}$$

and at node $j$, we get

$$f_j = f = k(u_j - u_i) = -ku_i + ku_j. \tag{12.4}$$

These two equations can be combined into a matrix equation

$$\begin{pmatrix} k & -k \\ -k & k \end{pmatrix} \begin{pmatrix} u_i \\ u_j \end{pmatrix} = \begin{pmatrix} f_i \\ f_j \end{pmatrix}, \qquad \text{or} \qquad \mathbf{Ku} = \mathbf{f}. \tag{12.5}$$

Here $\mathbf{K}$ is the stiffness matrix, $\mathbf{u}$ and $\mathbf{f}$ are the displacement vector and force vector, respectively. This is the basic spring element, and let us see how it works in a spring system such as shown in Figure 12.3 where three different springs are connected in series.

For a simple spring system shown in Figure 12.3, we now try to determine the displacements of $u_i (i = 1, 2, 3, 4)$. In order to do so, we

have to assemble the whole system into a single equation in terms of global stiffness matrix $\mathbf{K}$ and forces $\boldsymbol{f}$. As these three elements are connected in series, the system can be assembled element by element. For element $E_1$, its contribution to the global matrix is

$$\begin{pmatrix} k_1 & -k_1 \\ -k_1 & k_1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \tag{12.6}$$

which is equivalent to

$$\mathbf{K_1 u} = \mathbf{f}_{E_1}, \tag{12.7}$$

where

$$\begin{pmatrix} k_1 & -k_1 & 0 & 0 \\ -k_1 & k_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ . \\ . \end{pmatrix}, \tag{12.8}$$

and $\mathbf{f}_{E_1}^T = (f_1, f_2, 0, 0)$. Similarly, for element $E_2$, we have

$$\begin{pmatrix} k_2 & -k_2 \\ -k_2 & k_2 \end{pmatrix} \begin{pmatrix} u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} -f_2 \\ f_3 \end{pmatrix}, \tag{12.9}$$

or

$$\mathbf{K_2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & k_2 & -k_2 & 0 \\ 0 & -k_2 & k_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{12.10}$$

where we have used the balance at node 2. For element $E_3$, we have

$$\begin{pmatrix} k_3 & -k_3 \\ -k_3 & k_3 \end{pmatrix} \begin{pmatrix} u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} -f_3 \\ f_* \end{pmatrix}, \tag{12.11}$$

or

$$\mathbf{K}_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & k_3 & -k_3 \\ 0 & 0 & -k_3 & k_3 \end{pmatrix}, \tag{12.12}$$

where $f_4 = f_*$ has been used. We can now add the three sets of equations together to obtain a single equation

$$\begin{pmatrix} k_1 & -k_2 & 0 & 0 \\ -k_1 & k_1 + k_2 & -k_2 & 0 \\ 0 & -k_2 & k_2 + k_3 & -k_3 \\ 0 & 0 & -k_3 & k_3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ -f_2 + f_2 \\ -f_3 + f_3 \\ f_* \end{pmatrix},$$

or

$$\mathbf{K u} = \mathbf{f}, \tag{12.13}$$

where

$$\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_3 = \begin{pmatrix} k_1 & -k_1 & 0 & 0 \\ -k_1 & k_1 + k_2 & -k_2 & 0 \\ 0 & -k_2 & k_2 + k_3 & -k_3 \\ 0 & 0 & -k_3 & k_3 \end{pmatrix}, \qquad (12.14)$$

and

$$\mathbf{u}^T = (u_1, u_2, u_3, u_4), \qquad \mathbf{f} = \mathbf{f}_{E_1} + \mathbf{f}_{E_2} + \mathbf{f}_{E_3}. \qquad (12.15)$$

In general, the matrix $\mathbf{K}$ is singular or its rank is less than the total number of degrees of freedom, which is four in this case. This means that the equation has no unique solution. Thus, we need the boundary conditions to ensure a unique solution. In this spring system, if no boundary condition is applied at any nodes, then the applied force at the node 4 will make the spring system fly to the right. If we add a constraint by fixing the left node 1, then the system can stretch, and a unique configuration is formed. In our case where there are no applied forces at nodes 2 and 3, we have

$$\mathbf{f}^T = (0, 0, 0, f_*). \qquad (12.16)$$

---

**Example 12.1:** For $k_1 = 100$ N/m, $k_2 = 200$ N/m, and $k_3 = 50$N/m, and $f_* = 20$ N, the boundary at node 1 is fixed ($u_1 = 0$). Then, the stiffness matrix is

$$\mathbf{K} = \begin{pmatrix} 100 & -100 & 0 & 0 \\ -100 & 300 & -200 & 0 \\ 0 & -200 & 250 & -50 \\ 0 & 0 & -50 & 50 \end{pmatrix},$$

and the force column vector

$$\mathbf{f}^T = (0, 0, 0, 20).$$

The rank of $\mathbf{K}$ is 3, therefore, we need at least one boundary condition. By applying $u_1 = 0$, we now have only three unknown displacements $u_2, u_3, u_4$. Since $u_1 = 0$ is already known, the first equation for $u_1$ becomes redundant and we can now delete it so that the reduced stiffness matrix $\mathbf{A}$ is a $3 \times 3$ matrix. Therefore, we have

$$\mathbf{A} = \begin{pmatrix} 300 & -200 & 0 \\ -200 & 250 & 0 \\ 0 & -50 & 50 \end{pmatrix},$$

and the reduced forcing vector is

$$\mathbf{g}^T = (0, 0, 20).$$

The solution is

$$\mathbf{u} = \mathbf{A^{-1}g} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.7 \end{pmatrix}.$$

Therefore, the displacements are $u_2 = 0.2$m, $u_3 = 0.3$m, and $u_4 = 0.7$m.

Theoretically speaking, the force should be 20N everywhere in the spring systems since the mass of the springs is negligible. Let us calculate the force at nodes 2 and 3 to see if this is the case. At the node 2, the extension in element $E_1$ is $\Delta u = u_2 - u_1 = 0.2$ m, thus the force at node 2 is

$$f_2 = k_1 \Delta u = 100 \times 0.2 = 20\text{N}.$$

Similarly, at node 3 of element $E_2$, we have

$$f_3 = k_2(u_3 - u_2) = 200 \times 0.1 = 20\text{N},$$

which is the same at node 3 of element $E_3$

$$f_3 = k_3 \times (-\Delta u) = k_3(u_4 - u_3) = 50 \times 0.4 = 20\text{N}.$$

So the force is 20 N everywhere.

---

### 12.1.2 Bar Elements

The spring system we discussed earlier is limited in many ways, as a spring does not have any mass and its cross section is not explicitly included. A more complicated but realistic element is the bar element (also called truss element) as shown in Figure 12.4, which is a uniform rod with a cross section area $A$, Young's elastic modulus $E$, and a length $L$. A bar element can only support tension and compression, it cannot support bending. For this reason, it is also called a truss element.

The displacements at nodes $i$ and $j$ are $u_i$ and $u_j$, respectively. The forces at the corresponding nodes are $f_i$ and $f_j$. Now we have to derive its stiffness matrix. Assuming the bar is linearly elastic, the stress $\sigma$ is thus related to strain $\epsilon$ via $\sigma = E\epsilon$. Since $\epsilon = (u_j - u_i)/L$ and $\sigma = f/A$ where $F$ is the force in the bar element, we have

$$f = \frac{EA}{L}(\Delta u) = k(\Delta u), \tag{12.17}$$

where $\Delta u = u_j - u_i$ is the extension or elongation of the bar element. Now the equivalent spring stiffness constant is

$$k = \frac{EA}{L}. \tag{12.18}$$

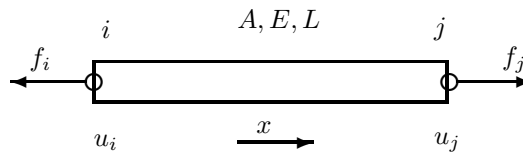



Figure 12.4: Bar element.

Therefore, the stiffness matrix $\mathbf{K}$ for this bar becomes

$$\mathbf{K} = \begin{pmatrix} k & -k \\ -k & k \end{pmatrix} = \frac{EA}{L}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{12.19}$$

We have up to now only discussed 1-D systems where all displacements $u_i$ or $u_j$ are along the bar direction, and each node has only one displacement (one degree of freedom). We now extend the study to 2-D systems. In 2-D, each node $i$ has two displacements $u_i$ (along the bar direction) and $v_i$ (perpendicular to the bar direction). Thus, each node has two degrees of freedom.

If we rotate the bar element by an angle $\theta$ as shown in Figure 12.5, we cannot use the standard addition to assemble the system. A transformation is needed from the global coordinates $(x, y)$ to the local coordinates $(x', y')$. From the geometrical consideration, the global displacements $u_i$ and $v_i$ at node $i$ are related to the local displacement $u'_i$ and (usually) $v'_i = 0$.

$$\begin{pmatrix} u'_i \\ v'_i \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} u_i \\ v_i \end{pmatrix}. \tag{12.20}$$

Using the similar transformation for $u_j$ and $v_j$, we get the transformation for the two-node bar element

$$\mathbf{u}' = \begin{pmatrix} u'_i \\ v'_i \\ u'_j \\ v'_j \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\theta & \sin\theta \\ 0 & 0 & -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} u_i \\ v_i \\ u_j \\ v_j \end{pmatrix},$$

which can be written as

$$\mathbf{u}' = \mathbf{R}\mathbf{u}, \tag{12.21}$$

where

$$\mathbf{R} = \begin{pmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\theta & \sin\theta \\ 0 & 0 & -\sin\theta & \cos\theta \end{pmatrix}. \tag{12.22}$$

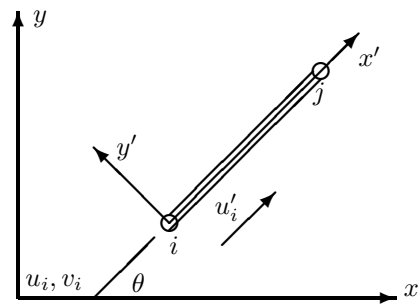



Figure 12.5: 2-D transformation of coordinates.

The same applies to transforming the force and the stiffness matrix,

$$\mathbf{f}' = \mathbf{R}\mathbf{f}, \qquad \mathbf{K}'\mathbf{u}' = \mathbf{f}'. \tag{12.23}$$

As the calculation is mainly based on the global coordinates, and the assembly should be done by transforming the local systems to the global coordinates, by combining the above two equations, we have

$$\mathbf{K}'\mathbf{R}\mathbf{u} = \mathbf{R}\mathbf{f}, \qquad \text{or} \qquad \mathbf{R}^{-1}\mathbf{K}'\mathbf{R}\mathbf{u} = \mathbf{K}\mathbf{u} = \mathbf{f}, \tag{12.24}$$

which is equivalent to a global stiffness matrix

$$\mathbf{K} = \mathbf{R}^{-1}\mathbf{K}'\mathbf{R}. \tag{12.25}$$

Since $\mathbf{R}$ is orthogonal, we have $\mathbf{R}^{-1} = \mathbf{R}^T$. From the stiffness matrix $\mathbf{K}'$ in the local coordinates

$$\mathbf{K}' = \frac{EA}{L}\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{12.26}$$

we have the stiffness matrix $\mathbf{K} = \mathbf{R}^T\mathbf{K}'\mathbf{R}$ or

$$\mathbf{K} = \frac{EA}{L}\begin{pmatrix} \cos^2\theta & \cos\theta\sin\theta & -\cos^2\theta & -\cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta & -\cos\theta\sin\theta & -\sin^2\theta \\ -\cos^2\theta & -\cos\theta\sin\theta & \cos^2\theta & \cos\theta\sin\theta \\ -\cos\theta\sin\theta & -\sin^2\theta & \cos\theta\sin\theta & \sin^2\theta \end{pmatrix}.$$

Bar elements can only elongate or shrink, they do not support bending or deflection. For more complicated elements, it is more convenient and even necessary to use a formal approach in terms of shape functions and weak formulations. Figure 12.6 shows several common elements.

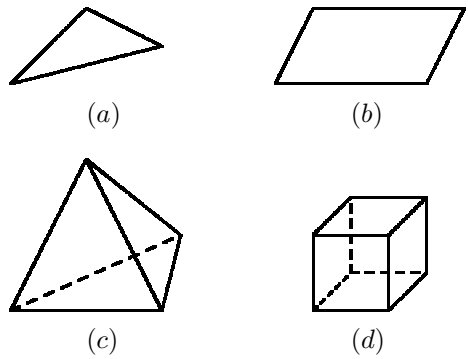



Figure 12.6: Common elements: (a) triangular; (b) quadrilateral; (c) tetrahedron; and (d) hexahedron.

## 12.2 Finite Element Formulation

### 12.2.1 Weak Formulation

Many problems are modelled in terms of partial differential equations, which can generally be written as

$$\mathcal{L}(u) = 0, \qquad \mathbf{x} \in \Omega, \tag{12.27}$$

where $\mathcal{L}$ is a differential operator, often linear. This problem is usually completed with the essential boundary condition (or prescribed values $\bar{u}$), $\mathcal{E}(u) = (u - \bar{u}) = 0$ for $\mathbf{x} \in \partial\Omega_E$, and natural boundary conditions $\mathcal{B}(u) = 0$ for $\mathbf{x} \in \partial\Omega_N$. Natural boundary conditions are usually concerned with flux or force.

Assuming that the true solution $u$ can be approximated by $u_h$ over a finite element mesh with an averaged element size or mean distance $h$ between two adjacent nodes, then we can approximate the above equation using

$$\mathcal{L}(u_h) \approx 0. \tag{12.28}$$

The ultimate goal is to construct a method of computing $u_h$ such that the error $|u_h - u|$ is minimised. Generally speaking, the residual $R(u_1, ..., u_M, \mathbf{x}) = \mathcal{L}(u_h(\mathbf{x}))$ varies with space and time. There are several methods of minimising $R$. Depending on the scheme of minimisation and the choice of shape functions, various methods can be formulated. These include the weighted residual method, the method of least squares, the Galerkin method and others.

Multiplying both sides of Eq.(12.28) by a test function or a proper weighting function $w_i$, integrating over the domain and using associated boundary conditions, we can write the general weak formulation of Zienkiewicz-type as

$$\int_{\Omega} \mathcal{L}(u_h) w_i d\Omega + \int_{\partial\Omega_N} \mathcal{B}(u_h) \bar{w}_i d\Gamma + \int_{\partial\Omega_E} \mathcal{E}(u_h) \tilde{w}_i d\Gamma_E \approx 0, \quad (12.29)$$

where $(i = 1, 2, ..., M)$, and $\bar{w}_i$ and $\tilde{w}_i$ are the values of $w_i$ on the natural and essential boundaries. If we can approximate the solution $u_h$ by the expansion in terms of shape function $N_i$

$$u_h(u,t) = \sum_{i=1}^{M} u_i(t) N_i(x) = \sum_{j=1}^{M} u_j N_j, \quad (12.30)$$

it requires that $N_i = 0$ on $\partial\Omega_E$ so that we can choose $\tilde{w}_i = 0$ on $\partial\Omega_E$. Thus, only the natural boundary conditions are included since the essential boundary conditions are automatically satisfied. In addition, there is not such limitation on the choice of $w_i$ and $\bar{w}_i$. If we choose $\bar{w}_i = -w_i$ on the natural boundary so as to simplify the formulation, we have

$$\int_{\Omega} \mathcal{L}(u_h) w_i d\Omega \approx \int_{\partial\Omega_N} \mathcal{B}(u_h) w_i d\Gamma. \quad (12.31)$$

### 12.2.2 Galerkin Method

There are many different ways to choose the test functions $w_i$ and shape functions $N_i$. One of the most popular methods is the Galerkin method where the test functions are the same as the shape functions, or $w_i = N_i$. In this special case, the formulation simply becomes

$$\int_{\Omega} \mathcal{L}(u_h) N_i d\Omega \approx \int_{\partial\Omega_N} \mathcal{B}(u_h) N_i d\Gamma. \quad (12.32)$$

The discretisation of this equation will usually lead to an algebraic matrix equation.

On the other hand, if we use the Dirac delta function as the test functions $w_i = \delta(\mathbf{x} - \mathbf{x}_i)$, the method is called the collocation method which uses the interesting properties of the Dirac function

$$\int_{\Omega} f(\mathbf{x}) \delta(\mathbf{x} - \mathbf{x}_i) d\Omega = f(\mathbf{x}_i), \quad (12.33)$$

together with $\delta(\mathbf{x} - \mathbf{x}_i) = 1$ at $\mathbf{x} = \mathbf{x}_i$ and $\delta(\mathbf{x} - \mathbf{x}_i) = 0$ at $\mathbf{x} \neq \mathbf{x}_i$.

### 12.2.3 Shape Functions

The main aim of the finite element method is to find an approximate solution $u_h(\mathbf{x}, t)$ for the exact solution $u$ at given nodal points,

$$u_h(\mathbf{x}, t) = \sum_{i=1}^{M} u_i(t) N_i(\mathbf{x}) \tag{12.34}$$

where $u_i$ are unknown coefficients or the values of $u$ at the discrete nodal point $i$. Functions $N_i$ $(i = 1, 2, ..., M)$ are linearly independent functions that vanish on the part of the essential boundary. At any node $i$, we have $N_i = 1$, and $N_i = 0$ at any other nodes, or

$$\sum_{i=1}^{M} N_i = 1, \qquad N_i(\mathbf{x}_j) = \delta_{ij}. \tag{12.35}$$

The functions $N_i(\mathbf{x})$ are referred to as basis functions, trial functions or more often shape functions in the literature of finite element methods.

#### Linear Shape Functions

For the simplest 1-D element with two nodes $i$ and $j$, the linear shape functions (shown in Figure 12.7.) can be written as

$$N_i = \frac{x_j - x}{L} = \frac{1 - \xi}{2}, \qquad N_j = \xi = \frac{x - x_i}{L} = \frac{1 + \xi}{2}, \tag{12.36}$$

where $\xi$ is the natural coordinate

$$\xi = \frac{x - x_o}{L/2}, \qquad L = |x_j - x_i|, \qquad x_o = \frac{x_i + x_j}{2}, \tag{12.37}$$

where $x_o$ is the midpoint of the element, and $\xi_i = -1$ at $x = x_i$ and $\xi_j = 1$ at $x = x_j$.

A linear shape function spans only two adjacent nodes $i$ and $j$, and it requires two coefficients in the generic form

$$N(\xi) = a + b\xi. \tag{12.38}$$

#### Quadratic Shape Functions

Suppose we want to get higher-order approximations, we can use, say, the quadratic shape functions which span three adjacent nodes $i$, $j$, and $k$. Three coefficients need to be determined

$$N(\xi) = a + b\xi + c\xi^2. \tag{12.39}$$

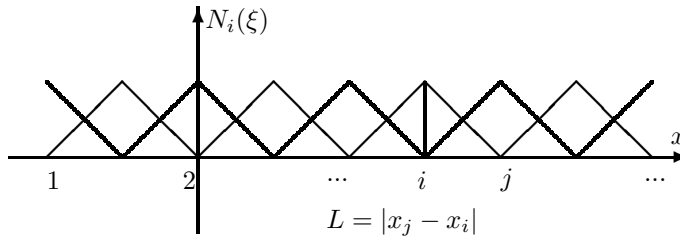



Figure 12.7: The 1-D linear shape functions.

Using the conditions $\xi_i = -1$ at $x = x_i$ and $\xi_j = 1$ at $x = x_j$, and the known displacements $u_i$, $u_j$ and $u_k$, we have

$$u_i = a + b \times (-1) + c \times (-1)^2, \tag{12.40}$$

$$u_j = a, \tag{12.41}$$

and

$$u_k = a + b \times (1) + c \times (1)^2, \tag{12.42}$$

whose solutions are

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} u_j \\ \frac{1}{2}(u_i - 2u_j + u_k) \\ \frac{1}{2}(u_k - u_i) \end{pmatrix}. \tag{12.43}$$

Substituting this into Eq.(12.39), we have

$$u = \frac{\xi(\xi - 1)}{2} u_i + (1 - \xi^2) u_j + \frac{\xi(\xi + 1)}{2} u_k, \tag{12.44}$$

which is equivalent to

$$u = N_i u_i + N_j u_j + N_k u_k, \tag{12.45}$$

where

$$\mathbf{N} = [N_i, N_j, N_k] = [\frac{\xi(\xi - 1)}{2}, (1 - \xi^2), \frac{\xi(\xi + 1)}{2}]. \tag{12.46}$$

**Lagrange Polynomials**

The essence of the shape functions is the interpolation, and the interpolation functions can be of many different types. Lagrange polynomials

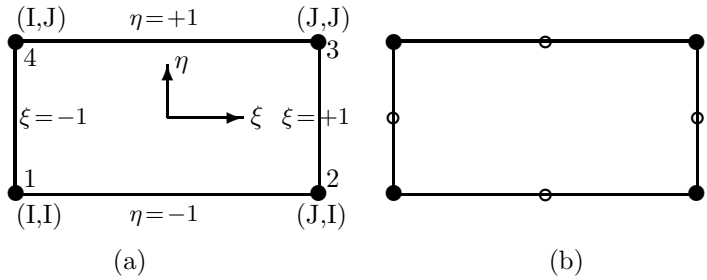



Figure 12.8: (a) A bilinear quadrilateral element; (b) A quadratic quadrilateral element.

are popularly used to construct shape functions. The $n-1$ order Lagrange polynomials require $n$ nodes, and the associated shape functions can generally be written as

$$N_i(\xi)=\prod_{j=1,j\neq i}^{n}\frac{(\xi-\xi_j)}{(\xi_i-\xi_j)}=\frac{(\xi-\xi_1)...(\xi-\xi_{i-1})(\xi-\xi_{i+1})...(\xi-\xi_n)}{(\xi_i-\xi_1)...(\xi_i-\xi_{i-1})(\xi_i-\xi_{i+1})...(\xi_i-\xi_n)},$$

where $\xi_j$ means that value of $\xi$ at node $j$. For $n=3$, it is straightforward to validate that

$$N_1(\xi)=\frac{\xi(\xi-1)}{2},\ \ N_2(\xi)=1-\xi^2,\ \ N_3(\xi)=\frac{\xi(\xi+1)}{2}. \qquad (12.47)$$

This method of formulating shape functions can be easily extended to 2D and 3D cases and for isoparametric elements. The derivative of $N_i(x)$ with respect to $\xi$ is given by

$$N_i'(\xi)=\sum_{k=1,k\neq i}^{n}\frac{1}{(\xi_i-\xi_j)}\Pi_{j=1,j\neq i}^{n}\frac{(\xi-\xi_j)}{(\xi_i-\xi_j)}. \qquad (12.48)$$

### 2D Shape Functions

The shape functions we discussed earlier are 1D shape functions (with one independent variables $x$ or $\xi$) for 1D elements. For 2D elements such as quadrilateral elements, corresponding shape functions with two independent variables: $x$ and $y$, or $\xi$ and $\eta$. Using the natural coordinates $\xi$ and $\eta$ shown in Fig. 12.8, we can construct various shape functions.

For a bilinear quadrilateral (Q4) element, we use bilinear approximations for the displacement field $u$ and $v$. If we use

$$u=\alpha_0+\alpha_1x+\alpha_2y+\alpha_3xy, \qquad (12.49)$$

$$v = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 xy, \tag{12.50}$$

and express them in terms of shape function $N_i$

$$u = \sum N_i u_i, \qquad v = \sum N_j v_j, \tag{12.51}$$

we can derive the shape functions by following the similar procedure as discussed above. We have

$$N_1 = \frac{(1-\xi)(1-\eta)}{4}, \qquad N_2 = \frac{(1+\xi)(1-\eta)}{4}, \tag{12.52}$$

$$N_3 = \frac{(1+\xi)(1+\eta)}{4}, \qquad N_4 = \frac{(1-\xi)(1+\eta)}{4}, \tag{12.53}$$

From the 1-D linear shape functions

$$N_I^{(2)}(\xi) = \frac{(1-\xi)}{2}, \qquad N_J^{(2)}(\xi) = \frac{(1+\xi)}{2}, \tag{12.54}$$

for a 2-node element (along $x$) where the superscript '(2)' means 2 nodes, we can also write another set of linear shape functions for a 2-node element in the $y$-direction. We have

$$N_I^{(2)}(\eta) = \frac{(1-\eta)}{2}, \qquad N_J^{(2)}(\eta) = \frac{(1+\eta)}{2}. \tag{12.55}$$

If we label the nodes by a pair $(I, J)$ in 2D coordinates, we have

$$N_i(\xi, \eta) = N_{IJ} = N_I^{(2)} N_J^{(2)}. \tag{12.56}$$

We can see that

$$N_1(\xi, \eta) = N_I^{(2)}(\xi) N_I^{(2)}(\eta),\ N_2(\xi, \eta) = N_J^{(2)}(\xi) N_I^{(2)}(\eta), \tag{12.57}$$

and

$$N_3(\xi, \eta) = N_J^{(2)}(\xi) N_J^{(2)}(\eta),\ N_4(\xi, \eta) = N_I^{(2)}(\xi) N_J^{(2)}(\eta). \tag{12.58}$$

In fact, higher-order shape functions for 2D and 3D elements can be systematically derived this way.

If we approximate the displacement field using higher-order approximations, then we are dealing with the quadratic quadrilateral (Q8) element because we have to use eight nodes (4 finite element nodes and 4 midpoints). In this case, the shape functions are much more complicated, for example, the shape function $N_2$ becomes

$$N_2 = \frac{(1-\xi)(1-\eta)}{4} - \frac{1}{4}[(1-\xi^2)(1-\eta) + (1+\xi)(1-\eta^2)]. \tag{12.59}$$

### 12.2.4 Estimating Derivatives and Integrals

#### Derivatives

Using the assumptions that $u_i(t)$ does not depend on space and $N_i(\mathbf{x})$ does not depend on time, the derivatives of $u$ can be approximated as

$$\frac{\partial u}{\partial \mathbf{x}} \approx \frac{\partial u_h}{\partial \mathbf{x}} = \sum_{i=1}^{M} u_i(t) N'(\mathbf{x}), \qquad \dot{u} \approx \frac{\partial u_h}{\partial t} = \sum_{i=1}^{M} \dot{u}_i N(\mathbf{x}), \tag{12.60}$$

where we have used the notations: $' = d/d\mathbf{x}$ and $\dot{} = \frac{\partial}{\partial t}$. The derivatives of the shape functions $N_i$ and higher-order derivatives are then calculated in a similar way.

#### Gauss Quadrature

In the finite element analysis, the calculation of stiffness matrices and application of boundary conditions such as in Eq.(12.32) involve the integration over elements. Such numerical integration is often carried out in terms of natural coordinates $\xi$ and $\eta$, and the Gauss integration or Gauss quadrature as discussed in Section 9.3 is usually used for evaluating integrals numerically. Gauss quadrature has relatively high accuracy. For example, the $n$-point Gauss quadrature for one-dimensional integrals

$$\mathcal{I} = \int_{-1}^{1} \psi(\xi) d\xi \approx \sum_{i=1}^{n} w_i \psi_i. \tag{12.61}$$

For the case of $n = 3$, we have

$$\int_{-1}^{1} \psi(\xi) d\xi \approx \sum_{i=1}^{3} w_i \psi_i = \frac{1}{9}[8\psi_2 + 5(\psi_1 + \psi_3)], \tag{12.62}$$

which is schematically shown in Figure 12.9 where the 2-D Gauss integration is over a quadrilateral element with point 3 at $(\xi_3, \eta_1) = (\sqrt{3/5}, -\sqrt{3/5})$ and point 9 at $(\xi_3, \eta_3)$=$(\sqrt{3/5}, \sqrt{3/5})$.

For two-dimensional integrals, we use $n^2$-point Gauss quadrature of order $n$, and we have

$$\mathcal{I} = \int_{-1}^{1}\int_{-1}^{1} \psi(\xi, \eta) d\xi d\eta = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \psi_{ij}, \tag{12.63}$$

where $\psi_{ij} = \psi(\xi_i, \eta_j)$. In the case of $n = 3$, we have 9 points (shown in Figure 12.9), and the quadrature becomes

$$\mathcal{I} = \int_{-1}^{1}\int_{-1}^{1} \psi(\xi, \eta) \approx \sum_{i=1}^{3}\sum_{j=1}^{3} w_i w_j \psi_{i+3*(j-1)}(\xi_i, \eta_j)$$

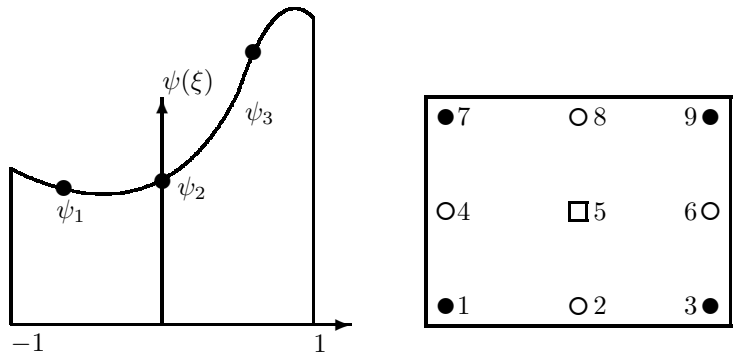



Figure 12.9: Gauss quadrature: a) 1-D integration with $|\xi_1 - \xi_2| = |\xi_2 - \xi_3| = \sqrt{3/5}$; and b) 2-D 9-point integration.

$$= \frac{1}{81}[25(\psi_1 + \psi_3 + \psi_7 + \psi_9) + 64\psi_5 + 40(\psi_2 + \psi_4 + \psi_6 + \psi_8)]. \quad (12.64)$$

## 12.3 Heat Transfer

Heat transfer problems are very common in engineering and earth sciences. The geometry in most applications is irregular. Thus, finite element methods are especially useful in this case.

### 12.3.1 Basic Formulation

The steady-state heat transfer is governed by the heat conduction equation or Poisson's equation

$$\nabla \cdot (k\nabla u) + Q = 0, \quad (12.65)$$

with the essential boundary condition

$$u = \bar{u}, \qquad \mathbf{x} \in \partial\Omega_E, \quad (12.66)$$

and the natural boundary condition

$$k\frac{\partial u}{\partial n} - q = 0, \qquad \mathbf{x} \in \partial\Omega_N. \quad (12.67)$$

Multiplying both sides of Eq.(12.65) by the shape function $N_i$ and using the formulation similar to the formulation (12.32) in terms of $u \approx u_h$,

we have

$$\int_{\Omega}[\nabla\cdot(k\nabla u)+Q]N_i d\Omega-\int_{\partial\Omega_N}[k\frac{\partial u}{\partial n}-q]N_i d\Gamma=0. \tag{12.68}$$

Integrating by parts and using Green's theorem, we have

$$-\int_{\Omega}(\nabla u_h\cdot k\cdot\nabla N_i)d\Omega+\int_{\partial\Omega}k\frac{\partial u_h}{\partial n}N_i d\Gamma$$

$$+\int_{\Omega}QN_i d\Omega-\int_{\partial\Omega_N}[k\frac{\partial u_h}{\partial n}-q]N_i d\Gamma=0. \tag{12.69}$$

Since $N_i=0$ on $\partial\Omega_E$, thus we have

$$\int_{\partial\Omega}[\,]N_i d\Gamma=\int_{\partial\Omega_N}[\,]N_i d\Gamma. \tag{12.70}$$

Therefore, the above weak formulation becomes

$$\int_{\Omega}(\nabla u_h\cdot k\cdot\nabla N_i)d\Omega-\int_{\Omega}QN_i d\Omega-\int_{\partial\Omega_N}qN_i d\Gamma=0. \tag{12.71}$$

Substituting $u_h=\sum_{j=1}^{M}u_jN_j(\mathbf{x})$ into the equation, we have

$$\sum_{j=1}^{M}[\int_{\Omega}(k\nabla N_i\cdot\nabla N_j)d\Omega]u_j-\int_{\Omega}QN_i d\Omega-\int_{\partial\Omega_N}qN_i d\Gamma=0. \tag{12.72}$$

This can be written in the compact matrix form

$$\sum_{j=1}^{M}K_{ij}U_j=f_i,\qquad \mathbf{KU}=\mathbf{f}, \tag{12.73}$$

where $\mathbf{K}=[K_{ij}],(i,j=1,2,...,M)$, $\mathbf{U}^T=(u_1,u_2,...,u_M)$, and $\mathbf{f}^T=(f_1,f_2,...,f_M)$. That is,

$$K_{ij}=\int_{\Omega}k\nabla N_i\nabla N_j d\Omega, \tag{12.74}$$

$$f_i=\int_{\Omega}QN_i d\Omega+\int_{\partial\Omega_N}qN_i d\Gamma. \tag{12.75}$$

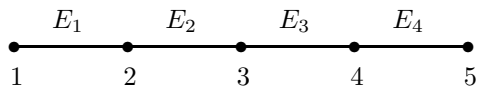



Figure 12.10: 1-D heat transfer with four elements and five nodes.

---

**Example 12.2(a):** As a simple example, we consider the 1-D steady-state heat conduction problem,

$$u''(x) + Q(x) = 0,$$

with boundary conditions

$$u(0) = \beta, \qquad u'(1) = q.$$

For a special case $Q(x) = Q =$ const, we have the analytical solution

$$u(x) = -\frac{Q}{2}x^2 + (Q+q)x + \beta. \tag{12.76}$$

Then Eq.(12.75) becomes

$$\sum_{j=1}^{M}(\int_0^1 N_i' N_j' dx)u_j = \int_0^1 QN_i dx + qN_i(1).$$

For the purpose of demonstrating the implementation procedure, let us solve this problem by dividing the interval into 4 elements and 5 nodes shown in Fig. 12.10. This will be discussed later in more detail.

---

### 12.3.2 Element-by-Element Assembly

In order to assembly the linear matrix system, we now use the popular element-by-element method. The stiffness matrix $\mathbf{K}$ in Eqs.(12.73) and (12.75) is the summation of the integral over the whole solution domain, and the domain is now divided into $m$ elements with each element on a subdomain $\Omega_e$ $(e = 1, 2, ..., m)$. Each element contributes to the whole stiffness matrix, and in fact, its contribution is a pure number. Thus, assembly of the stiffness matrix can be done in an element-by-element manner. Furthermore, $K_{i,j} \neq 0$ if and only if (or *iff*) nodes $i$ and $j$ belong to the same elements. In the 1-D case, $K_{i,j} \neq 0$ only for

$j = i-1, i, i+1$. In finite element analysis, the shape functions $N_j$ are typically localised functions, thus the matrix $\mathbf{K}$ is usually sparse in most cases. The element-by-element formulation can be written as

$$K_{i,j} = \sum_{e=1}^{m} K_{i,j}^{(e)}, \qquad K_{i,j}^{(e)} = \int_{\Omega_e} k \nabla N_i \nabla N_j d\Omega_e, \tag{12.77}$$

and

$$f_i = \sum_{e=1}^{m} f_i^{(e)}, \qquad f_i^{(e)} = \int_{\Omega_e} Q N_i d\Omega_e + \int_{\partial\Omega_{N\,e}} q N_i d\Gamma_e. \tag{12.78}$$

In addition, since the contribution of each element is a scalar or a simple number, the integration of each element can be done using the local coordinates and local node numbers or any coordinate system for the convenience of integration over an element. Then, the nonzero contribution of each element to the global system matrix $\mathbf{K}$ is simply assembled by direct addition to the corresponding global entry (of the stiffness matrix) of the corresponding nodes or related equations. In reality, this can be easily done using an index matrix to trace the element contribution to the global system matrix.

### 12.3.3 Application of Boundary Conditions

Boundary conditions can be essential, natural or mixed. The essential boundary conditions are automatically satisfied in the finite element formulation by the approximate solution. These include the displacement, rotation, and known value of the solution. Sometimes, they are also called the geometric boundary conditions. In our example, it is $u(0) = \beta$. Natural boundary conditions often involve the first derivatives such as strains, heat flux, force, and moment. Thus, they are also referred to as force boundary conditions. In our example, it is $u'(1) = q$.

The natural boundary conditions are included in the integration in the finite element equations such as (12.75). Thus no further imposition is necessary. On the other hand, although the essential boundary conditions are automatically satisfied in the finite element formulations, they still need to be implemented in the assembled finite element equations to ensure unique solutions. The simplest way is direct application. In this method, we simply use the expansion $u_h = \sum_{i=1}^{M} u_i N_i$, and apply directly the essential boundary conditions at point $i$ to replace the corresponding $i$th equation with $u_i = \bar{u}_i$ so that $i$th row of the stiffness matrix $\mathbf{K}$ in Eq.(12.73) becomes $(0, 0, ..., 1, ..., 0)$ and the corresponding $f_i = f(i) = \bar{u}_i$. All other points will be done in a similar manner.

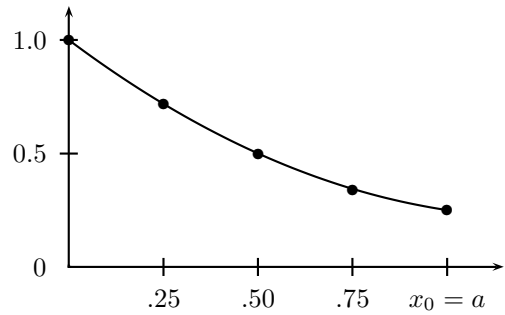



Figure 12.11: Comparison of numerical results (dots) with analytical solution (solid curve).

This method is widely used due to its simplicity and the advantage of time-stepping because it allows bigger time steps. The imposition of the essential boundary conditions can be done in many other ways including Lagrangian multiplier and penalty method.

---

**Example 12.2(b):** The assembly of the global system matrix for the example with 4 elements and five nodes (see Fig. 12.10) is shown below. For each element with $i$ and $j$ nodes, we have

$$N_i = 1 - \xi, \qquad N_j = \xi, \qquad \xi = \frac{x}{L}, \qquad L = h_e,$$

$$K_{ij}^{(e)} = [\int_0^L k N_i' N_j' dx] = \frac{k}{h_e}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad f_i^{(e)} = \frac{Q h_e}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

so that, for example in elements 1 and 2, these can extend to all nodes (with $h_i = x_{i+1} - x_i, i = 1, 2, 3, 4$),

$$K^{(1)} = \begin{pmatrix} k/h_1 & -k/h_1 & 0 & 0 & 0 \\ -k/h_1 & k/h_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad f^{(1)} = \frac{Q}{2}\begin{pmatrix} h_1 \\ h_1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$K^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & k/h_2 & -k/h_2 & 0 & 0 \\ 0 & -k/h_2 & k/h_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad f^{(2)} = \frac{Q}{2}\begin{pmatrix} 0 \\ h_2 \\ h_2 \\ 0 \\ 0 \end{pmatrix},$$

and so on. Now the global system matrix becomes

$$K = \begin{pmatrix} k/h_1 & -k/h_1 & 0 & 0 & 0 \\ -k/h_1 & \frac{k}{h_1}+\frac{k}{h_2} & -k/h_2 & 0 & 0 \\ 0 & -k/h_2 & \frac{k}{h_2}+\frac{k}{h_3} & -k/h_3 & 0 \\ 0 & 0 & -k/h_3 & \frac{k}{h_3}+\frac{k}{h_4} & -k/h_4 \\ 0 & 0 & 0 & -k/h_4 & k/h_4 \end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}, \qquad \boldsymbol{f} = \begin{pmatrix} Qh_1/2 \\ Q(h_1+h_2)/2 \\ Q(h_2+h_3)/2 \\ Q(h_3+h_4)/2 \\ Qh_4/2+q \end{pmatrix},$$

where the last row of $\boldsymbol{f}$ has already included the natural boundary condition at $u'(1) = q$.

For the simplified case of $k = 1$, $Q = -1$, $\beta = 1$, $q = -0.25$, and $h_1 = ... = h_4 = 0.25$, we have

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -4 & 8 & -4 & 0 & 0 \\ 0 & -4 & 8 & -4 & 0 \\ 0 & 0 & -4 & 8 & -4 \\ 0 & 0 & 0 & -4 & 4 \end{pmatrix}, \qquad \boldsymbol{f} = \begin{pmatrix} 1 \\ -0.25 \\ -0.25 \\ -0.25 \\ -0.375 \end{pmatrix}.$$

So the solution is

$$\mathbf{U} = \mathbf{K}^{-1}\boldsymbol{f} = \begin{pmatrix} 1.00 & 0.72 & 0.50 & 0.34 & 0.25 \end{pmatrix}^T.$$

The values of this numerical solution are plotted against the analytical solution (12.76) in Fig. 12.11. We can see that numerical solutions agree well with the analytical results.

## 12.4 Transient Problems

### 12.4.1 The Time Dimension

The problems we have discussed so far are static or time-independent because the time dimension is not involved. For time-dependent problems, the standard finite element formulation first produces an ordinary differential equation for matrices rather than algebraic matrix equations. Therefore, besides the standard finite element formulations, extra time-stepping schemes should be used in a similar manner to that in finite difference methods.

As the weak formulation uses the Green theorem that involves the spatial derivatives, the time derivatives can be considered as the source term. Thus, one simple and yet instructive way to extend the finite element formulation to include the time dimension is to replace $Q$ in Eq.(12.65) by $Q - \alpha u_t - \beta u_{tt} = Q - \alpha \dot{u} - \beta \ddot{u}$. Now we have

$$\nabla \cdot (k \nabla u) + (Q - \alpha \dot{u} - \beta \ddot{u}) = 0. \tag{12.79}$$

The boundary conditions and initial conditions are $u(\mathbf{x}, 0) = \phi(\mathbf{x})$, $u = \mathbf{u}, \mathbf{x} \in \partial\Omega_E$, and $k\frac{\partial u}{\partial n} - q = 0, \mathbf{x} \in \partial\Omega_N$. Using integration by parts and the expansion $u_h = \sum_{j=1}^{M} u_j N_j$, we have

$$\sum_{j=1}^{M} [\int_\Omega (k \nabla N_i \nabla N_j) d\Omega] + \sum_{j=1}^{M} \int_\Omega [(N_i \alpha N_j)\dot{u}_j + (N_i \beta N_j)\ddot{u}_j] d\Omega$$

$$- \int_\Omega N_i Q d\Omega - \int_{\partial\Omega_N} N_i q d\Gamma = 0, \tag{12.80}$$

which can be written in a compact form as

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}, \tag{12.81}$$

where

$$K_{ij} = \int_\Omega [(k \nabla N_i \nabla N_j)] d\Omega, \tag{12.82}$$

$$f_i = \int_\Omega N_i Q d\Omega + \int_{\partial\Omega_N} N_i q d\Gamma, \tag{12.83}$$

and

$$C_{ij} = \int_\Omega N_i \alpha N_j d\Omega, \qquad M_{ij} = \int_\Omega N_i \beta N_j d\Omega. \tag{12.84}$$

The matrices $\mathbf{K}, \mathbf{M}, \mathbf{C}$ are symmetric, that is to say, $K_{ij} = K_{ji}$, $M_{ij} = M_{ji}$, $C_{ij} = C_{ji}$ due to the interchangeability of the orders in the product of the integrand $k$, $N_i$ and $N_j$ (i.e., $\nabla N_i \cdot k \cdot \nabla N_j = k \nabla N_i \nabla N_j$, $N_i \alpha N_j = N_j \alpha N_i = \alpha N_i N_j$ etc). The matrix $\mathbf{C} = [C_{ij}]$ is the damping matrix similar to the damping coefficient of damped oscillations. $\mathbf{M} = [M_{ij}]$ is the general mass matrix due to a similar role acting as an equivalent mass in dynamics. In addition, before the boundary conditions are imposed, the matrix is usually singular, which may imply many solutions. Only after the proper boundary conditions have been enforced, the stiffness matrix will be nonsingular, thus unique solutions may be obtained. On the other hand, $\mathbf{M}$ and $\mathbf{C}$ will be always non-singular if they are not zero. For example, for the 1-D elements (with nodes $i$ and $j$),

$$K_{ij}^{(e)} = \frac{k}{h_e} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \det[K^{(e)}] = 0, \tag{12.85}$$

but

$$M_{ij}^{(e)} = \frac{\beta h_e}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}, \qquad \det[M^{(e)}] \neq 0,$$

$$C_{ij}^{(e)} = \frac{\alpha h_e}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}, \qquad \det[C^{(e)}] \neq 0, \tag{12.86}$$

Clearly, if $\mathbf{M} = 0$, it reduces to the linear heat conduction. If $\mathbf{C} = 0$, it becomes the wave equation with the source term.

### 12.4.2 Time-Stepping Schemes

From the general governing equation

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}, \tag{12.87}$$

we see that it is an ordinary differential equation in terms of time and matrices. Thus, in principle, all the time-stepping methods developed in the standard finite difference method can be used for this purpose. For a simple center difference scheme, we have

$$\dot{\mathbf{u}} = \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t}, \qquad \ddot{\mathbf{u}} = \frac{(\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1})}{(\Delta t)^2}. \tag{12.88}$$

so that Eq.(12.87) becomes

$$\mathbf{M}\frac{(\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1})}{(\Delta t)^2} + \mathbf{C}\frac{(\mathbf{u}^{n+1} - \mathbf{u}^{n-1})}{2\Delta t} + \mathbf{K}\mathbf{u}^n = \mathbf{f}. \tag{12.89}$$

Now the aim is to express $\mathbf{u}^{n+1}$ in terms of $\mathbf{u}^n$ and $\mathbf{u}^{n-1}$.

### 12.4.3 Travelling Waves

For the wave equation ($\mathbf{C} = 0$), we have

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}. \tag{12.90}$$

Using a central difference scheme for $\ddot{\mathbf{u}}$, we have

$$\mathbf{u}^{n+1} = \mathbf{M}^{-1}\mathbf{f}(\Delta t)^2 + [2\mathbf{I} - (\Delta t)^2\mathbf{M}^{-1}\mathbf{K}]\mathbf{u}^n - \mathbf{u}^{n-1}, \tag{12.91}$$

where $\mathbf{I}$ is an identity or unit matrix.

---

**Example 12.3:** For example, the 1-D wave equation

$$\frac{\partial^2 u}{\partial t^2} = c\frac{\partial^2 u}{\partial x^2}, \tag{12.92}$$

with the boundary conditions

$$u(0) = u(1) = 0, \qquad u(x,0) = e^{-(x-1/2)^2}, \tag{12.93}$$

can be written as

$$M_{ij} = \int_0^1 N_i N_j dx, \;\; K_{ij} = \int_0^1 c N_i' N_j' dx, \;\; \mathbf{f} = 0. \tag{12.94}$$

For $h = h_i = x_{i+1} - x_i =$ const, we have

$$\mathbf{K}= \begin{pmatrix} \frac{1}{h} & -\frac{1}{h} & 0 & ... & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & ... & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & ... & & \frac{2}{h} & -\frac{1}{h} \\ 0 & ... & & -\frac{1}{h} & \frac{1}{h} \end{pmatrix}, \;\; \mathbf{M}= \begin{pmatrix} h/2 & 0 & 0 & ... & 0 \\ 0 & h & 0 & ... & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & ... & & h & \\ 0 & ... & & & h/2 \end{pmatrix}.$$

For fixed boundary conditions at both ends $u(0) = u(L) = 0$, we have $K(1,1) = K(n,n) = 1$, and $K(1,2) = K(n,n-1) = 0$. Therefore, the global stiffness matrix becomes

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & ... & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & ... & & \frac{2}{h} & -\frac{1}{h} \\ 0 & ... & & 0 & 1 \end{pmatrix}.$$

Now the time march scheme is simply

$$\mathbf{u}^{n+1} = [2\mathbf{I} - (\Delta t)^2 \mathbf{M}^{-1}\mathbf{K}]\mathbf{u}^n - \mathbf{u}^{n-1}.$$

The initial profile $\mathbf{u}^0$ is derived from the $u(x,0) = \exp[-(x-1/2)^2]$. This problem can be solved using the Matlab (and Octave) program given at the end of this book (Appendix B).

# References


Bathe K. J., *Finite Element Procedures in Engineering Analysis*, Prentice Hall, (1982).

Cook R. D., *Finite Element Modelling For Stress Analysis*, Wiley & Sons, (1995).

Langtangen H. P., *Computational Partial Differential Equations: Numerical Methods and Diffpack Programming*, Springer, (1999).

Zienkiewicz O. C. and Taylor R. L., *The Finite Element Method*, vol. I/II, McGraw-Hill, 4th Edition, (1991).

# Part III

# Applications to Earth Sciences

# Chapter 13

# Reaction-Diffusion System

The partial differential equations we solved in the previous chapters using either finite difference methods or finite element methods are linear equations. The linear parabolic equation can easily be generalised to a nonlinear reaction-diffusion equation. Mathematically speaking, nonlinear equations are far more difficult to analyse, if not impossible. From the numerical point of view, some extra linearisation and approximations should be used for nonlinear terms. However, the finite difference schemes should still be useful for most nonlinear equations though they should be implemented more carefully.

The most interesting phenomenon associated with the reaction-diffusion system is probably the pattern formation due to its intrinsic instability under appropriate conditions. Beautiful patterns can automatically be formed under suitable conditions, including the patterns on animal skins (tigers, zebras), pattern of sea shells, mineral interaction in zebra stones, and the yellow and green bands of zebra grass leaves as shown in Chapter 1.

## 13.1 Mineral Reactions

Pattern formation based on the reaction-diffusion mechanism occurs commonly in geophysical and geological processes. On the one hand, patterns such as sand dunes, vegetation in grassland and plankton distribution are observed on a large scale. On the other hand, mineral banding such as gneiss can be seen almost everywhere. The layering and banding in rocks is formed by the segregation into separate bands or foliations during metamorphism and diagenetic processes due to dis-

solution and precipitation and the reaction-advection-diffusion transport mechanism, and granular minerals such as quartz and feldspar alternate with bands in which either platy mica or elongate amphibole minerals dominate.

An important class of mineral reactions is probably the diagenetic reaction in which the water-rich minerals such as smectite is transformed into more stable minerals such as illite. Such reactions are mainly carried out via the dissolution-precipitation processes. The details of the reaction will depend on many factors including temperature, pressure, the concentration of the minerals, and other minerals presented. For simplicity, we here provide a simplified schematic reaction model

$$\text{mineral A} \xrightarrow{r_1} \text{mineral C} + n[H_2O], \tag{13.1}$$

$$\text{mineral C} \xrightarrow{r_2} \text{mineral B}, \tag{13.2}$$

where we are only concerned with two minerals A and B, and the mineral C is the intermediate phase (usually aqueous), and $n$ is the number of water molecules produced in the reaction. $r_1$ and $r_2$ are the reaction rates of the two-step reactions, respectively.

[A] and [B] are the concentrations of A and B, respectively. Let $\boldsymbol{v}$ be the velocity of the flow. The mass conservation of both A and B leads to

$$\frac{\partial [A]}{\partial t} + \nabla \cdot ([A]\boldsymbol{v}) = D_A \nabla^2 [A] - r_1, \tag{13.3}$$

and

$$\frac{\partial [B]}{\partial t} + \nabla \cdot ([B]\boldsymbol{v}) = D_B \nabla^2 [B] + r_2, \tag{13.4}$$

where $D_A$ and $D_B$ are diffusion coefficients of minerals A and B, respectively. It is worth pointing out that these two equations have been written in terms of simplified solute transport. In reality, these reactions occur in porous media under the influence of temperature and pressure. The porosity of the porous media should be included (see discussion in Chapter 15).

The rates $r_1$ and $r_2$ may take complicated forms, depending on the detailed mineral chemistry. For example, we may have

$$r_1 = k_1[A](1 - e^{\Delta G_1/RT})^{m_1}, \qquad r_2 = k_2(e^{\Delta G_2/RT} - 1)^{m_2}, \tag{13.5}$$

where $k_1$ and $k_2$ are constants, and $\Delta G_i$ is the change of Gibbs free energy. $T$ is the temperature and $R$ is the universal gas constant. The exponent $m_1, m_2 = 1 \sim 2$ is also a known constant. For reactions to happen, it is necessary that $\Delta G_i < 0$. If the solution is very dilute, it usually follows that $\Delta G_i = RT(\ln c/c_0)$ where $c = [A]$ or $[B]$ and

$c_0$ is the concentration at equilibrium. For given $m_1 = 2, m_2 = 1$ and $c_0 = 1$, we have

$$r_1 = k_1[A](1 - \alpha[A])^2, \qquad r_2 = k_2(1 - [B]), \tag{13.6}$$

where $\alpha \in [0, 1]$. In the absence of advective flow (or $\boldsymbol{v} = 0$), the above equations become a system of reaction-diffusion of the generic form

$$\frac{\partial [A]}{\partial t} = D_A \nabla^2 [A] + f([A], [B]), \tag{13.7}$$

$$\frac{\partial [B]}{\partial t} = D_B \nabla^2 [B] + g([A], [B]), \tag{13.8}$$

where $f$ and $g$ are two known functions. These two equations form the basis of our discussion in this chapter. Let us first just look at a single equation by setting $u = [A]$ and $f(u) = f([A])$. Under appropriate conditions, it can have travelling wave solutions and instability.

## 13.2 Travelling Wave

The nonlinear reaction-diffusion equation

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u), \tag{13.9}$$

can have the travelling wave solution under appropriate conditions of $f(0) = f(1) = 0$, $f(u) > 0$ for $u \in (0, 1)$, and $f'(0) > 0$. For example, $f(u) = \gamma u(1 - u)$ satisfies these conditions, and the equation for this special case is called the Kolmogorov-Petrovskii-Piskunov (KPP) equation. By assuming that the travelling wave solution has the form $u(\zeta)$ and $\zeta = x - vt$, and substituting into the above equation, we have

$$Du''(\zeta) + vu'(\zeta) + f(u(\zeta)) = 0. \tag{13.10}$$

This a second-order ordinary differential equation that can be solved with the appropriate boundary conditions

$$u(-\infty) \to 1, \qquad u(\infty) \to 0. \tag{13.11}$$

The KPP theory suggests that the speed of the travelling wave is

$$v \geq 2\sqrt{Df'(0)}. \tag{13.12}$$

The KPP equation will be applied to study the transport process of bacteria in porous media in Chapter 15.

## 13.3 Pattern Formation

One of the most studied nonlinear reaction-diffusion equations in the 2-D case is the KPP equation

$$\frac{\partial u}{\partial t} = D(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) + \gamma q(u), \tag{13.13}$$

and

$$q(u) = u(1-u). \tag{13.14}$$

This KPP equation can describe a huge number of physical, chemical and biological problems. The most interesting feature of this nonlinear system is its ability to generate beautiful patterns.

We can solve it using the same finite difference scheme as in Chapter 10 by applying the periodic boundary conditions and using a random initial condition. Using the central difference for spatial coordinates and the explicit Euler scheme for time, we have

$$\frac{u_{i,j}^{(n+1)} - u_{i,j}^{(n)}}{\Delta t} = D[\frac{u_{i+1,j}^{(n)} - 2u_{i,j}^{(n)} + u_{i-1,j}^{(n)}}{(\Delta x)^2} + \frac{u_{i,j+1}^{(n)} - 2u_{i,j}^{(n)} + u_{i,j-1}^{(n)}}{(\Delta y)^2}]$$

$$+\gamma u_{i,j}^{(n)}(1 - u_{i,j}^{(n)}). \tag{13.15}$$

Using $\Delta x = \Delta y = \Delta t = 1$, we have

$$u_{i,j}^{(n+1)} = D[u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n)}] + (1-4D)u_{i,j}^{(n)} + \gamma u_{i,j}^{(n)}[1 - u_{i,j}^{(n)}],$$

which has been implemented using Matlab and Octave (about 15 lines), and the programs are given in Appendix B.

Figure 13.1 shows the pattern formation of the above equation on a $500 \times 500$ grid for $D = 0.2$ and $\gamma = 0.5$. We can see that rings and thin curves are formed, arising from the random initial condition: $u^{(0)} = \text{random}(n, n) \in [0, 1]$ where $n$ is the size of the grid.

If the formed patterns are plotted on a landscape surface as shown in Figure 13.2, the variations in the values of the field $u(x, y)$ can be easily demonstrated.

If you use the programs provided in Appendix B to run the simulations, you will see that the pattern emerges naturally from the initially random background. Once the pattern is formed, it evolves gradually with time, but the characteristics such as the shape and structure of the patterns do not change much with time. In this sense, one can see beautiful and stable patterns.

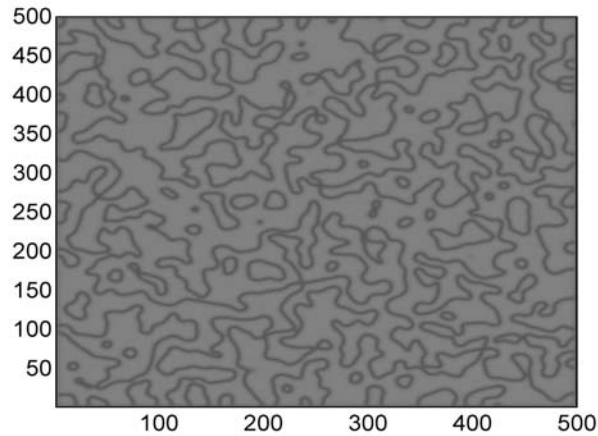



Figure 13.1: 2-D pattern formation for $D = 0.2$ and $\gamma = 0.5$.

## 13.4 Reaction-Diffusion System

The pattern formation in the previous section arises naturally from a single equation of nonlinear reaction-diffusion type. In many applications, we often have to simulate a system of nonlinear reaction-diffusion equations, and the variables are coupled in a complicated manner.

The pattern formation in the previous section comes from the instability of the nonlinear reaction-diffusion system. In order to show this, let us use the following mathematical model for enzyme inhibition and cooperativity:

$$\frac{\partial u}{\partial t} = D_u(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) + f(u, v), \tag{13.16}$$

$$\frac{\partial v}{\partial t} = D_v(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}) + g(u, v), \tag{13.17}$$

and

$$f(u, v) = \gamma u(1 - u), \tag{13.18}$$

$$g(u, v) = \beta(u - \alpha u^2 v), \tag{13.19}$$

where $D_u$ and $D_v$ are diffusion coefficients, while $\alpha, \beta, \gamma$ are all constants. This reaction-diffusion system may have instability if certain conditions are met.

The steady state solutions are obtained from $f(u_0, v_0) = 0$ and $g(u_0, v_0) = 0$. They are

$$u_0 = 1, \qquad v_0 = \frac{1}{\alpha}, \tag{13.20}$$

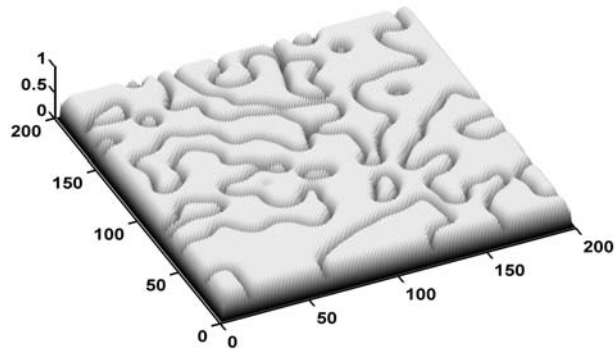



Figure 13.2: Surface of 2-D pattern formation.

or

$$u_0 = 0, \qquad v_0 = 0. \tag{13.21}$$

The latter solution $u_0 = v_0 = 0$ is trivial as it will give zero solution everywhere. Let $\boldsymbol{\psi} = (\psi_u, \psi_v)^T = (u - u_0, v - v_0)^T$ be a small perturbation such that $u = u_0 + \psi_u$ and $v = v_0 + \psi_v$. After substituting into Eqs.(13.16) and (13.17) and neglecting the nonlinear terms involving $\psi_u^2$ and $\psi_v^2$ so that $(u_0 + \psi_u)^2 = u_0^2 + 2u_0\psi_u + \psi_u^2 \approx u_0^2 + 2u_0\psi_u$, we then get the following linearised equations

$$\frac{\partial \psi_u}{\partial t} = D_u \nabla^2 \psi_u + [\gamma(1 - u_0) - \gamma u_0]\psi_u, \tag{13.22}$$

$$\frac{\partial \psi_v}{\partial t} = D_v \nabla^2 \psi_v + \beta[1 - 2\alpha u_0 v_0]\psi_u - \beta\alpha u_0^2 \psi_v. \tag{13.23}$$

where we have used the solutions for steady state. Writing them in a compact form for $\boldsymbol{\psi}$, we have

$$\frac{\partial \boldsymbol{\psi}}{\partial t} = D\nabla^2 \boldsymbol{\psi} + M\boldsymbol{\psi}, \tag{13.24}$$

where

$$D = \begin{pmatrix} D_u & 0 \\ 0 & D_v \end{pmatrix}, \tag{13.25}$$

and

$$M = \begin{pmatrix} \gamma(1 - 2u_0) & 0 \\ \beta(1 - 2\alpha u_0 v_0) & -\beta\alpha u_0^2 \end{pmatrix}. \tag{13.26}$$

Writing $\boldsymbol{\psi}$ in the form of Fourier components

$$\boldsymbol{\psi} = \sum e^{\lambda t + i\boldsymbol{k}\cdot\boldsymbol{x}} \boldsymbol{\psi}_k, \tag{13.27}$$

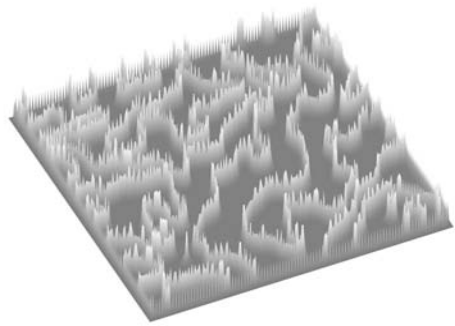


Figure 13.3: Pattern formation on a 200×200 grid for $D_u = 0.2$, $D_v = 0.1$, $\alpha = 1$, $\gamma = 0.5$, and $\beta = 0.5$.

where the summation is over all the wavenumber $\boldsymbol{k} = (k_x, k_y)^T$, and the dot product is $\boldsymbol{k} \cdot \boldsymbol{x} = k_x x + k_y y$. We finally have

$$|M - \lambda I - Dk^2| = 0, \tag{13.28}$$

where I is a $2 \times 2$ identity matrix, and $k^2 = \boldsymbol{k} \cdot \boldsymbol{k}$. This eigenvalue equation has two roots: $\lambda = \gamma(1-2u_0) - D_u k^2$ and $\lambda = -\beta\alpha u_0^2 - D_v k^2$. Since $\Re(\lambda) > 0$ implies that instability may occur, this requires that

$$D_u k^2 < \gamma(1 - 2u_0), \qquad D_v k^2 < -\beta\alpha u_0^2. \tag{13.29}$$

For example, for typical values of $\alpha = 1$, $\beta = \gamma = 0.5$, $D_u = 0.2$, and $D_v = 0.1$, the instability condition is indeed satisfied. If the unstable criteria are satisfied, any small random perturbation can generate complex patterns. It is worth pointing out that the linear stability analysis here just gives an indication of the potential parameter ranges for pattern formation. For more accurate analysis, we need the nonlinear stability theory.

Figure 13.3 shows the pattern formation of this system with periodic boundary conditions on a $200 \times 200$ grid where the results are mapped onto a surface. The values used in this simulation are $D_u = 0.2$, $D_v = 0.1$, $\alpha = 1$, $\gamma = 0.5$ and $\beta = 0.5$.

# References

Flake G. W., *The Computational Beauty of Nature: Computer Exploration of Fractals, Chaos, Complex Systems and Adapation*, MIT Press,( 2000).

Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

Keener J. and Sneyd J., *A Mathematical Physiology*, Springer-Verlag, New York, (2001).

Korn R. W., Pattern formation in the leaf of zebra grass, *J. Theor. Biol.*, **187**, 449-451 (1997).

Meinhardt H., *Models of Biological Pattern Formation*, Academic Press, London, (1982).

Murray J. D., *Mathematical Biology*, Springer-Verlag, New York, (1998).

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

Yang X. S., Computational modelling of nonlinear calcium waves, *Appl. Maths. Modelling*, **30**, 200-208 (2006).

Yang X. S. and Young Y., Cellular automata, PDEs and pattern formation (Chapter 18), in *Handbook of Bioinspired Algorithms*, edited by Olarius S. and Zomaya A. Y., Chapman & Hall/CRC, (2005).

# Chapter 14

# Elasticity and Poroelasticity

## 14.1 Hooke's Law and Elasticity

The basic Hooke's law of elasticity concerns an elastic body such as a spring, and it states that the extension $x$ is proportional to the load $F$, that is

$$F = kx, \tag{14.1}$$

where $k$ is the spring constant. However, this equation only works for 1-D deformations.

For a slender elastic body under a given tensile load $F$ shown in Fig. 14.1, the strain is defined as

$$\varepsilon_n = \frac{L - L_0}{L_0}, \tag{14.2}$$

where $L_0$ is the initial length and $L$ is the current length under load

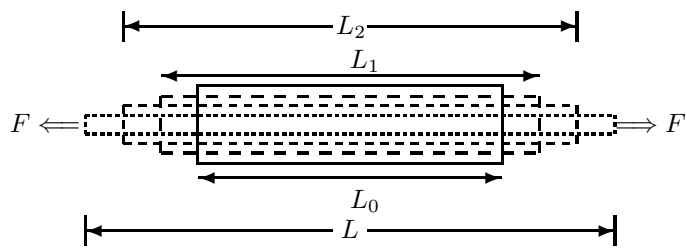



Figure 14.1: Strains at various stages of deformation.

$F$. The strain defined this way is called nominal strain $\varepsilon_n$. In reality, the deformation or elongation does not occur instantaneously. It could be considered in multiple stages from the initial $L_0$ to the final $L$. For simplicity, we only consider three stages $L_0$ to $L_1$, $L_2$ and $L$. We have

$$\varepsilon_{n1} = \frac{L_1 - L_0}{L_0}, \varepsilon_{n2} = \frac{L_2 - L_1}{L_1}, \varepsilon_{n3} = \frac{L - L_2}{L_2}. \tag{14.3}$$

As the system is linear elasticity, the total strain $\varepsilon_n$ should be equal to the sum of each stage, but

$$\varepsilon_{n1} + \varepsilon_{n2} + \varepsilon_{n3} \neq \varepsilon_n. \tag{14.4}$$

It seems that this definition of strain causes a problem.

Another definition of strain is the true strain $\varepsilon_t$ given by

$$\varepsilon_t = \ln(\frac{L}{L_0}). \tag{14.5}$$

If we use this definition, the strains at each stage are

$$\varepsilon_{t1} = \ln \frac{L_1}{L_0}, \;\; \varepsilon_{t2} = \ln \frac{L_2}{L_1}, \;\; \varepsilon_{t3} = \ln \frac{L}{L3}. \tag{14.6}$$

We can see that

$$\varepsilon_{t1} + \varepsilon_{t2} + \varepsilon_{t3} = \ln[\frac{L_1}{L_0} \times \frac{L_2}{L_1} \times \frac{L}{L_2}] = \ln \frac{L}{L_0} = \varepsilon_t, \tag{14.7}$$

which means that the total true strain is equal to the sum of true strains at each stage. This is a better definition, but it involves the logarithm.

From the definition of true strain, we know that

$$\varepsilon_t = \ln(\frac{L}{L_0}) = \ln[1 + \frac{(L - L_0)}{L_0}] = \ln(1 + \varepsilon_n). \tag{14.8}$$

When $\varepsilon_n \ll 1$, $\ln(1 + \varepsilon_n) \approx \varepsilon_n$, and we have $\varepsilon_t \approx \varepsilon_n$. In most elastic problems we meet, the strains are small, thus we can simply use the approximation $\varepsilon_t \approx \varepsilon_n$. In this case, two definitions of strains are essentially the same. However, for large strain deformation such as elongation of elastic rubber bands and polythene, the true strain is more convenient.

For the discussion of Hooke's law in the case of a bar of uniform cross section with a length $L$ and a cross-sectional area $A$, it is more convenient to use strain $\varepsilon$ and stress $\sigma$. The stress and strain are defined by

$$\sigma = \frac{F}{A}, \qquad \varepsilon = \frac{\Delta L}{L}, \tag{14.9}$$

where $\Delta L = L - L_0$ is the extension. The unit of stress is $\mathrm{N/m^2}$, while the strain is dimensionless, though it is conventionally expressed in m/m or percentage (%) in engineering. For the elastic bar, the stress-strain relationship is

$$\sigma = E\varepsilon, \tag{14.10}$$

where $E$ is the Young's modulus of elasticity. Written in terms $F$ and $x = \Delta L$, we have

$$F = \frac{EA}{L}\Delta L = kx, \qquad k = \frac{EA}{L}, \tag{14.11}$$

where $k$ is the equivalent spring constant for the bar. This equation is still only valid for any unidirectional compression or extension. For the 2-D and 3-D deformation, we need to generalise Hooke's law. For the general stress tensor (also called Cauchy stress tensor)

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}, \tag{14.12}$$

and strain tensor

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{yx} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{zx} & \varepsilon_{zy} & \varepsilon_{zz} \end{pmatrix} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix}, \tag{14.13}$$

it can be proved later that these tensors are symmetric, that is $\boldsymbol{\sigma} = \boldsymbol{\sigma}^T$ and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T$, which leads to

$$\sigma_{xy} = \sigma_{yx}, \quad \sigma_{xz} = \sigma_{zx}, \quad \sigma_{yz} = \sigma_{zy}, \tag{14.14}$$

and

$$\varepsilon_{xy} = \varepsilon_{yx}, \quad \varepsilon_{xz} = \varepsilon_{zx}, \quad \varepsilon_{yz} = \varepsilon_{zy}. \tag{14.15}$$

Therefore, we only have 6 independent components or unknowns for stresses and 6 unknown strain components.

The strain tensor is defined by the displacement $\mathbf{u}^T = (u_1, u_2, u_3)$

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{14.16}$$

where $x_1 = x$, $x_2 = y$, and $x_3 = z$. Sometimes, it is useful to write

$$\boldsymbol{\varepsilon} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T). \tag{14.17}$$

The generalised Hooke's law can be written as

$$\varepsilon_{xx} = \frac{1}{E}[\sigma_{xx} - \nu(\sigma_{yy} + \sigma_{zz})], \tag{14.18}$$

$$\varepsilon_{yy} = \frac{1}{E}[\sigma_{yy} - \nu(\sigma_{xx} + \sigma_{zz})], \tag{14.19}$$

$$\varepsilon_{zz} = \frac{1}{E}[\sigma_{zz} - \nu(\sigma_{xx} + \sigma_{yy})], \tag{14.20}$$

$$\varepsilon_{xy} = \frac{1+\nu}{E}\sigma_{xy}, \tag{14.21}$$

$$\varepsilon_{xz} = \frac{1+\nu}{E}\sigma_{xz}, \tag{14.22}$$

$$\varepsilon_{yz} = \frac{1+\nu}{E}\sigma_{yz}, \tag{14.23}$$

where $\nu$ is the Poisson's ratio, and it measures the tendency of extension in transverse directions (say, $x$ and $y$) when the elastic body is stretched in one direction (say, $z$).

Poisson's ratio can be defined as the ratio of the transverse contraction strain (normal to the applied load) to the axial strain in the direction of the applied tensile stress (see Fig. 14.2). That is

$$\nu = -\frac{\varepsilon_{\text{transverse}}}{\varepsilon_{\text{axial}}}. \tag{14.24}$$

For a perfectly incompressible material, $\nu = 0.5$, and $\nu = 0 \sim 0.5$ for most common materials. For example, $\nu = 0.25 \sim 0.3$ for steels, $\nu \approx 0.5$ for rubber, and $\nu \approx 0$ for the cork of a wine bottle. Some auxetic materials such as polymer foams or anti-rubbers have a negative Poisson's ratio $\nu < 0$. In fact, the isotropic upper limit of $\nu$ can reach $\nu \to 0.5$ which corresponds to $\lambda$ and $K \to \infty$, while the lower limit may approach $\nu \to -1$ which corresponds to $G \to \infty$ [see Eq.(14.34)]. We will see later that the Poisson's ratio of rocks will affect the speed of propagation of waves commonly discussed in earth sciences.

This generalised Hooke's law can be written concisely as

$$\varepsilon_{ij} = \frac{1+\nu}{E}\sigma_{ij} - \frac{\nu}{E}\sigma_{kk}\delta_{ij}, \tag{14.25}$$

where we have used Einstein's summation convention $\sigma_{kk} = \sigma_{xx} + \sigma_{yy} + \sigma_{zz}$. Another related quantity is pressure, which is defined by

$$p = -\frac{1}{3}\sigma_{kk} = -\frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3}. \tag{14.26}$$

The negative sign comes from the conventions that a positive normal stress results in tension, and negative one in compression, while the

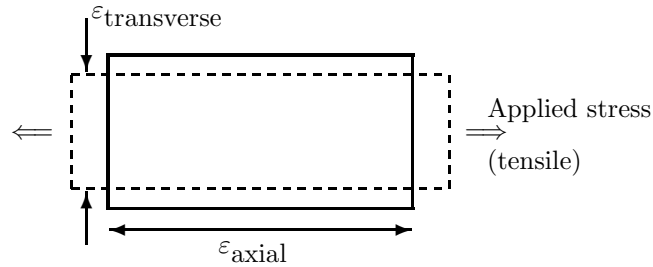



Figure 14.2: Poisson's ratio $\nu = -\varepsilon_{\text{transverse}}/\varepsilon_{\text{axial}}$.

positive pressure acts in compression. Sometimes, it is more convenient to express the stress tensor in terms of pressure and devitoric stress tensor $s_{ij}$

$$\sigma_{ij} = -p\delta_{ij} + s_{ij}. \tag{14.27}$$

If we want to invert Eq.(14.25), we must first express $\sigma_{kk}$ in terms of $\varepsilon_{kk}$ so that the right-hand side of the new expression does not contain the stress $\sigma_{kk}$. By contraction using $j \to i$, we have

$$\varepsilon_{ii} = \frac{1+\nu}{E}\sigma_{ii} - \frac{\nu}{E}\sigma_{kk}\delta_{ii} = \frac{1-2\nu}{E}\sigma_{ii}, \tag{14.28}$$

where we have used $\delta_{ii} = \delta_{11} + \delta_{22} + \delta_{33} = 1 + 1 + 1 = 3$ and $\sigma_{ii} = \sigma_{kk}$.

In engineering and poroelastic theory, the quantity

$$\varepsilon_{kk} = \varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz} = \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} + \frac{\partial u_3}{\partial z} = \nabla \cdot \boldsymbol{u}, \tag{14.29}$$

means a fractional change in volume, known as dilation or volumetric strain. It is often written as $\Delta = \varepsilon_{kk} = \nabla \cdot \mathbf{u}$. This gives that

$$\sigma_{ii} = \sigma_{kk} = \frac{E}{1-2\nu}\varepsilon_{kk}. \tag{14.30}$$

Substituting it into Eq.(14.25), we have

$$\varepsilon_{ij} = \frac{1+\nu}{E}\sigma_{ij} - \frac{\nu}{E}(\frac{E}{1-2\nu}\varepsilon_{kk})\delta_{ij}, \tag{14.31}$$

or after some rearrangement

$$\frac{1+\nu}{E}\sigma_{ij} = \varepsilon_{ij} + \frac{\nu}{1-2\nu}\varepsilon_{kk}\delta_{ij}, \tag{14.32}$$

which can be written as

$$\sigma_{ij} = 2G\varepsilon_{ij} + \lambda\varepsilon_{kk}\delta_{ij}, \tag{14.33}$$

where $\mu$ and $\lambda$ are Lamé constants. They are

$$G = \mu = \frac{E}{2(1+\nu)}, \qquad \lambda = \frac{\nu E}{(1+\nu)(1-2\nu)} = \frac{2\nu G}{(1-2\nu)}. \tag{14.34}$$

This stress-strain relationship can also be written as

$$\boldsymbol{\sigma} = 2G\boldsymbol{\varepsilon} + \lambda(\nabla \cdot \boldsymbol{u})\boldsymbol{\delta}. \tag{14.35}$$

Here $G = \mu$ is called the shear modulus, while $K$ is called the bulk modulus which is the ratio of pressure $-p$ to the volumetric strain (dilation) under hydrostatic loading. We have

$$K = \frac{-p}{\Delta}. \tag{14.36}$$

For hydrostatic loading, $\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = -p$, and the combination of Eqs.(14.28) and (14.29) leads to

$$\Delta = -\frac{(1-2\nu)}{3E}p, \tag{14.37}$$

which gives

$$K = \frac{E}{3(1-2\nu)}. \tag{14.38}$$

For most rocks, $\nu = 1/4 \sim 1/3$. If $\nu = 1/3$, then $K = E$, and $G = 3E/8$. For rubber, $\nu \approx 0.5$, $K \approx \infty$. This means that it is incompressible. That is why bicycle tyres should not be made of solid rubber, which would not be comfortable to ride. On the other hand, the Poisson's ratio for the cork of a wine bottle is almost zero. This means it can easily be inserted or removed from a wine bottle and yet it can withstand the pressure within the bottle. Thus, it is ideal for such a purpose.

## 14.2 Shear Stress

All materials have limited strength; above a certain limit, the material will start to yield and eventually fail. Yield is driven by shearing stresses and sliding along planes or grain boundaries in microcrystals, while fracture is driven by normal stresses (mode I) to separate atomic planes, creating free surfaces. Yield is usually associated with plastic deformation, and fracture can associated with both brittle and plastic materials. The prediction of failure planes, although not straightforward even for the most simple configuration, is not very difficult. For example, under uniaxial tension (or compression), the yield is by

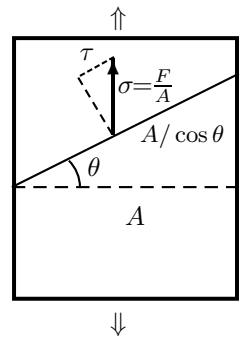



Figure 14.3: Uniaxial tension or compression and failure plane.

shearing and the ultimate failure plane lies on a plane at 45° where the shear stress reaches maximum (see Fig. 14.3). Let $\sigma_y$ be the tensile yield stress (normal stress) in the uniaxial tension when yield occurs, and let $\tau_y$ be the maximum shear stress at yield.

The area of the plane with an angle $\theta$ to the uniaxial tension is

$$A_\theta = \frac{A}{\cos\theta}, \tag{14.39}$$

and the shear stress (the component along the plane per unit area) is

$$\tau = \frac{\sigma \sin\theta}{1/\cos\theta} = \sigma \sin\theta\cos\theta = \frac{\sigma}{2}\sin 2\theta. \tag{14.40}$$

As the maximum $\sin 2\theta = 1$ occurs at $2\theta = \pi/2$, we have the relationship between yield stress $\sigma_y$ and shear stress $\tau_y$

$$\tau_y = \frac{\sigma_y}{2}, \tag{14.41}$$

which occurs on a plane at $\theta = 45°$. In nature, many fractures indeed occur at this angle, and even in a foundation built on a layer of soils, if failure occurs by, say, landslide. However, the real situation is more complicated in soils, and the angle may vary a lot due to the local stress state, cohesion and internal friction in soils.

## 14.3 Equations of Motion

For a general solid where the inertia is not negligible, we have

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + \mathbf{b} = \rho\frac{\partial^2\mathbf{u}}{\partial t^2}, \tag{14.42}$$

where $\rho$ is the density of the elastic body. In some books, the following form of body force $\mathbf{b} = \rho\mathbf{f}$ is used. In this case, the force $\boldsymbol{f}$ means the force per unit mass. Together with the generalised Hooke's law and relationship with displacement $\mathbf{u}$, we have the following set of equations of motion for an elastic body.

$$\frac{\partial \sigma_{ij}}{\partial x_j} + b_i = \rho \frac{\partial^2 u_i}{\partial t^2}, \tag{14.43}$$

$$\sigma_{ij} = 2G\varepsilon_{ij} + \lambda \varepsilon_{kk}\delta_{ij}, \tag{14.44}$$

$$\varepsilon_{ij} = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}). \tag{14.45}$$

There are 15 equations (6 for stresses, 6 for strains, 3 for displacements) and we have 15 unknowns (6 stress components, 6 strain components and 3 displacements). Therefore, the elastic field should be uniquely determined if appropriate boundary conditions are given. There are other conditions such as compatibility equations, and we will briefly discuss them later.

If we write the equations of motion using bold notation (usually for tensors and vectors) notation, we have

$$\nabla \cdot \boldsymbol{\sigma} + \mathbf{b} = \rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2},, \tag{14.46}$$

$$\boldsymbol{\sigma} = 2G\boldsymbol{\varepsilon} + \lambda(\nabla \cdot \boldsymbol{u})\boldsymbol{\delta}, \tag{14.47}$$

$$\boldsymbol{\varepsilon} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T). \tag{14.48}$$

If we substitute the generalised Hooke's law and displacement into the first equation (14.46), we have

$$\nabla \cdot [2G\boldsymbol{\varepsilon} + \lambda(\nabla \cdot \boldsymbol{u})\boldsymbol{\delta}] + \mathbf{b} = \rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2}, \tag{14.49}$$

or

$$\nabla \cdot [G(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T) + \lambda(\nabla \cdot \boldsymbol{u})\boldsymbol{\delta}] + \mathbf{b} = \rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2}, \tag{14.50}$$

which leads to

$$(G + \lambda)\nabla(\nabla \cdot \boldsymbol{u}) + G\nabla^2 \boldsymbol{u} + \mathbf{b} = \rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2}. \tag{14.51}$$

Using $G + \lambda = G/(1 - 2\nu)$ and after some rearrangements, we have

$$\rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2} = \frac{G}{1 - 2\nu}\nabla(\nabla . \boldsymbol{u}) + G\nabla^2 \boldsymbol{u} + \mathbf{b}, \tag{14.52}$$

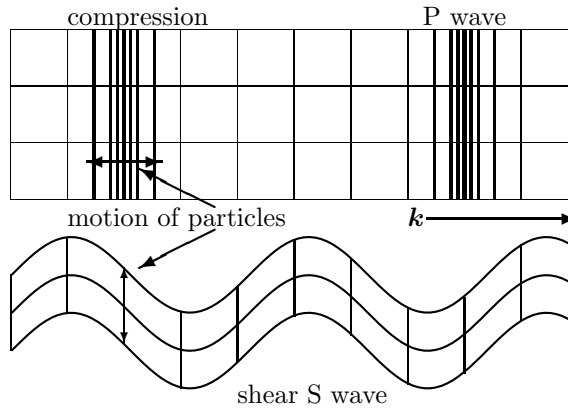



Figure 14.4: Seismic waves: P wave and S wave.

which is the well-known Cauchy-Navier equation. This equation supports both longitudinal wave ($P$ wave) and transverse wave ($S$ wave). As shown in Fig. 14.4, the P waves are the primary or compressional waves whose particle motion is parallel to the direction $\boldsymbol{k}$ of the propagating wave, while the S waves are the secondary or shear waves whose particle motion is perpendicular to the wave direction $\boldsymbol{k}$. For example, the sound waves in the air are compressional waves while the vibrations of a guitar string generate S waves.

In the simplest 1-D case without any (external) body force $\mathbf{b}$, we can take $\nabla \cdot \boldsymbol{u} = 0$ for $S$-wave and $\boldsymbol{u} \to u_1$, we simply have

$$\rho \frac{\partial^2 u_1}{\partial t^2} = G \frac{\partial^2 u_1}{\partial x^2}, \tag{14.53}$$

thus its wave speed is

$$v_S = \sqrt{\frac{G}{\rho}}. \tag{14.54}$$

For the $P$ waves in 1-D, the displacement field is non-rotational, *i.e.*, $\nabla \times (\nabla \times \boldsymbol{u}) = 0$. From the identity $\nabla(\nabla \cdot \boldsymbol{u}) = \nabla \times (\nabla \times \boldsymbol{u}) + \nabla^2 \boldsymbol{u}$, the 1-D Cauchy-Navier equation becomes

$$\rho \frac{\partial^2 u_1}{\partial t^2} = (\lambda + 2G) \frac{\partial^2 u_1}{\partial x^2}. \tag{14.55}$$

Then, the speed of $P$-wave is

$$v_P = \sqrt{\frac{(\lambda + 2G)}{\rho}}. \tag{14.56}$$

Since $\lambda+2G > G$, therefore, $P$-waves always travel faster than $S$-waves. In fact, using Eq.(14.34), we can write the ratio of the speed of $P$-waves to that of $S$-waves as

$$\frac{v_P}{v_S} = \sqrt{\frac{\lambda+2G}{G}} = \sqrt{\frac{K+\frac{4}{3}G}{G}} = \sqrt{\frac{2(1-\nu)}{1-2\nu}}. \tag{14.57}$$

We can see that this ratio is solely determined by the Poisson's ratio $\nu$. For the Earth's crust, $\nu = 0.2 \sim 0.33$, thus the inverse of the ratio

$$r_v = \frac{v_S}{v_P} = \sqrt{\frac{1-2\nu}{2(1-\nu)}} \approx 0.50 \sim 0.65. \tag{14.58}$$

For example, in the middle part of the Earth's crust, if we use a typical (or average) value of $E \approx 80$ GPa, $\nu \approx 0.25$ and $\rho \approx 2700$ kg/m$^3$, we have

$$v_S \approx \sqrt{\frac{G}{\rho}} = \sqrt{\frac{E}{2(1+\nu)\rho}} \approx 3400\text{m/s}. \tag{14.59}$$

Since $r_v \approx 0.57$, thus the speed of P waves would be typically 3400/0.57 $\approx$ 5900 m/s. In fact, P waves travel typically at speeds between 2 to 7.5 km/s in the Earth's crust while S waves travel commonly at 50% to 70% of the speed of P waves.

An interesting property of P and S waves in the Earth's crust is that the average ratio $r_v$ is relatively constant. Thus seismologists can use this fact to estimate the distance $d$ of the earthquake from observation stations from the recorded time delay $\tau = t_S - t_P = d/v_S - d/v_P$ from the first arrival of the P wave to the arrival of the S wave. The distance $d$ is most cases is about

$$d = \tau v_0, \qquad v_0 = \frac{1}{(\frac{1}{v_S} - \frac{1}{v_P})}.$$

For the distance range from 50km to 500km, $v_S \approx 3.4$ km/s and $v_P \approx 6$ km/s, we have a typical value of $v_0 \approx 8$ km/s.

---

**Example 14.1:** Although it is possible to estimate the epicentre of an earthquake from a single station using multiple component recording, it is more accurate to use three or more observation stations. Assume there are three seismic stations at $A$, $B$ and $C$, and they calculated the S-P time delay from their recorded seismographs and estimated that the distances were 200km, 100km, and 125km from $A$, $B$ and $C$, respectively. In order to locate the source of the earthquake, we first draw a circle centred at $A$ with a radius of 200km. The source could anywhere on the circle. Now

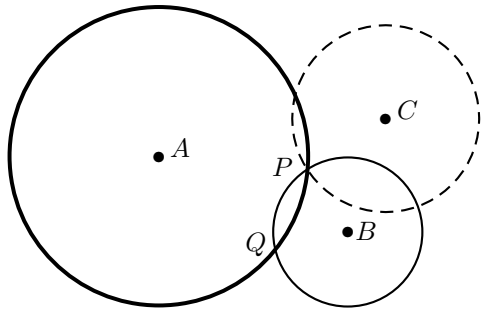



Figure 14.5: Locating the source $P$ of an earthquake from three observation stations: $A$, $B$ and $C$.

we draw a second circle centred at $B$ with a radius of 100km, and the second circle intersects with the first at two locations $P$ and $Q$. The two locations are the potential epicentres. In order to determine the location uniquely, we now draw a third circle centred at $C$ with a radius of 125km, which intersects with both of the other two circles at $P$. Therefore, $P$ is the estimated location of the earthquake. Of course, if we use more seismic stations we can estimate the epicentre more accurately. This is why seismologists usually have a network of seismic stations.

Furthermore, from the definitions of the strain components in terms of displacements $\boldsymbol{u}^T = (u_1, u_2, u_3) = (u, v, w)$, we have

$$\varepsilon_{xx} = \frac{\partial u}{\partial x}, \qquad \varepsilon_{yy} = \frac{\partial v}{\partial y}, \qquad \varepsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}). \tag{14.60}$$

By assuming the displacements are continuous and differentiable functions of positions, we differentiate $\varepsilon_{xx}$ with respect to $y$ twice, we have

$$\frac{\partial^2 \varepsilon_{xx}}{\partial y^2} = \frac{\partial^3 u}{\partial x \partial y^2}. \tag{14.61}$$

Similarly, differentiating $\varepsilon_{yy}$ with respect to $x$ twice, we have

$$\frac{\partial^2 \varepsilon_{yy}}{\partial x^2} = \frac{\partial^3 v}{\partial y \partial x^2}. \tag{14.62}$$

Now differentiating $\varepsilon_{xy}$ with respect to $y$ once, and with respect to $x$ once, we have

$$\frac{\partial \varepsilon_{xy}}{\partial x \partial y} = \frac{1}{2}[\frac{\partial^3 u}{\partial x \partial y^2} + \frac{\partial^3 v}{\partial y \partial x^2}] = \frac{1}{2}[\frac{\partial^2 \varepsilon_{xx}}{\partial y^2} + \frac{\partial^2 \varepsilon_{yy}}{\partial x^2}], \tag{14.63}$$

where we have used the interchangeability of partial derivatives $\frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x}$. This can be rearranged as

$$\frac{\partial^2 \varepsilon_{xx}}{\partial y^2} + \frac{\partial^2 \varepsilon_{yy}}{\partial x^2} = 2\frac{\partial^2 \varepsilon_{xy}}{\partial x \partial y}, \tag{14.64}$$

which is the compatibility equation. In the same fashion, we can derive other compatibility equations

$$\frac{\partial^2 \varepsilon_{zz}}{\partial y^2} + \frac{\partial^2 \varepsilon_{yy}}{\partial z^2} = 2\frac{\partial^2 \varepsilon_{yz}}{\partial y \partial z}. \tag{14.65}$$

$$\frac{\partial^2 \varepsilon_{xx}}{\partial z^2} + \frac{\partial^2 \varepsilon_{zz}}{\partial x^2} = 2\frac{\partial^2 \varepsilon_{xz}}{\partial x \partial z}. \tag{14.66}$$

## 14.4 Euler-Bernoulli Beam Theory

The Euler-Bernoulli beam theory is a simplified theory for calculating the deflection of beams under a distribution of load force using the linear isotropic theory of elasticity. The basic assumptions for the beam theory are: 1) the beam is isotropic and elastic; 2) the beam deformation is dominated by bending, while distortion and rotation are negligible; 3) the beam is long and slender with a constant cross section along the axis. Under these assumptions, we can now derive the governing equations.

Let $u(x,t)$ be the deflection of the beam (shown in Figure 14.6), $A$ be the area of the cross section, and $f(x,t)$ be the force per unit length. The first assumption implies that the bending moment $M$ is proportional to the curvature $\kappa$ of the bending. That is

$$M = EI\kappa, \qquad \kappa = \frac{\frac{\partial^2 u}{\partial x^2}}{[1 + (\frac{\partial u}{\partial x})^2]^{3/2}}, \tag{14.67}$$

where $E$ is the Young's modulus and $I$ is the area moment of the beam's cross section. In mechanics, $I$ is also called the second moment of area or the area moment of inertia. The area moment about a horizontal axis through the centroid is defined by

$$I = \int_\Omega y^2 dA, \tag{14.68}$$

which has a unit of $[\mathrm{m}]^4$, and it should not be confused with the mass moment of inertia $J$ (also often denoted as $I$, but we use $J$ here) about an axis, which is defined by

$$J = \int_\Omega r^2 dm = \int_\Omega \rho r^2 dxdydz \tag{14.69}$$

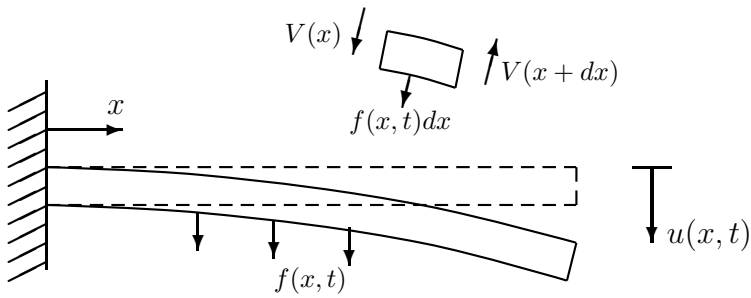



Figure 14.6: Beam bending.

with a unit of [Kg] $[m]^2$. Neither $E$ nor $I$ change along the $x$-axis. For a cylindrical rod with a radius of $R$, we have $I = \pi R^4/4$. For a rectangular beam with a base width of $b$ and a depth of $h$, we have $I = bh^3/12$.

The second assumption means that the shear $V(x)$ is related to the bending moment

$$\frac{\partial M}{\partial x} = V(x), \tag{14.70}$$

and the third assumption means $\frac{\partial u}{\partial x} \ll 1$. Therefore, we have

$$M \approx EI\frac{\partial^2 u}{\partial x^2}, \tag{14.71}$$

or

$$V \approx \frac{\partial}{\partial x}(EI\frac{\partial^2 u}{\partial x^2}). \tag{14.72}$$

For a small volume element (also shown in Figure 14.6), the mass of the element is $\rho A dx$ where $\rho$ is the density, and the acceleration is $\frac{\partial^2 u}{\partial t^2}$. The shear force variation is given by $V(x + dx) = V(x) + \frac{\partial V}{\partial x}dx$, and the total force is

$$V(x) - V(x + dx) + f(x, t)dx = [f(x, t) - \frac{\partial V}{\partial x}]dx. \tag{14.73}$$

Using the Newton's second law of motion, we have

$$f(x, t) - \frac{\partial V}{\partial x} = \rho A\frac{\partial^2 u}{\partial t^2}. \tag{14.74}$$

Substituting the above expression for $V$, we have

$$\rho A\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2}{\partial x^2}[EI\frac{\partial^2 u}{\partial x^2}] = f(x), \tag{14.75}$$

which is the Euler-Bernoulli equation. If there is no force $f(x,t) = 0$, the equation becomes a homogeneous form

$$\rho A \frac{\partial^2 u}{\partial t^2} + \frac{\partial^2}{\partial x^2}[EI\frac{\partial^2 u}{\partial x^2}] = 0, \tag{14.76}$$

which is a fourth-order wave equation. It governs the waves that travel along a beam, a lighting pole or any slender column.

For the elastostatic problem, $\frac{\partial^2 u}{\partial t^2} \approx 0$, we have

$$\frac{\partial^2}{\partial x^2}[EI\frac{\partial^2 u}{\partial x^2}] = q(x), \tag{14.77}$$

where $q(x) = f(x)$ is the applied force per unit length. This equation will be used to determine the deflection of a beam.

Let us now use the Euler-Bernoulli theory to calculate the shape of a cantilever with a uniform cross section under a point load.

---

**Example 14.2:** For a point load or concentrated load $P$ at $x = L$, we have $q = -P\delta(x-L)$ where $\delta(x-L)$ is a Dirac delta function with $\delta = 1$ when $x = L$ and $\delta = 0$ everywhere else. The integral of $\delta(x-L)$ is

$$H(x-L) = \int \delta(x-L)dx,$$

where $H(x-L)$ is the heaviside function with $H = 1$ when $x \geq L$ and $H = 0$ when $0 \leq x < L$. Using

$$\int H(x-L)dx = (x-L)H(x-L),$$

and integrating the equation $EI\frac{d^4u}{dx^4} = q$ twice, we have the moment $M$

$$M = EI\frac{d^2u}{dx^2} = Ax + B,$$

where $A$ and $B$ are two integration constants. At the free end $x = L$, The shear is constant (due to a point load $P$) and the moment is zero. Thus, we have $A = -P$, and $M(L) = AL + B = 0$, or $B = PL$. Now we have

$$EI\frac{d^2u}{dx^2} = -Px + PL.$$

Integrating it twice again, we have

$$EIu = -\frac{P}{6}x^3 + \frac{PL}{2}x^2 + Cx + D,$$

where $C$ and $D$ are two constants. Since the cantilever is fixed at $x = 0$, we have $u = \frac{du}{dx} = 0$ at $x = 0$, which leads to $C = 0$ from $u = 0$ and $D = 0$ from $u_x = 0$. Finally, we have

$$u = -\frac{P}{6EI}x^3 + \frac{PL}{2EI}x^2.$$

The end deflection at $x = L$ is

$$d = -\frac{PL^3}{6EI} + \frac{PL^3}{2EI} = \frac{PL^3}{3EI}.$$

Here we has derived that the end deflection for a point load $P$ at $x = L$ is given by

$$d = \frac{PL^3}{3EI}. \tag{14.78}$$

The deflection under both distributed and point loads can be obtained using superposition principles and the total deflection is their sum under appropriate constraints.

## 14.5 Airy Stress Functions

Some problems in geomechanics and fracture mechanics concern the solution of problems within a plane. In this case, we are dealing with plane strain and plane stress problems. For a plane stress problem, we assume that $\sigma_{zz} = 0$ (but $\varepsilon_{zz} \neq 0$), then the plane stress problem involves no stress components depending on $z$. That is to say $\sigma_{xz} = \sigma_{yz} = \sigma_{zz} = 0$. We have only three independent stress components $\sigma_{xx}, \sigma_{yy}$, and $\sigma_{xy}$. The generalised Hooke's law reduces to

$$\varepsilon_{xx} = \frac{1}{E}(\sigma_{xx} - \nu\sigma_{yy}),\ \varepsilon_{yy} = \frac{1}{E}(\sigma_{yy} - \nu\sigma_{xx}),\ \varepsilon_{xy} = \frac{1+\nu}{E}\sigma_{xy}. \tag{14.79}$$

However,

$$\varepsilon_{zz} = \frac{-\nu}{1-\nu}(\varepsilon_{xx} + \varepsilon_{yy}), \tag{14.80}$$

which is not zero in general.

For plane strain problems, it is assumed that $\varepsilon_{zz} = 0$. Thus, there are only three independent strain components $\varepsilon_{xx}, \varepsilon_{yy}$, and $\varepsilon_{xy}$, however, the stress $\sigma_{zz} = \nu(\sigma_{xx} + \sigma_{yy})$ is not zero. The compatibility equation becomes

$$\frac{\partial^2 \varepsilon_{xx}}{\partial y^2} + \frac{\partial^2 \varepsilon_{yy}}{\partial x^2} = 2\frac{\partial^2 \varepsilon_{xy}}{\partial x \partial y}. \tag{14.81}$$

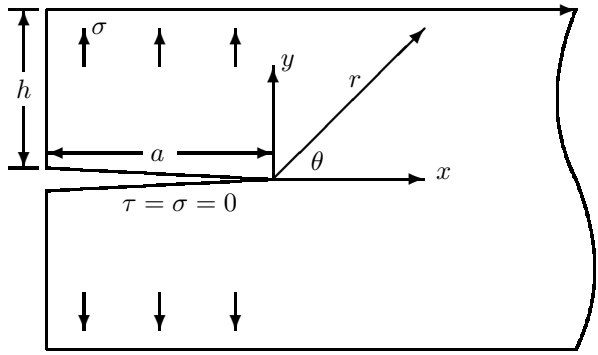



Figure 14.7: An edge crack in an elastic plate.

For plane strain problems with no body forces, the equilibrium equations are automatically satisfied if the stress components are related to a scalar function $\Phi$, called Airy's stress function. The Airy's stress function is defined by

$$\sigma_{xx} = \frac{\partial^2 \Phi}{\partial y^2}, \qquad \sigma_{yy} = \frac{\partial^2 \Phi}{\partial x^2}, \qquad \sigma_{xy} = -\frac{\partial^2 \Phi}{\partial x \partial y}. \tag{14.82}$$

In this case, the compatibility equation becomes

$$\nabla^2(\nabla^2 \Phi) = 0, \tag{14.83}$$

which is a biharmonic equation and can be written as

$$\nabla^4 \Phi = 0. \tag{14.84}$$

In cylindrical polar coordinates $(r, \theta, z)$, it becomes

$$[\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}]^2 \Phi = 0. \tag{14.85}$$

Now consider an edge crack in a semi-infinite elastic medium or solid as shown in Figure 14.7, assuming the elastic body deforms in plane strain. The far field is subjected to bounded stress $\sigma$ at infinity. In order to solve the governing equation to estimate the stress distribution, we assume that the size of the elastic body is much greater than the length $a$ of the edge crack. The surfaces of the crack shall be stress free, which leads to the following boundary conditions

$$\sigma_{\theta\theta} = \frac{\partial^2 \Phi}{\partial r^2} = 0, \ \sigma_{r\theta} = -\frac{\partial}{\partial r}(\frac{1}{r}\frac{\partial \Phi}{\partial \theta}) = 0, \text{ at } \ \theta = \pm\pi. \tag{14.86}$$

Let us try a solution of the form

$$\Phi = r^{n+1} f(\theta), \tag{14.87}$$

and substitute it into the governing biharmonic equation. We get

$$[\frac{d^2}{d\theta^2} + (n+1)^2][\frac{d^2}{d\theta^2} + (n-1)^2] f(\theta) = 0. \tag{14.88}$$

As the second-order equation $y'' + \lambda^2 y = 0$ has a general solution $y = A\sin\lambda\theta + B\cos\lambda\theta$, here we can use this method twice. The general solution takes the following form

$$f(\theta) = A\cos(n+1)\theta + B\sin(n+1)\theta + C\cos(n-1)\theta + D\sin(n-1)\theta.$$

The boundary conditions become

$$\sigma_{\theta\theta} = r^{n-1} n(n+1) f(\theta), \tag{14.89}$$

and

$$\begin{aligned}\sigma_{r\theta} = r^{n-1} n\{&(n+1)[A\sin(n+1)\theta - B\cos(n+1)\theta] \\ &+(n-1)[C\sin(n-1)\theta - D\cos(n-1)\theta]\},\end{aligned} \tag{14.90}$$

at $\theta = \pm\pi$. We know $n = 0$ is trivial as it means that the stress is zero everywhere. From the first equation, we have

$$\sin(2n\pi) = 0, \qquad n = \pm\frac{1}{2}, \pm 1, \pm\frac{3}{2}, ..., \tag{14.91}$$

and $r^n (n \geq 1)$ does not converge, therefore, they are not suitable solutions. The constraint now becomes $n \leq 0$, but the solution has singularity as $r \to 0$. This is, however, acceptable in the crack propagation as the stress concentrations do physically exist. Substituting the general solution into the boundary conditions with $n = 1/2$ and $\theta = \pm\pi$, we get $3A + C = 0$, and $B - D = 0$.

By defining the stress intensity factor $K_I$ for the crack, or $K_I = \frac{3A\sqrt{2\pi}}{4}$, which is for the opening (model I) of the crack, it is the stress limit at $\theta = 0$ so that

$$K_I = \lim_{r\to 0} \sigma_{\theta\theta}(r, \theta)\Big|_{\theta=0}. \tag{14.92}$$

Finally, the solution of stresses can be written as

$$\sigma_{rr} = \frac{K_I}{\sqrt{2\pi r}}(1 + \sin^2\frac{\theta}{2})\cos\frac{\theta}{2}, \tag{14.93}$$

$$\sigma_{\theta\theta} = \frac{K_I}{\sqrt{2\pi r}}\cos^3\frac{\theta}{2}, \tag{14.94}$$

$$\sigma_{r\theta} = \frac{K_I}{\sqrt{2\pi r}}\cos^2\frac{\theta}{2}\sin\frac{\theta}{2}. \tag{14.95}$$

Once we have the stress distribution, we can get the strains and displacements.

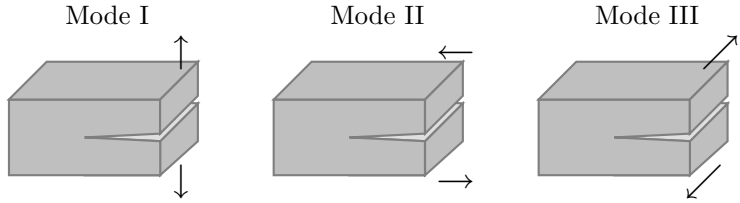



Figure 14.8: Fracture modes: Mode I (normal opening), mode II (in plane shear), and mode III (out of plane shear or tearing).

## 14.6 Fracture Mechanics

Any materials will have limited strength, and they can only carry finite stresses. Beyond a certain limit of stress, called failure stress, the material will fail, usually by fast fracture. The basic mechanism of fracture is by movement of dislocation (or flaws) in the material or by bond-breaking, and most likely a combination of these basic mechanisms.

At a crack tip, the stresses vary greatly with the distance $r$ and orientation $\theta$. Writing the above stress expressions in the previous section in terms of $\sigma_x$ and $\sigma_y$, and using $x = r\cos\theta$, $y = r\sin\theta$ as well as $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$, we finally have

$$\sigma_x = \frac{K_I}{\sqrt{2\pi r}}\cos\frac{\theta}{2}(1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2}), \tag{14.96}$$

$$\sigma_y = \frac{K_I}{\sqrt{2\pi r}}\cos\frac{\theta}{2}(1 + \sin\frac{\theta}{2}\sin\frac{3\theta}{2}), \tag{14.97}$$

and

$$\tau_{xy} = \frac{K_I}{\sqrt{2\pi r}}\cos\frac{\theta}{2}\cos\frac{3\theta}{2}\sin\frac{\theta}{2}. \tag{14.98}$$

The parameter $K_I$ is very important and is often called the stress intensity factor for mode I (normal opening) cracking as shown in Fig. 14.8. For other fracture modes, there are also corresponding stress intensity factors ($K_{II}$ and $K_{III}$).

From Eqs.(14.96) to (14.98), we know that the factor $1/\sqrt{2\pi r}$ approaches singularity as $r \to 0$. This means the stress $\sigma$ approaches infinity as $r \to 0$ at the crack tip. Combining Eqs.(14.92) with (14.94), we have

$$K_I = \sigma\sqrt{\pi a}, \tag{14.99}$$

where $a$ is the initial crack length. The expression for the factor $K_I$ implies that the overall intensity of the stress distribution (hence the name, stress intensity factor) around a crack is governed by the applied stress $\sigma$, the crack length $a$ and the specimen geometry. The geometry often introduces a geometrical factor or shape factor $Y$ in the stress intensity factor $K_I$

$$K_I = Y\sigma\sqrt{\pi a}. \tag{14.100}$$

For an edge crack at the boundary $Y = 1$. For an edge crack of length $2a$ in a large plane sheet, $Y = 1.12$. As all materials can withstand stress under a certain limit, if the stress exceeds a critical stress $\sigma_f$, called failure stress or strength, then fast fracture occurs when the initial crack starts to propagate and extend rapidly. Therefore, the crack tip can only withstand stresses up to a critical value of stress intensity, referred to as critical stress intensity factor $K_{IC}$. $K_{IC}$ is a material property and is a measure of material toughness so it is called fracture toughness. From the above equation, we have

$$K_{IC} = Y\sigma_f\sqrt{\pi a}. \tag{14.101}$$

For examples, we have $K_{IC} \approx 150$ MPa$\sqrt{\text{m}}$ for steels; $K_{IC} \approx 0.5$ MPa$\sqrt{\text{m}}$ for wood; and $K_{IC} \approx 0.7$ MPa$\sqrt{\text{m}}$ for glass.

There is an alternative approach to study the fast fracturing process, that is to use the energy balance approach first developed by A. Griffith in 1920. From Fig. 14.7, we know that the strain energy per unit volume for linear elastic media ($\sigma = E\varepsilon$) is given by

$$u = \frac{1}{2}\boldsymbol{\sigma} \cdot \boldsymbol{\varepsilon} = \frac{E\varepsilon^2}{2} = \frac{\sigma^2}{2E}. \tag{14.102}$$

Suppose that the region affected by crack is a triangular region with a width of $a$ and height $h = \beta a$. Thus, the total strain energy in the region is

$$U = -\frac{\sigma^2}{2E}\beta a^2, \tag{14.103}$$

where we use a negative sign to denote the fact that the strain energy can be released to create cracks. In order for the crack to propagate, energy is required to break the bond energy between atoms in the surface created by any advancing crack increment $\delta a$. Let $\gamma$ be the specific surface energy (surface energy per unit area) with a unit of J/m$^2$. The total surface energy is

$$\Gamma = 2\gamma a, \tag{14.104}$$

for a crack of length $a$. Here the factor 2 comes from the fact that two free surfaces have been created. The total energy in the system is

$$\Pi = U + \Gamma = 2\gamma a - \frac{\sigma^2}{2E}\beta a^2. \tag{14.105}$$

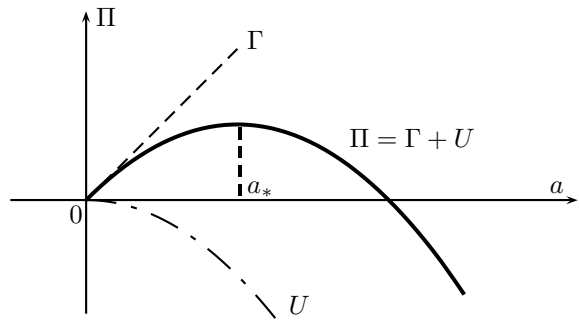



Figure 14.9: Energy in fast fracture in an elastic media.

We see that there are two competing terms (see Fig. 14.9), whose stationary conditions will define a critical length $a_*$ of initial cracks at the failure stress $\sigma_f$. We have

$$\frac{\partial \Pi}{\partial a} = 2\gamma - \frac{\sigma_f^2}{E}\beta a_* = 0, \tag{14.106}$$

or

$$a_* = \frac{2\gamma E}{\beta \sigma_f^2}, \qquad \sqrt{2E\gamma} = \sigma_f \sqrt{\beta a_*}. \tag{14.107}$$

If we compare these results with Eq.(14.99) with $a = a_*$ and $\sigma = \sigma_f$, then we have

$$\beta = \pi, \qquad K_{IC} = \sqrt{2\gamma E}. \tag{14.108}$$

In order to generalise the concept, it is more conventional to use the critical strain energy release rate or fracture energy $G_f$ (instead of $2\gamma$) to denote that the rate of strain energy release is sufficient to create all free surfaces so as to sustain the crack propagation. This means that $G_f = 2\gamma$ or the rate of strain energy release is just balanced or completely converted to surface energy. Crack propagation occurs only if the energy needed to create a unit surface is sufficiently supplied (or balanced) by the release rate of strain energy, and this is known as Griffith's criterion. Therefore, we have

$$\sigma_f = \sqrt{\frac{EG_f}{\pi a}}, \qquad K_{IC} = \sqrt{G_f E}, \tag{14.109}$$

which is an important relationship because it links critical strain energy release rate, failure stress and the length of cracks (or flaws in the

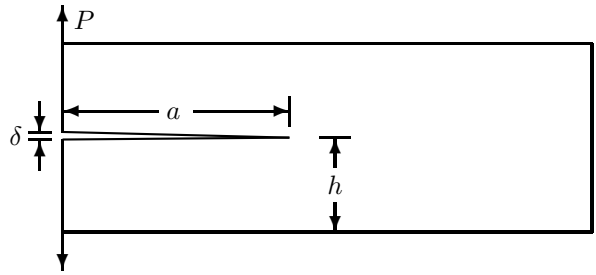



Figure 14.10: Normal opening by splitting a double cantilever beam with a thickness of $b$ and fracture energy.

system). This relationship is only valid for plane stress. For plane strain, we have

$$K_{IC}^2 = EG_f(1-\nu^2). \tag{14.110}$$

For $\nu \approx 0.1 \sim 0.3$, $(1-\nu^2) \approx 0.91 \sim 0.99$, thus $K_{IC}^2 \approx EG_f$ can be used for both cases.

From the data of $K_{IC}$ and $E$, we can calculate $G_f$ or vice versa. Using the values given earlier, we have $E \approx 210$ GPa, $K_{IC} \approx 150$ MPa$\sqrt{\text{m}}$ and 107 kJ m$^{-2}$ for steels; $E \approx 2$ GPa, $K_{IC} \approx 0.5$ MPa$\sqrt{\text{m}}$ and 0.12 kJ m$^{-2}$ for wood; and $E \approx 70$ GPa, $K_{IC} \approx 0.7$ MPa$\sqrt{\text{m}}$, and 0.007 kJ m$^{-2}$ for glass. This is why glass is so brittle and can fracture easily, and why steel is very stiff and resistant to fracturing.

Near the tip of a crack, there is a small region where crack propagation is not governed by fast fracture in the elastic media, but plasticity comes to play an important role. This small region is called the process zone with a size $r_p$ which can be estimated as follows. Though the bond rupture is important in fast fracture process, however, the largest part of the fracture energy is actually associated with plastic flow near the crack tip, where the stress is so high that material starts to yield once the stress reaches the yield stress $\sigma_Y$.

When the stress is near the yield stress $\sigma_Y$, plastic flow comes into action and plays an important role in resisting crack propagation. From $K_I = \sigma\sqrt{\pi a}$, we know that $\sigma = \sigma_Y$ and $K_I = K_{IC}$, and we have

$$\sigma = \sigma_Y = \frac{K_{IC}}{\sqrt{\pi r_p}}, \qquad \text{or} \qquad r_p \approx \frac{K_{IC}^2}{\pi\sigma_Y^2}. \tag{14.111}$$

Strictly speaking, Eqs.(14.96) to (14.98) are only valid in certain regions, typically when $r_p < r < a/4$. If $r$ is too large or too small, the stress is no longer correct.

Now a simple question is how to measure fracture energy for a given material. There are many ways to achieve this, and the simplest way

is to use the splitting of a double cantilever beam shown in Fig. 14.10.

---

**Example 14.3:** For the double cantilever beam, the total strain energy $U$ is

$$U = \frac{1}{2}P\delta = \frac{1}{2}CP^2,$$

where $\delta$ is the total deformation. Here we have also used the compliance $C = \delta/P$ which is the ratio of deformation to an applied load $P$. The strain energy release rate $G_f$ is equivalent to

$$G_f = \frac{\partial U}{\partial s} = \frac{P^2}{2b}\frac{\partial C}{\partial a}\Big|_{a=a_*},$$

where $s$ is the surface area and $ds = b\ da$ with $b$ being the thickness of the beam. As the double cantilever system is symmetric, the end deflection of of a cantilever is $\delta/2$ which is given by Eq.(14.78), and we have

$$\frac{\delta}{2} = \frac{Pa^3}{3EI} = \frac{4Pa^3}{Ebh^3},$$

where we have used $I = bh^3/12$. As $C = \delta/P = 2a^3/(3EI)$ and $\frac{\partial C}{\partial a} = 24a^2/Ebh^3$, the critical energy release rate $G_f$ at the critical load $P_*$ for the crack to just propagate, is thus given by

$$G_f = \frac{P_*}{2b}\frac{2a^2}{EI} = \frac{12P_*a^2}{b^2h^3E}.$$

---

The effect of grain size on the yield strength $\sigma_y$ is often called the Hall-Petch equation

$$\sigma_y = \sigma_0 + \frac{A}{\sqrt{d}}, \tag{14.112}$$

where $\sigma_0$ is the lattice resistance, and $d$ is the grain size or crystallite size. $A$ is a constant which is also called the dislocation locking coefficient in materials sciences. This relationship is experimentally validated from the millimetre scale to the submicrometre scale. It is expected that it is also valid for the larger grain size. Consequently, fine-grain materials are stronger than coarse-grained materials.

For ductile materials, ahead of the crack tip there is a substantial amount of plastic flow, and this plasticity dissipates a lot of energy as compared with the limited damage before bonds break in brittle elastic materials. This higher energy dissipation requires a higher strain energy release rate. Subsequently, ductile materials usually have higher fracture energy. However, the concept of stress intensity factor is not valid for ductile materials, and thus more advanced techniques such as J-integral should be used.

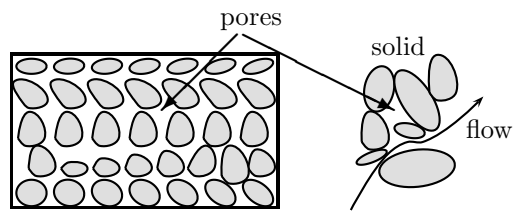



Figure 14.11: Irregular solid grains and pores in porous media.

## 14.7 Biot's Theory

Poroelasticity is widely used in earth sciences, from geomechanics to petroleum engineering, from hydrology to environmental fluid dynamics. Linear poroelasticity can be considered as the extension of the theory of elasticity to theory of porous media. Biot's theory is as important to poroelasticity as Hooke's law to linear elasticity.

### 14.7.1 Biot's Poroelasticity

In linear elasticity, there is a relationship between strain and stress, and thus there exist pairs of two basic elastic parameters such as Young's modulus $E$ and Poisson's ratio $\nu$ or the Lamé constant $\mu = G$ and $\lambda$ or the bulk modulus and one other parameter. In a porous medium, the situation is more complicated as it consists of irregular solid grains (solid matrix) and pores filled with fluid (see Fig. 14.11). The porous system will deform under stress and the volume of pores will change, and the pore pressure (the pressure of fluid in the pores) will also change. In addition, as the volume of pores changes, the fluid content will subsequently change. These processes are coupled and thus more difficult to model. In linear poroelasticity, there are two more quantities: pore pressure $p$ and the increment of fluid content $\theta = \delta m_f/\rho_f$ per unit volume where $\delta m_f$ is the change in fluid mass content and $\rho_f$ is the density of the fluid at reference state. Similar to the stress-strain relationship in linear elasticity, we have to relate $\theta$ to $p$ in some way. Therefore, extra parameters are needed.

In order to introduce these parameters, we now start with the simplest isotropic fluid-filled porous medium with isotropic applied stress field $\sigma$. The volume strain or dilation $\Delta$ is also related to the fluid content $\xi$. In this simplest case, Biot's theory provides a relationship

concerning these quantities

$$\begin{pmatrix} \Delta \\ \theta \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \sigma \\ p \end{pmatrix}, \tag{14.113}$$

where $A_{ij}$ is the coefficient matrix. From the reciprocity in strain energy where the final state of the system should be independent of the order of applying forces, which leads to the Maxwell's reciprocal theorem, the matrix $A_{ij}$ should be symmetric. That is $A_{ij} = A_{ji}$. In our case, $A_{12} = A_{21}$. The coefficients can be defined by fixing one independent variable and varying the other, and they are given by

$$A_{11} \equiv \frac{\partial \Delta}{\partial \sigma}\Big|_p = \frac{1}{K}, \qquad A_{12} \equiv \frac{\partial \Delta}{\partial p}\Big|_\sigma = A_{21} \equiv \frac{\partial \theta}{\partial \sigma}\Big|_p = \frac{1}{H}, \tag{14.114}$$

and

$$A_{22} \equiv \frac{\partial \theta}{\partial p}\Big|_\sigma = \frac{1}{R}, \tag{14.115}$$

where the notation $|_p$ emphasises the fact that $p$ is held constant. $K$ is the drained bulk modulus, and the coefficient $1/K$ is the drained compressibility of the material measured under drained condition at constant pressure. $S_\sigma = 1/R$ is the unconstrained specific storage coefficient at constant stress, and $1/H$ is called the compressibility under constant stress or the poroelastic expansion coefficient. These notations might be unfamiliar; however, they are essentially the same notations used by Biot in 1941, and such notations are widely used in the literature of poroelasticity.

Now equation (14.113) becomes

$$\begin{pmatrix} \Delta \\ \theta \end{pmatrix} = \begin{pmatrix} 1/K & 1/H \\ 1/H & 1/R \end{pmatrix} \begin{pmatrix} \sigma \\ p \end{pmatrix}. \tag{14.116}$$

Other parameters that are widely used but can be derived from the above three parameters are the Skempton's coefficient $B$ and the constrained specific storage coefficient $S_\Delta$ which is the specific storage coefficient at constant strain. They are defined by

$$B \equiv -\frac{\partial p}{\partial \sigma}\Big|_\theta = \frac{R}{H}, \tag{14.117}$$

$$S_\Delta \equiv \frac{\partial \theta}{\partial p}\Big|_\Delta = \frac{1}{M} = S_\sigma - \frac{K}{H^2} = S_\sigma - \frac{\alpha^2}{K}, \tag{14.118}$$

where $\alpha = K/H$ is called the Biot-Willis coefficient which is the ratio of volume of fluid to the change in bulk volume at constant pore pressure. The standard Poisson's ratio $\nu$ is defined under the drained condition in

poroelasticity at constant pressure and stress $\nu \equiv -\varepsilon_{\text{transverse}}/\varepsilon_{\text{axial}}$. There is a Poisson's ratio defined under undrained condition, called undrained Poisson's ratio $\nu_u$, which is given by

$$\nu_u \equiv -\frac{\varepsilon_{\text{transverse}}}{\varepsilon_{\text{axial}}}\bigg|_\theta = \frac{3\nu + \alpha B(1-2\nu)}{3 - \alpha B(1-2\nu)}. \tag{14.119}$$

### 14.7.2 Effective Stress

Terzaghi was the first to study the behaviour of soil and formulate his consolidation theory. From a series of experimental studies, he concluded that the solid grains are almost incompressible and pore pressure $p$ has almost no effect in the volume of the soil. What controls the changes in the volume and strength of a soil is Terzaghi's effective stress $\sigma_e$ defined as the difference of the total stress $\sigma_{ij}$ and the pore pressure, that is $\sigma_e = \sigma - p$. As the conventions of the stress and pressure are different in engineering, fluid dynamics and soil mechanics, we use the following conventions: stress is negative for compression and positive for tension, while pressure $p$ is positive if it is greater than ambient pressure in compression. Then, we have effective stress

$$\sigma'_{ij} \equiv \sigma_{eij} = \sigma_{ij} + \alpha p \delta_{ij}, \tag{14.120}$$

where $\alpha$ is $O(1)$ coefficient. In Terzaghi's definition, we have $\alpha = 1$. The notations $\sigma'$ and $\sigma_e$ are interchangeable. $\sigma_e$ emphasises the fact that it is effective stress, while $\sigma'$ is more convenient when using subscript or index such as $\sigma'_{ij}$. Skempton in 1960 suggested that value $\alpha = 1 - a$ where $0 < a \ll 1$ is affected by the grain-to-grain interfacial contact area, though it might not be always the case for more fully compacted rocklike porous media. In the simplest one-dimensional case, we have the effective mean stress $\sigma'_{kk}/3$

$$\frac{\sigma'_{kk}}{3} = \frac{\sigma_{kk}}{3} + \alpha p. \tag{14.121}$$

We can also write them in terms of effective pressure $p_e = -\sigma'_{kk}/3$

$$p_e = P - \alpha p, \tag{14.122}$$

where $P = -\sigma_{kk}/3$ is the total pressure.

## 14.8 Linear Poroelasticity

### 14.8.1 Poroelasticity

As the effective stress is the most important concept in poroelasticity, and since fluid cannot support shear, the constitutive laws should be

formulated in terms of effective stress for fluid-saturated porous media. In fact, effective stress $\sigma_e$ or $\sigma'_{ij} = \sigma_{ij} + \alpha p \delta_{ij}$ is to poroelasticity as total stress $\sigma$ to linear elasticity in a solid. For groundwater flow problems, we can assume that the solid matrix grains are almost incompressible. With these assumptions, we can simplify the formulation for poroelastic flow problems, and we have a special form of Biot's constitutive laws.

Since the Terzaghi effective stress is essentially the grain-to-grain stress which acts only in the solid matrix, and the solid matrix itself can be described using linear elasticity, we can now rewrite the linear elasticity in terms of effective stress. If we replace the total stress $\sigma_{ij}$ in the linear elasticity $\sigma_{ij} = 2G\varepsilon_{ij} + \lambda \varepsilon_{kk} \delta_{ij}$ by the effective stress $\sigma'_{ij}$, we have

$$\sigma'_{ij} = 2G\varepsilon_{ij} + \lambda \varepsilon_{kk} \delta_{ij}, \tag{14.123}$$

or in terms of total stress and pressure

$$\sigma_{ij} = 2G\varepsilon_{ij} + \lambda \varepsilon_{kk} \delta_{ij} - \alpha p \delta_{ij}, = 2G\varepsilon_{ij} + \frac{2G\nu}{1-2\nu} \varepsilon_{kk} \delta_{ij} - \alpha p \delta_{ij}. \tag{14.124}$$

This equation is equivalent to six equations for six independent components of the stress and strain. From Eq.(14.116) and $\sigma = \sigma_{kk}/3$, the extra constitutive equation becomes

$$\theta = \frac{1}{3H} \sigma_{kk} + \frac{p}{R} = \frac{\alpha}{3K} \sigma_{kk} + \frac{\alpha}{KB} p. \tag{14.125}$$

Both Eqs.(14.124) and (14.125) form the basic equations in Biot's three-dimensional poroelasticity.

If we want to invert Eq.(14.124), we have to obtain $\sigma_{kk}$ by contraction using $k = i = j$ and $\delta_{kk} = 3$, we have

$$\sigma_{kk} = 2G\varepsilon_{kk} + \frac{2G\nu}{1-2\nu} \varepsilon_{kk} \delta_{kk} - \alpha p \delta_{kk}, \tag{14.126}$$

or

$$\sigma_{kk} = \frac{2G(1+\nu)}{1-2\nu} \varepsilon_{kk} - 3\alpha p = 3K\varepsilon_{kk} - 3\alpha p, \tag{14.127}$$

where we have used $G = E/2(1+\nu)$ and $K = E/3(1-2\nu)$. Now we have

$$\varepsilon_{kk} = \frac{1}{3K} \sigma_{kk} + \frac{\alpha}{K} p. \tag{14.128}$$

---

**Example 14.4:** The above equation can be written in terms of volume stress and effective pressure. We have

$$\varepsilon_{kk} = \frac{1}{K} [\frac{\sigma_{kk}}{3} + \alpha p],$$

Since $-\sigma_{kk}/3 = P$ is the total pressure and $P - \alpha p$ is the effective pressure $p_e$ using Eq.(14.122), we have

$$\varepsilon_{kk} = -\frac{1}{K} p_e,$$

or

$$p_e = -K\varepsilon_{kk} = -K\Delta = -K\nabla \cdot \boldsymbol{u}.$$

This relationship highlights the fact that the volume strain is solely determined by the effective pressure in the porous medium.

---

In later discussion of consolidation, the relationship

$$p_e = -K\nabla \cdot \boldsymbol{u}, \tag{14.129}$$

can be considered as the constitutive relationship of the porous medium. In viscous creep and rock deformation, this relationship can be modified to include viscous creep laws.

Substituting the expression for $\varepsilon_{kk}$ back into Eq.(14.124) and using $2G\nu/3K(1-2\nu) = \nu/(1+\nu)$, we have

$$\sigma_{ij} = 2G\varepsilon_{ij} + \frac{\nu}{1+\nu}\sigma_{kk}\delta_{ij} + [\frac{2G\nu}{K(1-\nu)} - 1]\alpha p\delta_{ij}. \tag{14.130}$$

Dividing both sides of the above equation by $2G$ and noting that

$$\frac{\nu}{K(1-2\nu)} - \frac{1}{2G} = -\frac{1}{3K}, \tag{14.131}$$

we have

$$\frac{1}{2G}\sigma_{ij} = \varepsilon_{ij} + \frac{\nu}{2G(1+\nu)}\sigma_{kk}\delta_{ij} - \frac{1}{3K}\alpha p\delta_{ij}. \tag{14.132}$$

Rearranging this equation, we have

$$\begin{aligned}\varepsilon_{ij} &= \frac{1}{2G}\sigma_{ij} - \frac{\nu}{2G(1+\nu)}\sigma_{kk}\delta_{ij} + \frac{\alpha}{3K}p\delta_{ij} \\ &= \frac{1+\nu}{E}\sigma_{ij} - \frac{\nu}{E}\sigma_{kk}\delta_{ij} + \frac{\alpha}{3K}p\delta_{ij}.\end{aligned} \tag{14.133}$$

Using $\sigma_{kk} = 3K\varepsilon_{kk} - 3\alpha p$ from Eq.(14.127), Eq.(14.125) becomes

$$\theta = \frac{3K}{3H}\varepsilon_{kk} + (\frac{1}{R} - \frac{\alpha}{H})p. \tag{14.134}$$

Using $\alpha = K/H$ and $1/R - \alpha/H = \alpha/K_u B$, we have

$$\theta = \alpha\varepsilon_{kk} + \frac{\alpha}{K_u B}p, \tag{14.135}$$

where

$$K_u = \frac{K}{1 - \alpha B}, \tag{14.136}$$

is the undrained bulk modulus under the constraint that $\Delta$ is constant.

### 14.8.2 Equation of Motion

Using the poroelasticity constitutive relationship (14.124) and following the same procedure of obtaining the equation of motion (14.52) in linear elasticity $\rho\ddot{\boldsymbol{u}} = \frac{G}{(1-2\nu)}\nabla(\nabla \cdot \boldsymbol{u}) + G\nabla^2\boldsymbol{u} + \boldsymbol{b}$, we will arrive at the following equation of motion for poroelasticity

$$\rho\frac{\partial^2 \boldsymbol{u}}{\partial t^2} = \frac{G}{(1-2\nu)}\nabla(\nabla \cdot \boldsymbol{u}) + G\nabla^2\boldsymbol{u} + \alpha\nabla p + \boldsymbol{b}, \tag{14.137}$$

where the only extra term is the $\alpha\nabla p$. In the case of $\boldsymbol{b} = 0$ the pressure $p$ usually increases with depth, we have $p = \rho_f g h$ where $h$ is the change of fluid (water) head from initial head, often called the head in hydrology. $\rho_f$ is the density of the fluid and $g$ is the acceleration due to gravity. Therefore, we have

$$\rho\frac{\partial^2 \boldsymbol{u}}{\partial t^2} = \frac{G}{(1-2\nu)}\nabla(\nabla \cdot \boldsymbol{u}) + G\nabla^2\boldsymbol{u} + \alpha\rho_f g\nabla h. \tag{14.138}$$

Now we have an extra variable $p$ (or $h$) to be determined, and thus an extra equation is needed. This extra equation forms the basis of flow in porous media and will be discussed in detail in the next chapter.

## References

Biot M. A., General theory of three-dimensional consolidation, *J. Appl. Phys.*, **12**, 155-164 (1941).

Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

Greenkorn R. A., *Flow phenomena in porous media*, Marcel Dekker Inc., (1983).

Parton V. Z., *Fracture Mechanics: From Theory to Practice*, Gorden & Breach Sci. Publishers, (1992).

Smith G. N., *Elements of Soil Mechanics for Civil and Mining Engineers*, 5th Edition, Granada Publishing, (1982).

Wang H. F., *Theory of Linear Poroelasticity: with applications to geomechanics and hydrogeology*, Princeton University Press, (2000).

Terzaghi K., *Theoretical Soil Mechanics*, John Wiley, New York, (1943).

# Chapter 15

# Flow in Porous Media

The theory of linear poroelasticity discussed in the last chapter can be used to study various phenomena in porous media. As applications, we will study groundwater flow, the consolidation of soil, pollutant transport and viscous creep.

## 15.1 Groundwater Flow

For groundwater flow, we can often assume that the solid matrix is incompressible and its deformation is small or even negligible. This will simplify the governing equations tremendously.

### 15.1.1 Porosity

The void rate $e$ is defined by

$$e = \frac{\text{volume of voids } (V_v)}{\text{volume of solids } (V_s)}. \tag{15.1}$$

Since the porosity $\phi$ is the ratio of the volume of voids to the total volume $(V)$, thus we have

$$\phi = \frac{V_v}{V} = \frac{V_v}{V_s + V_v} = \frac{e}{1+e}, \tag{15.2}$$

or conversely $e = \frac{\phi}{1-\phi}$.

### 15.1.2 Darcy's Law

Darcy's law describes the flow through a porous medium under a pressure gradient. It is as important to porous media as Fick's law is

to diffusion and heat conduction. Let $\boldsymbol{q}$ be the flux of the fluid flow through the medium. Darcy's law states that the flux $\boldsymbol{q}$ is proportional to the pressure gradient

$$\boldsymbol{q} = -\frac{k}{\mu}\nabla p, \tag{15.3}$$

where $k$ is the permeability of the porous system depending on microstructure of the solid rocks, and $\mu$ is the viscosity of the fluid.

The permeability $k$ is usually a function of porosity $\phi$, but it often assumes that porosity $\phi$ does not change with time $t$ in groundwater flow. Typical values of permeability are $10^{-9} \sim 10^{-11}$ m$^2$(or 1000 darcy to 10 darcy) for sands (where 1 darcy = $10^{-12}$m$^2$), $10^{-12} \sim 10^{-14}$ m$^2$ for sandstone, $10^{-14} \sim 10^{-16}$ m$^2$ for limestone, and $10^{-18}$ m$^2$ or even smaller for shales. However, the viscosity $\mu$ of fluid at room temperature is about $10^{-3}$ N s m$^{-2}$ and it varies with temperature.

Since gravity always appears in the groundwater flow, the pressure should be replaced by $p \to p + \rho_f g z$ where $\rho_f$ is the density of the fluid, $g$ is the acceleration due to gravity, and $z$ is pointing upwards. Therefore, Darcy's law can be written the generic form

$$\boldsymbol{q} = -\frac{k}{\mu}\nabla(p + \rho_f g z) = -\frac{k}{\mu}\nabla p - \kappa\vec{\boldsymbol{k}}, \tag{15.4}$$

where $\vec{\boldsymbol{k}}$ is a unit vector pointing upwards in the $z$-axis direction. The ratio of $k$ to $\mu$ defines a hydraulic conductivity $\kappa$

$$\kappa \equiv \frac{\rho_f g k}{\mu}. \tag{15.5}$$

In groundwater flow, it is usual to link the pressure to a head $h$

$$p + \rho_f g z = \rho_f g h, \tag{15.6}$$

or

$$h = z + \frac{p}{\rho_f g}. \tag{15.7}$$

Thus, Darcy's law can be written in terms of the head $h$ as

$$\boldsymbol{q} = \kappa\nabla h. \tag{15.8}$$

---

**Example 15.1:** Darcy's flow velocity $\boldsymbol{q}$ is usually very small. In fact, on the label of some popular mineral water bottles, it is claimed that the mineral water has filtered through a 1500m layer of ancient limestone for about 5000 years. Let us see if this is true.

We know that the permeability for limestone is about $k = 10^{-14} \sim 10^{-16}$ m$^2$, so we take a moderate value $k = 10^{-15}$ m$^2$ and the viscosity of

water $\mu = 10^{-3}$ N s m$^{-2}$. For the pressure gradient, we use the hydrostatic gradient $\nabla p \approx \rho g \approx 10^4$ N m$^{-3}$. Thus, the Darcy's flow velocity is about the order

$$\boldsymbol{q} \approx \frac{k}{\mu}\nabla p \approx \frac{10^{-15}}{10^{-3}} \times 10^4 \approx 10^{-8}\text{m/s}.$$

This seems too small, however, as there are about $365 \times 24 \times 3600 \approx 3.15 \times 10^7$ seconds in a year. This velocity is about

$$\boldsymbol{q} \approx 10^{-8} \times 3.15 \times 10^7 \approx 0.315\text{m/year}.$$

Thus, for the water to filter through a layer of limestone with a thickness of 1500m, it would take

$$t \approx \frac{1500}{0.315} \approx 4700\text{years}.$$

So the value of 5000 years is about right. This means that the mineral water (of this particular brand) is at least 5000 years old!

### 15.1.3 Flow Equations

The conservation of fluid mass or continuity equation can be written as

$$\frac{\partial \theta}{\partial t} + \nabla \cdot \boldsymbol{q} = Q, \tag{15.9}$$

where $Q$ is the fluid source which could vary with spatial coordinates and even time $t$. Substituting Darcy's law (15.3) into Eq.(15.9), we have

$$\frac{\partial \theta}{\partial t} = \nabla.(\frac{k}{\mu}\nabla p) + Q. \tag{15.10}$$

In the case of $k/\mu$ is a constant, we have

$$\frac{\partial \theta}{\partial t} = \frac{k}{\mu}\nabla^2 p + Q. \tag{15.11}$$

Using the constitutive equation (14.125), we get

$$\frac{\alpha}{3K}\frac{\partial \sigma_{kk}}{\partial t} + \frac{\alpha}{KB}\frac{\partial p}{\partial t} = \frac{k}{\mu}\nabla^2 p + Q. \tag{15.12}$$

As $\alpha/KB = 1/R = S_\sigma$ is the specific storage at constant stress, we can rewrite the above equation as

$$S_\sigma \frac{\partial p}{\partial t} = \frac{k}{\mu}\nabla^2 p - \frac{S_\sigma B}{3}\frac{\partial \sigma_{kk}}{\partial t} + Q, \tag{15.13}$$

which is the general equation for three-dimensional groundwater flow. This is an inhomogeneous diffusion equation for pore pressure. We can also express this equation in terms of $h$ by using Eq.(15.7) and assuming $z$ and $t$ are independent, and we have

$$S\frac{\partial h}{\partial t} = \kappa\nabla^2 h - \gamma\frac{\partial \sigma_{kk}}{\partial t} + Q, \tag{15.14}$$

where $S = S_\sigma \rho_f g$ and $\gamma = S_\sigma B/3$. Under the condition of constant stress $\sigma_{zz}$ so that $\frac{\partial \sigma_{kk}}{\partial t} = 0$, we have a simpler equation

$$S\frac{\partial h}{\partial t} = \kappa\nabla^2 h + Q. \tag{15.15}$$

Sometimes it is more convenient to write this governing equation in the cylindrical polar coordinates $(r, \varphi, z)$; we have

$$S\frac{\partial h}{\partial t} = \kappa[\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial h}{\partial r}) + \frac{1}{r^2}\frac{\partial^2 h}{\partial \varphi^2} + \frac{\partial^2 h}{\partial z^2}] + Q. \tag{15.16}$$

In the simplified case when $Q = 0$ and $h$ has rotational symmetry, then $h$ does not depend on $\varphi$ and $z$, and we have

$$S\frac{\partial h}{\partial t} = \kappa\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial h}{\partial r}). \tag{15.17}$$

---

**Example 15.2:** In order to measure the hydraulic conductivity $\kappa$, the pumping out test is often used. In a well with a radius of $r = a$, a constant pumping rate $-Q_p$ (volume per unit time) where we use the negative to denote the flow is for pumping out (see Fig. 15.1). In the cylindrical coordinates, the rate of water flow is

$$-Q_p = \int_0^h \int_0^{2\pi} J d\varphi dz,$$

where the flux $J$ is

$$J = -\kappa r\frac{\partial h}{\partial r}.$$

This leads to

$$Q_p = 2\pi\kappa h r\frac{\partial h}{\partial r}.$$

Let $h_0$ be the water level in the well, which means $h = h_0$ at $r = a$. If we can measure the water level $h_*$ by using a borehole at any location $r = R > a$, we can determine the average hydraulic conductivity of the layer (between initial water head level to $h_0$).

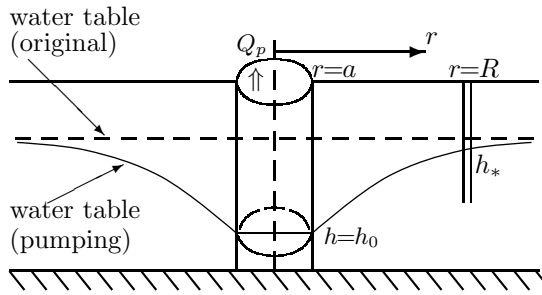



Figure 15.1: Measuring hydraulic conductivity by the pumping out test.

Integrating both sides of the above equation, we have

$$2\pi\kappa \int_{h_0}^{h_*} h dh = \int_a^R Q_p \frac{dr}{r},$$

which gives

$$\pi\kappa(h_*^2 - h_0^2) = Q_p \ln \frac{R}{a}.$$

Thus, we have an expression for $\kappa$

$$\kappa = \frac{Q_p \ln(R/a)}{\pi(h_*^2 - h_0^2)}.$$

Obviously, if there are more observation points for the water head at various locations, the error of $\kappa$ can be minimised.

---

There are other methods for measuring hydraulic conductivity such as the rock sampling and pumping in test.

In the simplest 1-D case ($z$-axis only) where $Q = 0$ and the stress $\sigma_{zz}$ is constant, we simply have

$$\frac{\partial h}{\partial t} = \frac{\kappa}{S} \frac{\partial^2 h}{\partial z^2}, \tag{15.18}$$

which is the 1-D equation for groundwater flow under constant stress. We can see that this is essentially the 1-D diffusion equation or heat conduction equation we met in earlier chapters. Thus, the solution techniques for given appropriate boundary conditions will be the same as those for the diffusion equation. Let us look at an example.

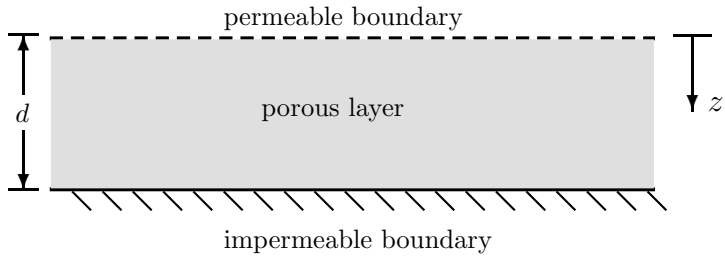



Figure 15.2: 1-D groundwater flow in a permeable layer.

---

**Example 15.3:** For the 1-D ground-water flow in a permeable layer shown in Fig. 15.2, the top boundary is permeable at $z = 0$ so that $h = h_0$ (fixed water head), and the bottom boundary is an impermeable base at $z = d$ with $\frac{\partial h}{\partial z} = 0$ (zero flux boundary). The initial condition is $h(z, t = 0) = 0$. The present problem becomes

$$\frac{\partial h}{\partial t} = \frac{\kappa}{S}\frac{\partial^2 h}{\partial z^2},$$

with

$$h(z = 0, t) = h_0, \;\; \frac{\partial h}{\partial z}\Big|_{z=d} = 0, \;\; h(z, t = 0) = 0.$$

This problem can be converted into the same problem (5.29) with boundary conditions (5.30) discussed earlier. Now let us set $h = h_0 + u$ and use $\frac{\partial h}{\partial t} = \frac{\partial u}{\partial t}$ and $\frac{\partial h}{\partial z} = \frac{\partial u}{\partial z}$, we then have

$$\frac{\partial u}{\partial t} = \frac{\kappa}{S}\frac{\partial^2 u}{\partial z^2},$$

with boundary conditions

$$u(z = 0, t) = 0, \;\; \frac{\partial u}{\partial z}\Big|_{z=d} = 0, \;\; u(z, t = 0) = -h_0.$$

This is essentially the same problem as (5.29) if we transform $z \to x$, $d \to L$, $\kappa/S \to \kappa$, and $\psi = -h_0$. Using solution (5.43), we have

$$u = -\frac{4h_0}{\pi}\sum_{n=0}^{\infty}\frac{1}{(2n+1)}\sin\frac{(2n+1)\pi z}{2d}e^{-(2n+1)^2\pi^2\kappa t/(4Sd^2)}].$$

Finally, the solution to the original problem becomes

$$h(z,t) = u + h_0 = h_0[1 - \frac{4}{\pi}\sum_{n=0}^{\infty}\frac{1}{2n+1}\sin\frac{(2n+1)\pi z}{2d}e^{-\frac{(2n+1)^2\pi^2\kappa t}{(4Sd^2)}}].$$

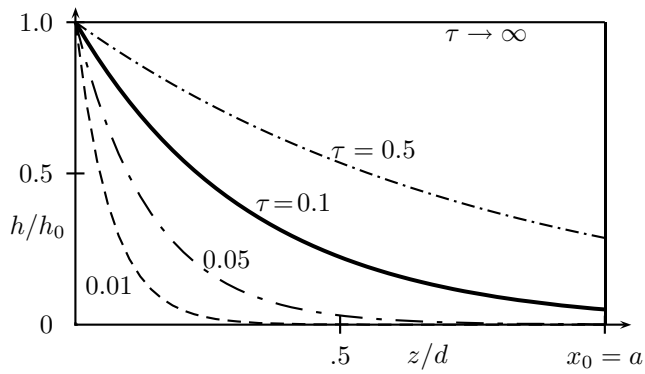



Figure 15.3: Evolving water heads for different dimensionless time $\tau = \kappa t/Sd$ in 1-D groundwater flow.

These results are shown in Fig. 15.3 where $\tau = \kappa t/Sd$ is the dimensionless time.

## 15.2 Pollutant Transport

An important issue in hydrology and environmental science is to understand how any potential pollutant will be transported in porous media. Let $c$ be the concentration of the pollutant or solute in the fluid, and conservation of mass for the solute leads to the following transport equation

$$\frac{\partial(\phi c)}{\partial t} + \nabla \cdot (c\boldsymbol{q}) = \nabla \cdot [D(\phi)\nabla c] + r, \tag{15.19}$$

where $r$ is the pollutant source, and $D(\phi)$ is the diffusion coefficient which is usually a function of porosity $\phi$. The diffusion coefficient depends on the tortuosity $\tau_D$ which is a measure of the zigzag path in porous media compared with the straight path in the same domain. In the simple case, we have $D \propto 1/\tau_d^2$ because the diffusion path becomes tortuous as affected by the porosity $\phi$, and thus the effective diffusion coefficient is a function of $\phi$. Generally speaking, $D(\phi) = D_0(\phi/\phi_0)^n$ where $n$ a constant and $D_0$ is the diffusion coefficient at the initial or reference porosity $\phi_0$. As before, $\boldsymbol{q}$ is the Darcy's flow velocity which can be written in terms of porosity $\phi$ and relative velocity $\boldsymbol{v}$ of fluid to the solid. That is $\boldsymbol{q} = \phi\boldsymbol{v}$, and thus $\boldsymbol{v}$ can be viewed as the convection

velocity for pollutant transport. Now we have

$$\frac{\partial(\phi c)}{\partial t} + \nabla \cdot (c\phi \boldsymbol{v}) = \nabla \cdot [D(\phi)\nabla c] + r. \tag{15.20}$$

Since $D_0 \approx 10^{-8} \sim 10^{-10}$ m$^2$ s$^{-1}$ (in rocks) is usually small, the pollutant transport process is primarily dominated by advection with the Darcy's flow which is usually faster than $10^{-8}$ m/s.

The governing equation for pollutant transport can also be used to study the bacterial transport in saturated/unsaturated soils and rocks. The transport of micro-organisms in porous media is crucially important for water and waste water management. The fundamental mobility of bacteria is based on the following simplified mechanism: the bacteria tend to deposit and attach to the surface of the solid grains, and they can also be released from the solid surface and be transported via the advection fluid flow in the porous medium. For simplicity, we model the deposition and release of the bacteria particles to or from the solid as the first-order kinetic process, while modelling the transport in fluid as the reaction-diffusion process. For a given species of bacteria, let $c$ and $s$ be the bacterial concentrations in the fluid and the solid surface, respectively. The governing equations can be written as

$$\frac{\partial(\phi c)}{\partial t} = D\nabla^2(\phi c) - \nabla(\phi c \boldsymbol{v}) - (r_0 + r_1)\phi c + \frac{r_2\phi}{A}s + r\phi c(c_0 - c), \tag{15.21}$$

$$\frac{\partial s}{\partial t} = D_s \nabla s + r_1 A c - r_2 s, \tag{15.22}$$

where $r$ is the growth rate of the bacteria species, and $rc(c_0 - c)$ is the typical logistic growth function. $c_0$ is the attainable maximum concentration in water as limited by food supply. $D_s$ is the diffusion coefficient for solid phase which is very small and can thus be taken to be zero ($D_s = 0$). $r_0$ is the deposit rate at the water-air surface, $r_2$ is the bacteria release rate. $r_1$ is the filtration factor including the absorption of bacteria onto the solid surfaces. The coefficient $A$ depends on the porosity $\phi$, degree of saturation $\Theta$, specific surface area $A_s$, and the bulk density $\rho$ of the porous medium. That is

$$A = \frac{\phi\Theta}{A_s\rho}. \tag{15.23}$$

The rate $r_0$ and $r_1$ also depend on the concentration and air content $(1 - \beta)$ so that

$$r_0 = r_a(1 - \beta), \qquad r_1 = r_s\Theta(1 - \frac{s}{s_{max}}), \tag{15.24}$$

where $r_a$ is the deposition constant for the bacteria at the air-water interface, while $r_s$ is the deposition constant on the solid surface. $s_{max}$ is the maximum attainable bacterial concentration on the solid surface due to competition for food.

For simplicity without losing generality, we can assume that the flow velocity $\boldsymbol{v}$ is a known constant vector. We also assume that both $\phi$ and $A$ are constant. With these assumptions, the 1-D governing equations become

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial z^2} - v\frac{\partial c}{\partial z}$$

$$-[r_a(1-\beta) + r_s\Theta(1-\frac{s}{s_{max}})]c + \frac{r_2}{A}s + rc(c_0 - c), \tag{15.25}$$

$$\frac{\partial s}{\partial t} = r_s\Theta(1-\frac{s}{s_{max}})Ac - r_2 s. \tag{15.26}$$

These two equations are coupled and form a system of nonlinear reaction-diffusion equations. The solution of the nonlinear system is usually difficult, however, let us discuss a simplified case as an example.

---

**Example 15.4:** Let us study the bacteria transport in a simple 1-D case. For the case of solid absorption limit, the deposit process is dominant and the release rate is essentially zero. Thus, we can assume $r_2 = 0$ and the air content is sufficient so that $\beta \to 0$. Let $c_0$ be the attainable maximum concentration of bacteria in water. We can dimensionalise the equations by dividing both $s$ and $c$ by $c_0$, which is equivalent to setting $c = c_0 C$ and $s = c_0 S$. We also choose a typical time scale $D/v^2$ and a length scale $D/v$ so that $t = (D/v^2)\tau$ and $z = (D/v)Z$, we have

$$\frac{\partial C}{\partial \tau} = D\frac{\partial^2 C}{\partial Z^2} - v\frac{\partial C}{\partial Z} - [\delta_1 + \delta_2(1-\gamma S)] + \Re C(1-C),$$

$$\frac{\partial S}{\partial \tau} = \delta_3(1-\gamma S)C,$$

where

$$\delta_1 = \frac{r_a D}{v^2}, \qquad \delta_2 = \frac{r_s\Theta D}{v^2},$$

$$\delta_3 = \frac{r_s\Theta A D}{v^2}, \qquad \Re = \frac{rD}{v^2}.$$

where $\gamma = c_0/s_{max}$. We can further assume that the concentration $c_0$ is small compared with the solid concentration at the absorption limit $s_{max}$ so that $\gamma \to 0$. If we use $\gamma = 0$, the equation for $S$ is simply

$$\frac{\partial S}{\partial \tau} = \delta_3 C,$$

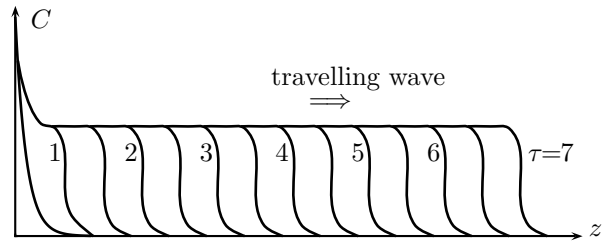



Figure 15.4: Travelling wave of bacteria mobility ($t = \tau D/v^2$).

which is decoupled from the equation for $C$ because $S$ can be determined once $C$ is known. Then, the equation for $C$ becomes

$$\frac{\partial C}{\partial \tau} = \frac{\partial^2 C}{\partial Z^2} - \frac{\partial C}{\partial Z} - (\delta_1 + \delta_2)C + \Re C(1 - C).$$

Using $\zeta = Z - \tau$ as the new variable, we can write the advection reaction-diffusion equation for $C$ as an ordinary diffusion equation, and we have

$$\frac{dC}{d\zeta} = D\frac{d^2 C}{d\zeta^2} - \delta C + \Re C(1 - C),$$

where $\delta = \delta_1 + \delta_2$. This a nonlinear ODE which can be solved numerically using the finite difference method discussed earlier in this book. The results are shown in Fig. 15.4 where $\delta = \Re = 0.1$ are used. We can see that a travelling wave of the bacteria concentration arises.

## 15.3 Theory of Consolidation

Consolidation is a well-studied phenomenon in porous media. In groundwater flow and pollutant transport, we often assume that the porosity does not change with time. In fact, porosity does change with time under appropriate conditions. This is subject to consolidation and compaction. Consolidation is a term often used in geotechnical engineering while compaction is often met in basin modelling and rock mechanics. Terzaghi first presented the theory of consolidation in 1925 and most of the work was about the prediction of settlement rates or consolidation of soil based on the following assumptions: 1) soil is homogeneous and fully saturated, and consolidation or compaction, is one-dimensional (vertical); 2) Darcy's flow is in one direction and the permeability $k$ is constant; 3) volume changes are caused solely by the effective stress,

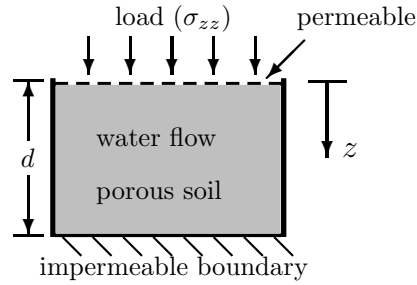



Figure 15.5: Terzaghi's consolidation test.

which is in turn associated with the changes in void ratio or porosity only. Assuming the conditions quoted above, and from Eq.(15.13), we can simplify pore pressure by setting $Q = 0$ as there are no sources or sinks. If the stress $\sigma_{kk}$ is constant or varies little, then we have

$$S_\sigma \frac{\partial p}{\partial t} = \frac{k}{\mu}\nabla^2 p, \tag{15.27}$$

or

$$\frac{\partial p}{\partial t} = c_v \nabla^2 p, \qquad c_v = \frac{k}{S_\sigma \mu}, \tag{15.28}$$

where $c_v$ is the consolidation coefficient. This is the (simplified) governing equation for pore pressure dissipation.

---

**Example 15.5:** Terzaghi's consolidation problem: the classical example is the consolidation test carried out by Terzaghi for a layer of fluid-saturated soil with a thickness of $d$. The top at $z = 0$ is loaded with a constant stress $\sigma_{zz} = -\sigma_{33} = \text{const}$. The application of this load will initiate an instantaneous response in the whole column under essentially undrained conditions, which means an almost constant pore pressure $p_0 = \Gamma\sigma_{33}$ where $\Gamma$ is the load efficiency.

The initial pore pressure is now $p(z, t = 0) = p_0$. The boundary condition at the top is that $p(0, t) = 0$. The bottom boundary is impermeable or $dp/dz|_{z=d} = 0$.

This problem is essentially the same as problem (5.29) with boundary conditions (5.30) discussed earlier. Therefore, the solution is the same as (5.43), and we have

$$p = \frac{4\Gamma\sigma_{zz}}{\pi}\sum_{n=0}^{\infty}\frac{1}{(2n+1)}\sin[\frac{(2n+1)\pi z}{2d}]e^{-(2n+1)^2 c_v \pi^2 t/4d^2}.$$

This solution suggests that $p \to 0$ or completely drained equilibrium as $t \to \infty$. The variations of pressure occur at a typical time scale $t_*$ when

the smallest exponent (when $n = 0$) is unity so that the largest time factor is $e^{-1}$. That is

$$\frac{c_v \pi^2 t_*}{4d^2} \approx 1,$$

which leads to

$$t_* \approx \frac{4d^2}{c_v \pi^2},$$

which means that the time for pore pressure dissipation will quadruple if the column depth is doubled.

---

In rocks and geotechnical engineering, it is common to express the consolidation in terms of porosity $\phi$ or void ratio $e$ rather than pore pressure. Let us now derive the consolidation in terms of $\phi$. From the fluid continuity equation (15.10), by setting $Q = 0$ and adding the term due to gravity, we have

$$\frac{\partial \theta}{\partial t} = \nabla \cdot [\frac{k}{\mu}(\nabla p + \rho_f g)], \tag{15.29}$$

where $\theta$ is the increment of fluid per unit volume, and for a fully liquid-saturated porous medium, the change of fluid content $\theta$ can be attributed to the change in porosity $\phi$. Thus, we can write $\theta = \varrho\phi$, where $\varrho$, the mass per unit fluid (pore) volume, is a constant, and $\varrho \approx 1$ can be used in most applications. Now we get

$$\varrho \frac{\partial \phi}{\partial t} = \nabla \cdot [\frac{k}{\mu}(\nabla p + \rho_f g)]. \tag{15.30}$$

From the previous chapter, we know that

$$p_e = P - \alpha p, \tag{15.31}$$

where $P$ is the total overburden pressure given by

$$P = \int_z^L \rho g dz = \int_z^L [\phi\rho_f + (1-\phi)\rho_s] g dz + P_0, \tag{15.32}$$

where $P_0$ is the external load at the boundary and $P_0$ can be considered as constant in most cases. $\rho_s$ and $\rho_f$ are the densities of the solid and fluid, respectively. Now we have

$$\nabla P = -\rho g = -[\phi\rho_f + (1-\phi)\rho_s] g. \tag{15.33}$$

By substituting $p$ in terms of $P$ and $p_e$, the governing equation (15.30) becomes

$$\varrho \frac{\partial \phi}{\partial t} = \nabla \cdot \{\frac{k}{\mu}[\frac{1}{\alpha}(-\nabla p_e + \nabla P) + \rho_f g]\}$$

$$= \nabla \cdot \{\frac{k}{\mu}[-\frac{1}{\alpha}\nabla p_e - \frac{\rho g}{\alpha} + \rho_f g]\} = \nabla \cdot \{\frac{k}{\alpha\mu}[-\nabla p_e + r]\}, \qquad (15.34)$$

where

$$r = -[\phi\rho_f + (1-\phi)\rho_s] + \alpha\rho_f g. \qquad (15.35)$$

In order to complete this equation we need to have a constitutive relationship between the effective pressure $p_e$ and the porosity $\phi$.

In the previous chapter, we know that the effective pressure $p_e$ is related to the volume strain $\Delta$. As a fluid such as water is incompressible, the volume strain in the porous medium is essentially the volume strain of the solid matrix, thus we can write the constitutive relation (14.129) as

$$p_e = -K_s \nabla \cdot \boldsymbol{U}, \qquad (15.36)$$

where $K_s$ is the bulk modulus of the solid matrix. We also use $\boldsymbol{U}$ to replace $\boldsymbol{u}$ as the displacement so as to avoid any possible confusion between the displacement and the velocity $\boldsymbol{v}$. By taking the material derivative of both sides of the above equation, we have

$$\dot{p}_e \equiv \frac{Dp_e}{Dt} = -K_s \nabla \cdot \dot{\boldsymbol{U}} = -K_s \nabla \cdot \boldsymbol{v}_s, \qquad (15.37)$$

where $\dot{\boldsymbol{U}} = \boldsymbol{v}_s$ is the velocity of the solid grains. For deformable media, the time derivative should be the material derivative given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \boldsymbol{v}_s \cdot \nabla, \qquad (15.38)$$

where $\boldsymbol{v}_s$ is the velocity of the solid. The material derivative is also called the Lagrange derivative as it follows the motion of the particle. If the consolidation is slow, then $\boldsymbol{v}_s \to 0$, so $D/Dt \approx \partial/\partial t$. Indeed, most processes in earth sciences are slow, so we do not have to worry about the difference between the material derivative and the standard time derivative.

---

**Example 15.6:** Let us now try to derive a constitutive relationship between $p_e$ and $\phi$ from first principles. The conservation of mass for the solid phase is

$$\frac{\partial[\rho_s(1-\phi)]}{\partial t} + \nabla \cdot [\rho_s(1-\phi)\boldsymbol{v}_s] = 0,$$

where the density of the solid grains $\rho_s$ =constant is often used. Dividing both sides of this equation by $\rho_s$ and rearranging it, we have

$$\frac{\partial(1-\phi)}{\partial t} + \boldsymbol{v}_s \cdot \nabla(1-\phi) = \frac{D(1-\phi)}{Dt} = -(1-\phi)\nabla \cdot \boldsymbol{v}_s.$$

Using Eq.(15.37), we have

$$\frac{D(1-\phi)}{Dt} = -(1-\phi)\frac{\dot{p}_e}{-K_s} = \frac{(1-\phi)}{K_s}\frac{Dp_e}{Dt},$$

or

$$\frac{1}{K_s}\frac{Dp_e}{Dt} = \frac{1}{(1-\phi)}\frac{D(1-\phi)}{Dt} = \frac{D\ln(1-\phi)}{Dt}.$$

The left hand of this equation is a function of $p_e$ only and the right-hand side is a function of $(1-\phi)$ only. This implies that $p_e$ is a function of $1-\phi$, that is $p_e = p_e(1-\phi)$. In fact, by integrating both sides with respect to $t$, we have

$$\frac{p_e}{K_s} = \ln(1-\phi) + A,$$

where $A$ is an integration constant. This is equivalent to

$$\phi = 1 - e^{-A+p_e/K_s} = 1 - Be^{p_e/K_s},$$

where $B = e^{-A}$. This indeed shows that $p_e$ is a function of $\phi$ only.

---

The relationship

$$p_e = p_e(\phi), \tag{15.39}$$

is often referred to as the constitutive relationship between the effective pressure and porosity. The actual relationship depends on the type of rocks or soil, and the type of problem. For example, the following relationship, known as the normal consolidation line, is widely used in consolidation of soil layers

$$e = e_0 - C_c \ln(p_e/p_0), \tag{15.40}$$

where $e$ is the void ratio, $C_c$ is the compression index, and $e_0$ is the void ratio at the initial effective pressure $p_0$. However, for the compaction in basin modelling, the following equation

$$\frac{p_e}{\bar{p}_0} = \ln\frac{\phi_0}{\phi} - (\phi_0 - \phi), \tag{15.41}$$

is often used. Here $\phi_0$ is the initial porosity while the $\bar{p}_0$ is the reference effective pressure.

Using the relationship Eq.(15.39), we can write Eq.(15.34) as

$$\varrho\frac{\partial\phi}{\partial t} = \nabla\cdot\{\frac{k}{\alpha\mu}[-p_e'(\phi)\nabla\phi + r]\}, \tag{15.42}$$

where $p_e'(\phi) = dp_e(\phi)/d\phi < 0$ because the increase of $p_e$ will decrease the porosity. This is the generalised consolidation equation under external load $P_0$ and its own overburden load due to gravity. It is also called the compaction equation in earth sciences.

The overburden load (or its own weight) only becomes significant when the the thickness $L$ of the consolidation layer is very large, and thus the overburden weight is comparable with or even much more significant than the external load $P_0$. This is true for some applications in earth sciences such as compaction of sediments in sedimentary basins and oil reservoirs, magma transport, and rock mechanics. However, for applications in geotechnical engineering, the overburden weight is not important, and this is equivalent to setting $g = 0$ or $r = 0$. In this latter case, we obtain a simplified consolidation equation

$$\frac{\partial \phi}{\partial t} = \nabla \cdot [\Lambda(\phi) \nabla \phi], \tag{15.43}$$

where

$$\Lambda(\phi) = -\frac{k(\phi)}{\mu} \frac{1}{\alpha \varrho} p_e'(\phi) > 0. \tag{15.44}$$

We can see that this consolidation equation (15.43) is essentially a nonlinear diffusion equation. In most applications, $\alpha \approx 1$ and $\varrho \approx 1$ are good approximations.

The permeability varies with porosity in complicated ways, depending on the type of rocks and soils. The most common function is the Carman-Kozeny relationship

$$k(\phi) = A \frac{\phi^m}{(1-\phi)^2}, \tag{15.45}$$

where $A$ is a constant but could depend on the grain size. $m$ is the exponent and typically $m = 2 \sim 8$ for rocks. It is often written as the simplified form

$$k(\phi) = k_0 (\frac{\phi}{\phi_0})^m, \tag{15.46}$$

where $k_0$ is the permeability at the initial porosity $\phi_0$.

## 15.4 Viscous Creep

### 15.4.1 Power-Law Creep

Power-law creep belongs to plastic deformation which involves many different mechanisms at the atomic levels. The basic equation is

$$\dot{e} = A \frac{\sigma^n}{d^m} \exp[-\frac{Q}{RT}], \tag{15.47}$$

where $\dot{e}$ is the strain rate and $A$ is a constant. $\sigma$ is the mean effective stress and $d$ is grain size. $R = 8.3144$ J/mole K is the universal gas

constant and $T$ is the absolute temperature. $Q$ is the activation energy required to get crystal dislocations moving, and its value is typically $100-500$ kJ/mole. $n$ and $m$ are constants depending on the mechanism and materials. The value $n$ is typically between 1 and 8. For dislocation creep, $n = 2 \sim 8$ with $m = 0$ (which means there is a grain size effect). For diffusion-controlled creep such as viscous creep, $n = 1$, and $m = 2 \sim 3$.

In the simplest case of linear creep law ($n = 1$), we have the simplified creep law

$$\sigma = \frac{d^m}{Ae^{-Q/RT}}\dot{e}. \tag{15.48}$$

Since $p_e = -\sigma$ and $\dot{e} = \nabla \cdot \dot{\boldsymbol{U}} = \nabla \cdot \boldsymbol{v}_s$ where $\boldsymbol{v}_s$ is the velocity of the solid. We now have

$$p_e = -\frac{d^m}{Ae^{-Q/RT}}\nabla \cdot \boldsymbol{v}_s = -\xi \nabla \cdot \boldsymbol{v}_s. \tag{15.49}$$

This is a linear constitutive equation for viscous creep.

It is worth pointing out that the creep law present here is different from the constitutive equation (14.129) or $\dot{p}_e = -K_s \nabla \cdot \boldsymbol{v}_s$ [see Eq.(15.37)] for the linear poroelasticity, and the mechanisms are also very different. In linear poroelasticity, the volume strain is related to the effective pressure, while in creep the strain rate is linked to the effective pressure and thus strain is continuously changing with time. The viscosity of the fluid is relatively low, about $10^{-3}$ N s m$^{-2}$ or $10^{-3}$ Pa s, but molten chocolate and toothpaste are highly viscous with a viscosity of about $10 \sim 10^3$ Pa s. Pitch has a viscosity of about $2 \times 10^8$ Pa s. Solid rock can behave like a fluid on a geological time scale, and its viscosity is in the order of $10^{20} \sim 10^{22}$ Pa s. We will try to derive the linear creep law from first principles in the rest of this section.

### 15.4.2 Derivation of creep law

The approach for deriving the law of viscous creep depends on the underlying mechanism. The classical theoretical consideration assumes a grain-boundary diffusion film of constant thickness and diffusivity, while other theories use the concept of a roughened, fluid-invaded nonequilibrium contact structure. Coble's classical treatment of grain boundary diffusion creep includes the kinetics of quartz dissolution/precipitation reaction. This 1-D approximation is only valid for a closed system when the thickness $w$ of the water film is small with respect to the grain diameter ($d$). However, this mechanism is somewhat biased toward grain-boundary diffusion-controlled pressure solution creep. This shortcoming of the creep laws can be overcome by

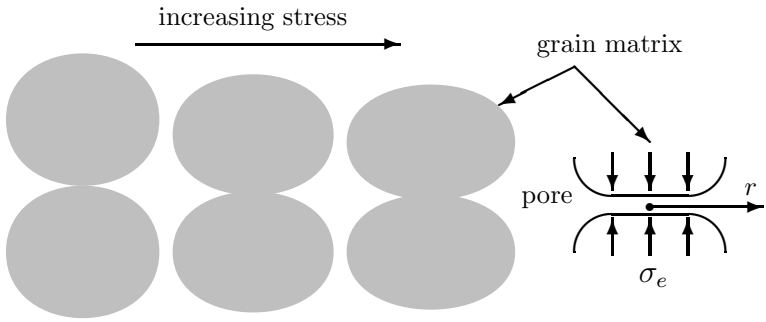



Figure 15.6: The contact areas increase as effective stress increases due to the pressure-enhanced diffusion along grain boundaries.

using solute transport and pressure solution deformation under possible open system conditions.

Now let us consider the intergranular contact region as a disc with a radius $r = L$ shown in Fig. 15.6. Let $J(r)$ be the radial component of solute mass flux, $\dot{e}$ be the average strain rate, and $v$ is the uniform shortening velocity of the upper grain relative to the lower grain due to the pressure solution creep. The kinetic relation between $v$ and $\dot{e}$ is

$$v = \dot{e}\bar{d}. \tag{15.50}$$

For simplicity, we assume that the film thickness $w$ does not change with time and the diffusion is near steady-state. Mass conservation gives

$$2\pi r J(r) + \rho_s \pi r^2 v = 0, \tag{15.51}$$

where $\rho_s$ is the density of grain solid. The flux $J(r)$ obeys Fick's Law

$$J(r) = -D_{gb} w \frac{dc}{dr}, \tag{15.52}$$

where $D_{gb}$ is the diffusivity of the solute in water along grain boundaries with a thickness $w$. $D_{gb}$ also varies with temperature $T$. In fact, we have

$$D_{gb}(T) = D_0 e^{-\frac{E_a}{RT}}, \tag{15.53}$$

where $D_0$ is the diffusivity at reference temperature $T_0$. $R$ is the universal gas constant. $E_a$ is the effective activation energy with a value of $5 \sim 100$ kJ/mol depending on the porous materials.

Let $c_0$ be the equilibrium concentration of the grain materials dissolved in pore fluid. Combing Eqs.(15.51) and (15.52), we have

$$\frac{dc}{dr} = \frac{\rho_s v}{2 D_{gb} w}. \tag{15.54}$$

Integrating it once and using the boundary conditions: $c_r = 0$ at $r = 0$, $c = c_0$ at $r = L$, we have the following steady state solution

$$c(r) = c_0 - \frac{\rho_s v}{4D_{gb}w}(L^2 - r^2). \tag{15.55}$$

The parabolic change of concentration $c(r)$ implies that the stress $\sigma(r)$ should be heterogeneously distributed in the contact region.

Experimental studies show that both concentration and thin film thickness depend on the effective stress $\sigma$, and they obey the following constitutive laws

$$c = c_0 \exp(-\frac{\nu_m \sigma_e}{RT}) \quad \text{and} \quad w = w_0 \exp(-\frac{\sigma_e}{\sigma_0}), \tag{15.56}$$

where $w_0$, $\sigma_0$ are constants depending on the properties of the thin film, and $\nu_m$ is the molar volume (of quartz). From the relation (15.56), we have

$$\sigma_e(r) = -\frac{RT}{\nu_m} \ln \frac{c(r)}{c_0}, \tag{15.57}$$

where we have used the condition $\sigma_e(r) = 0$ at $r = L$. Let $\sigma$ be the averaged effective stress, then

$$\pi L^2 \sigma = \int_0^L 2\pi \sigma_e(r) r dr. \tag{15.58}$$

Combining (15.57) and (15.58), we have

$$\sigma = -\frac{2RT}{\nu_m L^2} \int_0^L r \ln[1 - \frac{\rho_s \dot{e} \bar{d}}{4c_0 D_{gb} w}(L^2 - r^2)] dr. \tag{15.59}$$

Using (15.50) and integrating by parts, we have

$$\sigma = -\frac{RT}{\nu_m}[(1 - \frac{1}{BL^2}) \ln(1 - BL^2) - 1], \tag{15.60}$$

where

$$B = \frac{\rho_s \dot{e} \bar{d}}{4c_0 D_{gb} w}. \tag{15.61}$$

By defining a critical effective stress $\sigma_c$ (and equivalently a critical creep rate $\dot{e}_c$ ) when $BL^2 = 1$

$$\sigma_c = \frac{RT}{\nu_m}, \quad \dot{e}_c = \frac{4c_0 D_{gb} w}{\rho_s L^2 \bar{d}}, \tag{15.62}$$

(15.60) can be rewritten as

$$\frac{\sigma}{\sigma_c} = [1 - (1 - \frac{\dot{e}_c}{\dot{e}}) \ln(1 - \frac{\dot{e}}{\dot{e}_c})]. \tag{15.63}$$

A typical value of $\sigma_c$ is about 95 MPa with values of $T \sim 300$ K, $R \sim 8.31$ J mol$^{-1}$ K$^{-1}$, and $\nu_m \sim 2.6 \times 10^{-5}$ m$^3$ mol$^{-1}$.

Clearly, if $|\sigma| \ll \sigma_c$, we have

$$\dot{e} = \frac{4\nu_m c_0 D_{gb} w}{RT\rho_s \bar{d} L^2}\sigma = \frac{16\nu_m c_0 D_{gb} w}{RT\rho_s \bar{d}^3}\sigma, \tag{15.64}$$

or

$$\dot{e} = \frac{1}{\xi}\sigma, \qquad \xi = \frac{RT\rho_s \bar{d}^3}{16\nu_m c_0 D_{gb} w}, \tag{15.65}$$

which is a linear creep law in terms of stress. Here we have used $L = \bar{d}/2$. A different choice of $L = O(\bar{d})$ will only introduce an additional shape factor into the above relation. Under upper-crustal stress conditions $\sigma < 100$ MPa, the above approximation is valid as we expected. At higher stress states, we can use $|\sigma| \gg \sigma_c$, then (15.63) becomes

$$\dot{e} = \frac{4c_0 D_{gb} w}{\rho_s \bar{d} L^2}[1 - e^{-\frac{\nu_m \sigma}{RT}}]. \tag{15.66}$$

Let $L^2 = 4\bar{d}^2/\alpha_s$, and $\alpha_s = O(1)$ is a shape factor. The above relation (15.66) becomes

$$\dot{e} = \frac{\alpha_s c_0 D_{gb} w}{\rho_s \bar{d}^3}[1 - e^{-\frac{\nu_m \sigma}{RT}}], \tag{15.67}$$

which is consistent with Dewers and Hajash's empirical law derived from a quartz compaction experiment. It is worth pointing out that the creep law (15.67) degenerates into (15.65) when $\nu_m \sigma/RT \ll 1$, but it may be inaccurate when $|\sigma| \sim \sigma_c$.

Now let us estimate the order of $\xi$. For typical values of $R \approx 8.31$ J mod$^{-1}$ K$^{-1}$, $T \approx 300$ K, $c_0 \approx 10^{-4}$ M, $w_0 D_{gb} \approx 10^{-19}$ m$^3$ s$^{-1}$, $\bar{d} \approx 10^{-4}$ m, $\rho_s = 2.5 \times 10^3$ kg m$^{-3}$, and $\nu_m = 2 \times 10^{-5}$ m$^3$ mol$^{-1}$, we have

$$\xi = \frac{RT\rho_s \bar{d}^3}{16\nu_m c_0 D_{gb} w}$$

$$\approx \frac{8.31 \times 300 \times 2.5 \times 10^3 \times (10^{-4})^3}{16 \times 2 \times 10^{-5} \times 10^{-4} \times 10^{-19}} \approx 1.9 \times 10^{21} \text{Pa s}, \tag{15.68}$$

which suggests that rocks and thus the Earth's crust are highly viscous on a geological time scale.

Although mountains are formed due to the action of plate tectonics, the plate itself is not perfectly rigid and the solid rocky mountain can behave like a viscous chocolate paste on a geological time scale. In fact, recent observations in the Greek islands suggest that some of the mountainous surface has moved from 2m to 5m in the last hundred years. Let us see if this is possible.

**Example 15.7:** The spreading of a mountain: assuming that the mountain (idealised as a cone) is about $h = 1000$ m high and the radius of the base is 1000 m, the stress level is thus

$$\sigma \approx \rho g h = 2.5 \times 10^3 \times 10 \times 1000 \approx 2.5 \times 10^7 \text{Pa}.$$

If we use $\xi \approx 2 \times 10^{21}$ Pa s from earlier estimation (15.68), we know from Eq.(15.67) that the strain rate is

$$\dot{e} = \frac{\sigma}{\xi} \approx \frac{2.5 \times 10^7}{2 \times 10^{21}} \approx 2.5 \times 10^{-14} \text{m/s} \approx 4 \times 10^{-7} \text{m/year}.$$

This seems too small, but we have to know that it is the strain rate, not the displacement. The strain increment in a year is about

$$e = \dot{e} \times t(\text{year}) = 4 \times 10^{-7} \text{m}.$$

As the strain is defined as $e = \delta h/h$, the displacement in a year is about

$$\delta h \approx 1000 \times e \approx 4 \times 10^{-4} \text{m} = 0.4 \text{mm}.$$

This is indeed too small. However, over a period of a million years, the total displacement is about

$$\delta h \times 10^6 \approx 400 \text{m}.$$

Over a million years, erosion occurs at a considerable rate. In addition, the plate has moved for about $20$ km assuming that the rate of plate motion is about 2cm or $2 \times 10^{-2}$ m/year.

For the surface movement in Greek islands to be a few centimetres a year, we have to use 2 orders lower (or $1/100$) of the viscosity $\boldsymbol{x} \approx 2 \times 10^{-19}$. Therefore, if we believe the observed data is linked to the viscous mechanism, this means that the apparently rigid mountains are more fluid-like than we expected.

Now let us look at the driving force of plate tectonics using a similar estimation. We know that the plate is moving at about the speed $1 \sim 9$ cm/year, depending on the location, so we will choose a moderate speed $v = 5$ cm/year or $v \approx 1.6 \times 10^{-9}$ m/s. As the unit of strain rate is $\text{s}^{-1}$, so for the mantle of a thickness of $L = 3000\text{km} = 3 \times 10^6$ m, the strain rate is

$$\dot{e} = \frac{v}{L} \approx \frac{1.6 \times 10^{-9}}{3 \times 10^6} \frac{\text{m/s}}{\text{m}} \approx 5 \times 10^{-16} \frac{1}{\text{s}}.$$

Using the viscosity $\xi = 2 \times 10^{21}$, we can estimate the driving force by mantle convection is about

$$\sigma \approx \xi \dot{e} = 2 \times 10^{21} \times 5 \times 10^{-16} \approx 10^6 \text{Pa},$$

which is just about 10 atmospheric pressures. Thus, the driving shear force by mantle convection on such a large scale is surprisingly small. Such force can be provided by thermal convection of the mantle or even by pulling force due to the density difference of the subduction zone. Of course, the reality is much more complicated than this. The main point here is that for plate tectonic phenomena to happen, the large scale convection does not necessarily require a big force.

---

The constitutive relation (15.65) is only valid for the one-dimensional case. We can extend it to a more general form $\sigma = \xi \dot{e}_{kk}$. Noting that $p_e = -\sigma$ and $\dot{e}_{kk} = \nabla \cdot \dot{\boldsymbol{U}} = \nabla \cdot \boldsymbol{v}_s$, we have the following compaction law or creep law

$$p_e = -\xi \nabla . \boldsymbol{v}_s. \tag{15.69}$$

This creep relationship is widely used in studies of magma transport and plate tectonics and other branches of Earth sciences.

## 15.5 Hydrofracture

Hydrofracturing is a major mechanism to create fractures in porous rocks, soils and oil reservoirs. It is also a major technique for groundwater remediation. For example, injecting water under high pressure into a bedrock formation via a low yield water well will increase the size and extent of the existing fractures in the bedrock, resulting in an increase of the permeability. This will in turn increase the ultimate yield capacity of the water well. This same procedure also applies to the boreholes for oil reservoirs. In this final section, we will briefly review the basic mechanism of hydrofracturing and related processes such as magma transport and diagenetic reactions.

### 15.5.1 Hydrofracture

For a fluid-saturated porous medium, we know that the effective pressure $p_e$ is related to the total pressure $P$ and the pore pressure $p$

$$p_e = P - p. \tag{15.70}$$

The overpressure is the difference of the pore pressure and the hydrostatic pressure. If there is substantial overpressure induced by either diagenesis or injected water or any other mechanisms, it is possible that the pore pressure could significantly exceed its hydrostatic pressure. It is even possible that the pore pressure is greater than the total pressure $P$ in certain regions.

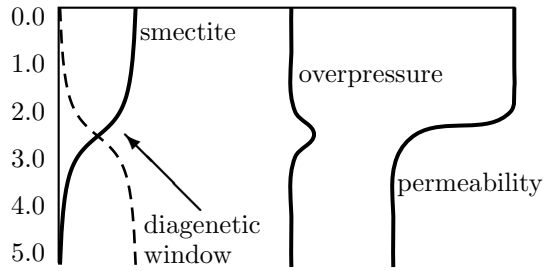



Figure 15.7: Hydrofracture via overpressuring due to the fluid generation from the dehydration of water-rich minerals.

From the above equation, we know that if $p > P$, we have $p_e < 0$. In reality, the physics requires that the effective pressure should be non-negative. What happens here is that fracture is initiated in rocks so as to dissipate the overpressure. This is the basic mechanism for hydrofracturing.

Most porous materials such as soil and rocks have limited (usually low) tensile strength, and small amounts of tensile stress due to overpressure will result in fracturing. From the fracture mechanics discussed in the previous chapter, we know that the fracture formation depends on the state of stress, pore fluid pressure, and the properties of the rocks.

From the Hall-Petch relationship Eq.(14.112), we know that coarse-grained rock is weaker than fine-grained rock. For example, it is observed that in many sedimentary basins, sandstones tend to be fractured while mudstones are not.

In groundwater remediation, the high-pressure water is injected into the ground to create the microfractures by the hydrofracturing process, resulting in an increase in permeability (and thus hydraulic conductivity) in surrounding porous rock layers. This will increase the water yield in wells. The same procedure can also be applied to create fractures in oil reservoirs so as to increase the production of a low yield borehole.

### 15.5.2 Diagenesis

Diagenesis is an important process in sedimentary basins, and there are many diagenetic reactions. The major process is the dewatering or dehydration reaction of water-rich minerals such as smectite (S) which

is transformed into more stable illite (I), releasing its bound water. This reaction can schematically be represented as

$$S \longrightarrow I + n\ H_2O, \tag{15.71}$$

where $n \approx 5$ the number of moles of water molecules released per mole of smectite. Let $\phi$, $S$ and $I$ be the porosity, volume fractions of smectite, and volume fraction of illite, respectively. For the simplified model given by Fowler and Yang in 2003, the governing equations become

$$\frac{\partial S}{\partial t} = -e^{\beta(T_0+\alpha z-T_*)}S, \tag{15.72}$$

$$\frac{\partial I}{\partial t} + (1-r)\frac{\partial S}{\partial t} = 0, \tag{15.73}$$

$$\frac{\partial \phi}{\partial t} = \lambda \frac{\partial^2 \phi}{\partial z^2} + re^{-\beta(T_0+\alpha z-T_*)}S, \tag{15.74}$$

where $\lambda$, $r$ and $\beta$ are constants. $z$ is the depth from the ocean floor. $T_0 \approx 300$K is the temperature at the ocean floor, and $\alpha = 30$K/km is the temperature gradient. $T_* \approx 90°$C $\approx 363$K is the critical temperature to switch on the diagenetic reaction, and this means that the diagenetic reaction typically occurs at a depth of about $2.5 \sim 3.5$ km. Here the equations have been written in non-dimensional form, and the parameters are given by

$$\lambda = \frac{D\tau}{d^2}, \qquad r = A_0\tau, \tag{15.75}$$

where $A_0$ is the Arrhenius coefficient and $D$ is the diffusion coefficient. The length scale is $d = 1000$ m and the time scale is $\tau = 1$ million years (or $0.3 \times 10^{14}$ s). For typical values of $D \approx 3 \times 10^{-8}$ m$^2$/s and $A_0 \approx 3 \times 10^{-15}$ s$^{-1}$, we have $\lambda \approx 1$ and $r \approx 0.1$.

For the values of $\lambda = 1$, $\beta = 2$, and $r = 0.1$, we can solve the above equation numerically. The results are shown in Fig. 15.7 where we can see that there is a diagenetic window for dehydration. In this window, the overpressure is high because the bound water is released here. This overpressuring will lead to the increase of porosity via hydrofracturing, and this fracturing process is reflected by the increase of permeability starting from the diagenetic window up to the ocean floor. This hydrofracture is a mechanism for creating vertical cracks in sedimentary basins.

### 15.5.3 Dyke and Diapir Propagation

Another important fracturing process related to overpressuring is the propagation of the magma dykes. Magma rise is driven by the buoyancy

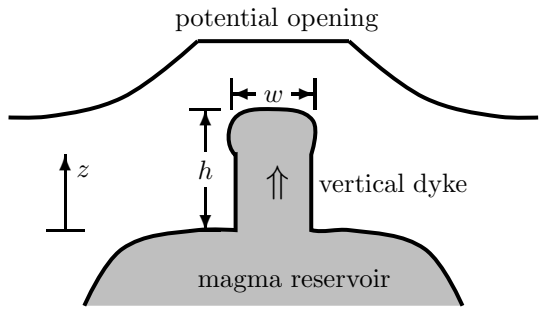



Figure 15.8: Formation of vertical dykes by magma column rising from a magma reservoir.

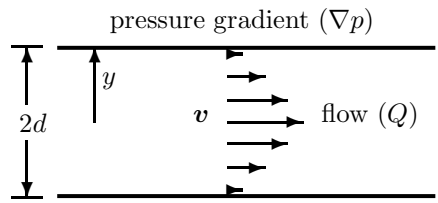



Figure 15.9: Laminar flow between two plates.

force due to density difference, and the resistance forces are the viscous shear stress and fracture toughness. Dykes are formed through magma intrusion into brittle rocks with a very small aspect ratio of width to length, typically several metres wide and up to hundreds of kilometres long. For a potential vertical dyke connected to a large magma reservoir (see Fig. 15.8), its height and width are $h$ and $w$, respectively. We can idealise it as the viscous laminar flow or Poiseuille flow between two parallel plates $2d$ apart (see Fig. 15.9). The governing equation is the viscous flow equation Eq.(5.25)

$$\xi \nabla^2 \boldsymbol{v} = \nabla p, \tag{15.76}$$

where $\nabla p$ is the pressure gradient and $\xi$ is the viscosity of the fluid.

For the one-dimensional flow with $\nabla p = \frac{\partial p}{\partial x} = \Delta p$=constant (pressure drop per unit length), the equation simply becomes

$$\frac{\partial^2 v}{\partial y^2} = \frac{\Delta p}{\xi}. \tag{15.77}$$

The boundary condition at $y = d$ is $v = 0$ (non-slip condition), while

the shear $\tau = \xi \frac{\partial v}{\partial y} = 0$ at $y = 0$ by symmetry. Integrating the above equation twice, we have

$$v = \frac{\Delta p}{2\xi} y^2 + Ay + B. \tag{15.78}$$

From $\frac{\partial v}{\partial y} = 0$, we have $A = 0$. From $v = 0$ at $y = d$, we have $B = -\frac{\Delta p}{2\mu} d^2$. Thus, the solution becomes

$$v = -\frac{\Delta p}{2\xi}(d^2 - y^2). \tag{15.79}$$

The total flow rate per unit thickness (perpendicular to the page) is the sum of all the layers or the integration. We have

$$Q = \int_{-d}^{d} v dy = 2 \int_{0}^{d} v dy = -\frac{2\Delta p}{3\xi} d^3. \tag{15.80}$$

Now coming back to our original problem, we know that the buoyancy force is given by $f_b = \delta \rho g z$ where $\delta\rho$ is the density difference and $g$ is the acceleration due to gravity. Therefore, the pressure gradient is $\Delta p = \frac{\partial p}{\partial z} = -\delta \rho g$ because pressure decreases as $z$ increases. Using $d = w/2$, we have

$$Q = \frac{\delta \rho g}{12\xi} w^3. \tag{15.81}$$

Since the total width is $w$, the averaged flow velocity $\bar{v}$ is simply

$$\bar{v} = \frac{Q}{w} = \frac{\delta \rho g}{12\xi} w^2. \tag{15.82}$$

As the rock-magma viscosity and the magma viscosity are different, an effective or combined viscosity should be used for the magma ascent. For silicic magma, the density difference is about $100 \sim 200$ kg/m$^3$. The viscosity of magma is controlled by composition (such as Si and water content), volatility of gas (such as water vapour), temperature, and pressure. The viscosity varies from $10^2$ Pa s (mafic or basaltic magma) to $10^7$ Pa s (felsic magma) and up to $10^{12}$ Pa s (such as rhyolite). For a typical width of $w = 2$ m, the average velocity of magma ascent is $\bar{v} = \frac{\delta \rho g}{12\xi} w^2 \approx 3 \times 10^{-10} \sim 5$m/s. Well, this estimate varies too widely and is probably not much use. However, it does give us an indication of how the flow rate varies with the width of the dyke. In fact, the magma flow velocity in most cases is in the range of $10^{-3} \sim 20$ m/s or 4 m/hr to 70km/hr. The wide range of this velocity is mainly controlled by pressure and more importantly by viscosity. For example, the dry rhyolite magma has a viscosity of about $10^{11}$ Pa s at

1100K, if its water content increases up to 5%, then its viscosity will reduce to $10^5$ Pa s. In addition, the crystallinity or the content of the suspended solid crystals (expressed as volume fraction $\psi$) will affect the the effective viscosity of the magma, which obeys the Einstein-Roscoe equation

$$\xi = \xi_0(1 - \beta\psi)^{-2.5}, \tag{15.83}$$

where $\xi_0$ is the viscosity when $\psi = 0$ and $\beta \approx 1.67$ is a constant.

Here we use the linear viscous relationship; if we use a nonlinear relationship, then the width-dependence is even more complicated. If the magma supply is limited, the cooling time scale is typically $\tau = d^2/\kappa$ [see Eq.(1.17) in Chapter 1] where $\kappa$ is the thermal diffusivity. For a typical value of $\kappa \approx 1 \times 10^{-6}$ m$^2$/s, we have $\tau \approx 10$ days for $d = 1$m and $\tau \approx 11$ years for $d = 20$ m. If the temperature is too low, the magma may solidify and block the flow though the channel. If the viscosity is too high, it may also clog the channel. In general, magma with low viscosity flows quickly and covers a large area (say, hundreds of square kilometres), while magma with high viscosity flows slowly and covers a relative small region. Volatile gases escape more quickly in low viscosity magma, while gas pressure may build up in high viscosity magma, causing potential violent eruptions. However, if overpressure is high enough, the high pressure of the underlying magma may overcome the blockage and the magma will erupt in an explosive manner. If the overpressure dissipation is enough, the channel may be clogged again.

It is worth pointing out that the above estimation is mainly valid for the flow in existing fissures where flow and fracture can easily be formed. If there is no existing fissure, then the fracture of the rock near the head of the magma column is governed by the local stress so that the whole column acts as an edge crack. Therefore, the fast fracture criterion discussed in the previous chapter still applies

$$K_{IC} = \sigma\sqrt{\pi h}, \tag{15.84}$$

where $\sigma$ is the principal stress, and $K_{IC}$ is the critical stress intensity factor of the surrounding rock. If the column height $h$ is relatively large, then a small stress concentration will lead to further crack propagation.

Another interesting related phenomenon is the propagation and emplacement of diapirs. The motion of a diapir is similar to the process inside a lava lamp. As an example, let us look at a diapir in detail.

---

**Example 15.8:** A diapir is formed by intrusion under buoyancy and pressure difference. Typical scale of diapirs is several kilometres. For a diapir ascent, the Poiseuille flow in a cylinder is a crude model. From

Eq.(4.30), we know that $\Delta P/L = -\delta\rho g$. Thus, the total flow rate is

$$Q = \frac{\pi g \delta\rho}{128\xi} w^4, \tag{15.85}$$

where $w$ is the diameter of the diapir, and $\xi$ is the effective combined viscosity. As the area of the circular diapir is $A = \pi(w/2)^2$, the average ascent velocity is $\bar{v} = \frac{g\delta\rho}{32\xi} w^2$. For the typical values of $w = 2$km, $\xi = 10^{17}$ Pa s, and $\delta\rho = 200$ kg/$m^3$, we have

$$\bar{v} = \frac{g\delta\rho}{32\xi} w^2 \approx \frac{9.8 \times 200}{32 \times 10^{17}} (2000)^2 \approx 1.23 \times 10^{-9} \text{m/s} \approx 0.04 \text{m/year},$$

which is about 4 cm/year. It is a significant speed geologically.

An interesting implication of this estimation is that $\bar{v} \propto w^2$. This means that larger diapirs will rise much faster than small diapirs. This is because larger diapirs have a much greater force of buoyancy to overcome the resistance of overlying rock layers. Obviously, as the diapir body moves up, the temperature becomes lower, the density of the ascent body becomes closer to the density of the surrounding rock. Subsequently, the density difference is getting smaller, so is the buoyancy. Studies show that the diapir ascent rate varies from $0.05 \sim 7$m/year (at a depth between 40 and 10 km) to a slower rate $0.01 \sim 0.1$ m/year (at depths between 5 and 10 km). Alternatively, we know that the time scale of cooling is $\tau = d^2/\kappa$, thus the diapir has moved up $h = \bar{v}\tau$ during the period of $\tau$. For a diapir with $d = 1000$m and $\kappa \approx 10^{-6}$ m$^2$/s, we have $\tau \approx 32,000$ years, so that $h = 1250$ m. This again implies that large diapirs move quickly on the geological time scale. The final emplacement of a diapir depends on many factors and the most important factor is probably the density of the surrounding rock.

The cylindrical model may be too far from the reality. A better model to consider is that of the diapir as a sphere rising through the viscous rock (like a rising balloon). Since both the diapir and the surrounding rock are highly viscous, the actual model is modified as

$$v = \frac{\alpha}{3} \frac{\delta\rho g a^2}{\xi_r}, \qquad \alpha = \frac{\xi_r + \xi_d}{\xi_r + \frac{3\xi_d}{2}}, \tag{15.86}$$

where $a = w/2$ is the radius of the spherical diapir. $\xi_r$ and $\xi_d$ are the viscosities of the surrounding rock and the diapir, respectively. In the case of $\xi_d \gg \xi_r$, we have $\alpha \approx 2/3$. Then, the above equation reduces to the Stokes equation $v \approx \frac{2\delta\rho g a^2}{9\xi_r}$ which is the settling velocity of a rigid sphere in a viscous fluid with viscosity $\xi_r$. However, $\xi_d$ is usually two or

three orders lower than $\xi_r$, and their typical values are $\xi_d \approx 10^{16} \sim 10^{19}$ Pa s and $\xi_r \approx 10^{19} \sim 10^{22}$ Pa s. More sophisticated mathematical models are needed to get a better estimate and this is still an area of active research.

# References

Biot M. A., General theory of three-dimensional consolidation, *J. Appl. Phys.*, **12**, 155-164 (1941).

Britannica Encyclopedia, (2001).

Burov E., Jaupart C. and Guillou-Frottier L., Ascent and emplacement of buoyant magma bodies, *J. Geophys. Res.*, **108**, 01904(2003).

Clemens J. D., Observations on the origins and ascent mechasims of granitic magmas, *J. Geol. Society London*, **155**, 845-851 (1998).

Fowler A. C., A compaction model for melt transport in the Earth's asthenosphere, in: *Magma Transport & Storage*, Ed. Ryan., (1991).

Fowler A. C., *Mathematical models in the applied sciences*, Cambridge University Press, (1997).

Greenkorn R. A., *Flow phenomena in porous media*, Marcel Dekker Inc., (1983).

Lewis R., Pao W. and Yang X. S., Instability and reaction-diffusion transport in bacteria, *Com.Num.Meth.Eng.*, **20**, 718-729 (2000).

Murch B. W. and Skinner B. J., *Geology Today - Understanding Our Planet*, John Wiley & Sons, (2001).

Ryan M. P., *Magma Transport and Storage*, Wiley & Sons, (1991).

Skempton A. W., Effective stress in soils, concrete and rocks, in: *Pore Pressure and Suction in Soils*, Butterworths, London, (1960).

Wang H. F., *Theory of Linear Poroelasticity: with applications to geomechanics and hydrogeology*, Princeton University Press, (2000).

Terzaghi K., *Theoretical Soil Mechanics*, John Wiley, New York, (1943).

# Appendix A

# Mathematical Formulae

## A.1 Differentiation and Integration

### A.1.1 Differentiation

**Differentiation Rules:**

$$(uv)' = u'v + uv', \qquad (\frac{u}{v})' = \frac{u'v - uv'}{v^2}$$

$$f[g(x)]' = f'[g(x)] \cdot g'(x)$$

**Leibnitz's Theorem:**

$$\frac{d^n}{dx^n}(uv) = u^{(n)}v + bu^{(n-1)}v' + ... + \binom{n}{k}u^{(n-k)}v^{(k)} + ... + uv^{(n)},$$

$$\binom{n}{k} = {}^nC_k = \frac{n!}{k!(n-k)!}$$

### A.1.2 Integration

**Integration by parts**

$$\int_a^b u\frac{dv}{dx}dx = [uv]\Big|_a^b + \int_a^b v\frac{du}{dx}dx$$

**Differentiation of an integral**

$$\frac{d}{dx}\int_{a(x)}^{b(x)} u(x,y)dy = [u(x,b)\frac{db}{dx} - u(x,a)\frac{da}{dx}] + \int_{a(x)}^{b(x)} \frac{\partial u(x,y)}{\partial x}dy$$

### A.1.3 Power Series

$$e^z = 1 + z + \frac{z^2}{2!} + ... + \frac{z^n}{n!}... \quad (z \in \mathcal{C})$$

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - ..., \qquad \cos z = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - ...$$

$$\sinh z = z + \frac{z^3}{3!} + \frac{z^5}{5!} + ..., \qquad \cosh z = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + ...$$

### A.1.4 Complex Numbers

$$e^{i\theta} = \cos\theta + i\sin\theta, \qquad z = x + iy = re^{i\theta} = r(cos\theta + i\sin\theta)$$

**De Moivre's formula**:

$$[r(\cos\theta + i\sin\theta)]^n = r^n(\cos n\theta + i\sin n\theta)$$

## A.2 Vectors and Matrices

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta = a_x b_x + a_y b_y + a_z b_z$$

$$\mathbf{a}\times\mathbf{b} = \mathbf{n}|\mathbf{a}||\mathbf{b}|\sin\theta = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}$$

**Vector Triple**

$$\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}$$

$$\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \mathbf{b}\cdot(\mathbf{c}\times\mathbf{a}) = \mathbf{c}\cdot(\mathbf{a}\times\mathbf{b}) = -\mathbf{a}\cdot(\mathbf{c}\times\mathbf{b})$$

$$\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = (\mathbf{a}\cdot\mathbf{c})\mathbf{b} - (\mathbf{a}\cdot\mathbf{b})\mathbf{c}$$

**Gauss' Divergence Theorem**

$$\iiint_V \nabla\cdot\boldsymbol{u}dV = \iint_S \boldsymbol{u}\cdot d\mathbf{A}$$

**Stokes' Theorem**

$$\iint_S (\nabla\times\mathbf{u})\cdot d\mathbf{A} = \oint_\Gamma \mathbf{u}\cdot d\mathbf{l}$$

**Green's Theorems**

$$\int_V (\psi\nabla^2\phi - \phi\nabla^2\psi)dV = \int_S (\psi\frac{\partial\phi}{\partial n} - \phi\frac{\partial\psi}{\partial n})|dS|$$

$$\oint (udx + vdy) = \int\int (\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y})dxdy$$

**Identities**

$$\nabla \cdot \nabla \times \mathbf{u} = 0, \qquad \nabla \times \nabla\phi = 0$$

$$\nabla \times (\phi\mathbf{u}) = \phi\nabla \times \mathbf{u} + (\nabla\phi) \times \mathbf{u}$$

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

$$\nabla \times (\nabla \times \mathbf{u}) = \nabla(\nabla \cdot \mathbf{u}) - \nabla^{\mathbf{2}}\mathbf{u}$$

**Matrices**

$$(\mathbf{AB...Z})^{\mathbf{T}} = \mathbf{Z}^{\mathbf{T}}...\mathbf{B}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}, \ \ (\mathbf{AB...Z})^{-\mathbf{1}} = \mathbf{Z}^{-\mathbf{1}}...\mathbf{B}^{-\mathbf{1}}\mathbf{A}^{-\mathbf{1}}$$

$$|\mathbf{AB....Z}| = |\mathbf{A}||\mathbf{B}|...|\mathbf{Z}|, \qquad |\mathbf{A}| = \det\mathbf{A}$$

**Trace and Determinants**

$$\mathbf{Av}_i = \lambda_i\mathbf{v}_i, \qquad \mathrm{eig}(\mathbf{AB}) = \mathrm{eig}(\mathbf{BA})$$

$$\mathrm{tr}(\mathbf{A}) = \sum_i \mathbf{A}_{ii} = \sum_i \lambda_i, \qquad \lambda_i = \mathrm{eig}(\mathbf{A})$$

**Exponential Matrices**

$$e^{\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!}\mathbf{A}^n = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + ...$$

$$e^{t\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!}(t\mathbf{A})^n = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2}\mathbf{A}^2 + ...$$

## A.3 Asymptotic Expansions

**Gaussian Function**

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(x-\mu)^2}{2\sigma^2}] \to \delta(x) \qquad \sigma \to \infty.$$

**Binomial Distribution**

$$g(k;n,p) = \binom{n}{k}p^k(1-p)^{n-k}, \ \ (k = 0, 1, 2, ...n)$$

$$g(k \to x; n \to \infty, p) \sim f(x; \lambda = np), \qquad (\lim_{n\to\infty} np = \lambda)$$

$$g(x; n \to \infty, p)|_{np\gg 1} \sim p(x; \mu, \sigma), \;\; \mu = np, \; \sigma^2 = np(1-p)$$

**Poisson Distribution**

$$f(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x \in \mathcal{N}$$

$$f(x; \lambda \gg 1) \approx p(x; \mu, \sigma), \qquad \mu = \lambda, \;\; \sigma^2 = \lambda$$

**Stirling's Formula**

$$n! \approx (\frac{n}{e})^n \sqrt{2\pi n}, \qquad n \gg 1$$

**Error Function**

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\eta^2} d\eta, \qquad \mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$$

$$\mathrm{erf}(x) \sim 1 - \frac{e^{-x^2}}{x\sqrt{\pi}}, \; (x \to \infty), \;\; \mathrm{erf}(x) \sim \frac{2}{\sqrt{\pi}}[x - \frac{x^3}{3} + \frac{x^5}{10} - ...], \; (x < \infty).$$

# References


Abramowitz M. and Stegun I. A., *Handbook of Mathematical Functions*, Dover Publication, (1965).

Courant R. and Hilbert, D., *Methods of Mathematical Physics*, 2 volumes, Wiley-Interscience, New York, (1962).

Korn G. A. and Korn T. M., *Mathematical Handbook for Scientists and Engineers*, Dover Publication, (1961).

Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

Pearson C. E., *Handbook of Applied Mathematics*, 2nd Ed, Van Nostrand Reinhold, New York, 1983.

Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, 3rd Ed, Cambridge University Press, (2006).

Selby S. M., *Standard Mathematical Tables*, CRC Press, (1974).

# Appendix B

# Matlab and Octave Programs

The numerical algorithms discussed in this book can be implemented in any programming languages, however, we do have to make a choice when implementing them, though they should not be tied to a particular programming language. In this sense, this appendix can be used alone and/or with reference to the main text of this book.

We will implement some of the algorithms using both Matlab and Octave. The reasons for our choice are twofold: 1) Matlab is a popular script programming language with many powerful functions for manipulating matrice and vectors as well as visualisation, and it is widely accessible (www.mathworks.com); 2) Octave is an open source software (www.octave.org) that has many functions that are compatible with Matlab and thus most Matlab programs can run directly on an Octave platform (or with minimal modifications).

## B.1 Gaussian Quadrature

For an integral with finite integration limits $a$ and $b$, the 7-point Gaussian quadrature is given by

$$I = \int_a^b f(x)dx = \frac{(b-a)}{2}\int_{-1}^{1} f[\frac{(b-a)(\zeta+1)}{2}+a]d\zeta =$$

$$\approx \frac{b-a}{2}\sum_{i=1}^{n} w_i f[\frac{(b-a)(\zeta_i+1)}{2}+a], \tag{B.1}$$

where we have used $x = \frac{(b-a)(\zeta+1)}{2} + a$. The Gaussian points are

$$\zeta_1 \approx -0.94910,\ \zeta_2 \approx -0.74153,\ \zeta_3 \approx -0.40584,$$

$$\zeta_4 = 0.0,\ \zeta_5 = -\zeta_3,\ \zeta_6 = -\zeta_2,\ \zeta_7 = -\zeta_1, \tag{B.2}$$

and their corresponding weighting coefficients are

$$w_1 \approx 0.12948,\ w_2 \approx 0.27970,\ w_3 \approx 0.38183,$$

$$w_4 \approx 0.41795,\ w_5 = w_3,\ w_6 = w_2,\ w_7 = w_1. \tag{B.3}$$

This following implemented program works in both Matlab and Octave. To run the program, type in (only about 16 lines) and save to a file, Gauss_quad.m in any directory, say, D:/programs.

For Matlab (Octave), first launch Matlab (or Gnu Octave), then change to the directory where the file was saved (using >cd D:/programs), and type in

>Guass_quad ('exp(-x.^2)*2/sqrt(pi)',0,1)

It will display

$$I = 0.84270079279326.$$

We know that the exact value is $\mathrm{erf}(1) = 0.84270079294971...$, and we see that the 7-point Guassian quadrature is accurate to the 9th decimal place. The function in single quotation can be any valid Matlab or Octave expression, however, the power .^ (instead of ^) and ./ (instead of /) operators should be used as they are element-wise operations. Alternatively, the vectorization command or vectorize can be used. It is worth pointing out that when you install the Gnu Octave, you probably have to install Gnuplot as well (www.gnuplot.info) if you have not installed it on your computer.

**Gauss_quad.m**

```
% Numerical Integration of I=\int_a^b f(x) dx
% by 7-point Gaussian Quadrature
% Programmed by X S Yang (Cambridge University)
% Usage: Gauss_quad(function_str,a,b)
%   E.g. Gauss_quad('(sin(x)./x).^2',0,pi);
function [I]=Gauss_quad(fstr,a,b)                % line 1
% Power (.^) and division (./) in fstr should be used
format long;                                     % line 2
if nargin<3,
 disp('Usage:Gauss_quad(integrand,a,b)');
 disp('E.g., Gauss_quad(''exp(-x.^2)*2./sqrt(pi)'',-1,1);');
end
% Default function and values if no input
```

```
if nargin<1,                                          % line 7
   help Gauss_quad.m;
   fstr='exp(-x.^2)*2/sqrt(pi)';
   a=-1.0; b=1.0;
end                                                   % line 11
% Converting the input integrand, fstr, to a function f(x)
f=inline(fstr,0);
% Seven-point integration scheme so zeta_1 to zeta_7
zeta=[-0.9491079123; -0.7415311855; -0.4058451513; 0.0;
       0.4058451513; 0.7415311855; 0.9491079123];
% Weighting coefficients
w=[0.1294849661; 0.2797053914; 0.3818300505; 0.4179591836;
   0.3818300505; 0.2797053914; 0.1294849661];
% Index for the seven points
Index=1:7;                                            % line 15
I=(b-a)/2*sum(w(Index).*f((b-a).*(zeta(Index)+1)/2+a));
disp(' '); disp('The integral is '); I                % line 16
```

## B.2 Newton's Method

The roots of a function $f(x) = 0$ can be found using Newton's iteration method

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{B.4}$$

where the initial value $x_{n=0}$ can be a random guess. So we use $x_0 =$randn where randn is a random number drawn from a normal distribution. In case of multiple roots, it will only produce a single root as the initial guess is random. To get multiple roots, you can call the function many times to obtain all the roots.

**Newton_matlab.m**

```
% Finding roots of f(x)=0 via the Newton's iteration method
% Programmed by X S Yang (Cambridge University)
% Usage: Newton(function_str); E.g. Newton('x.^5-pi');
% [Notes: Since the initial guess is random, so in case
%  of multiple roots, only a single root will be given.]
function [root]=Newton(fstr)                          ! line 1
format long;                                          ! line 2
% Default function and values if no input
if nargin<1,
   help Newton.m;
   fstr='x.^5-pi';                                    ! line 5
end
% Tolerance (to the tenth decimal place)
```

```
   delta=10^(-10);
% Converting the input fstr to a function f(x)
f=inline(fstr);
% Defining x as the independent variable
syms x;
% Find the f'(x) from f(x) using diff
fprime=inline(char(diff(f(x),x)));                ! line 10
% Initial random guess
xn=randn;    deltax=1;
% Iteration until the prescribed accuracy
while (deltax>delta)
    root=xn-f(xn)/fprime(xn);
    deltax=abs(root-xn);
    xn=root;                                      ! line 15
end
disp(strcat(fstr, ' has a root')); root           ! line 17
```

For example, type in >Newton('x.^5-pi'), it will produces a root

$$\text{root} = 1.2572741156,$$

which is accurate to the 10th decimal place. However, as we used some symbolic function in Matlab to obtain $f'(x)$, the Octave version is slight different, you have to supply the $f'(x)$. For the same example, we now have to type in >Newton_octave('x.^5-pi','5*x.^4').

**Newton_octave.m**

```
% Finding roots of f(x)=0 via the Newton's iteration method
% Programmed by X S Yang (Cambridge University)
% Usage: Newton(function,fprime);
% E.g. Newton('x.^5-3.1415','5*x.^4');
% [Notes: Since the initial guess is random, so in case
%  of multiple roots, only a single root will be given.]
function [root]=Newton(fstr,fprimestr)            ! line 1
format long;                                      ! line 2
% Default function and values if no input
if nargin<1,
   help Newton.m;
   fstr='x.^5-3.1415';   fprimestr='5*x.^4';      ! line 5
end
% Tolerance (to the tenth decimal place)
   delta=10^(-10);
% Converting fstr & fprime to functions f(x) & f'(x)
str=strcat('(',fstr); str=strcat(str,')/(');
str=strcat(str,fprimestr); str=strcat(str,')');
fdivfp=inline(str);                               ! line 10
% Initial random guess
```

```
xn=randn;    deltax=1;
% Iteration until the prescribed accuracy
while (deltax>delta)
    root=xn-fdivfp(xn);
    deltax=abs(root-xn);
    xn=root;                                        ! line 15
end
disp(strcat(fstr, ' has a root')); root             ! line 17
```

## B.3 Pattern Formation

The nonlinear reaction-diffusion equation (Chapter 13)

$$\frac{\partial u}{\partial t} = D(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) + \gamma u(1-u), \tag{B.5}$$

can be discretised using the finite difference method. We have

$$\frac{u_{i,j}^{(n+1)} - u_{i,j}^{(n)}}{\Delta t} = D[\frac{u_{i+1,j}^{(n)} - 2u_{i-1,j}^{(n)} + u_{i,j}^{(n)}}{(\Delta x)^2} + \frac{u_{i,j+1}^{(n)} - 2u_{i,j}^{(n)} + u_{i,j-1}^{(n)}}{(\Delta y)^2}]$$

$$+\gamma u_{i,j}^{(n)}(1-u_{i,j}^{(n)}). \tag{B.6}$$

Using $\Delta x = \Delta y = \Delta t = 1$, we have

$$u_{i,j}^{(n+1)} = D[u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n)}]$$

$$+(1-4D)u_{i,j}^{(n)} + \gamma u_{i,j}^{(n)}(1-u_{i,j}^{(n)}). \tag{B.7}$$

This scheme is implemented using both Matlab and Octave. This simple program solves this nonlinear reaction-diffusion equation and generates beautiful patterns.

**Pattern_matlab.m**

```
% Pattern formation:  a 15 line matlab program
% PDE form: u_t=D*(u_{xx}+u_{yy})+gamma*q(u)
% where q(u)='u.*(1-u)';
% The solution of this PDE is obtained by the
% finite difference method, assuming dx=dy=dt=1.
% Written by X S Yang (Cambridge University)
% Usage: pattern(200)   or  simply >pattern
% ------------------------------------------------
function pattern(n)                    % line 1
% Input number of time steps
if nargin<1, n=200; end                % line 2
```

```
% ------------------------------------------------
% Initialize parameters
% ---- time=100, D=0.2; gamma=0.5; --------------
time=100; D=0.2; gamma=0.5;            % line 3
% ---- Set initial values of u randomly ---------
u=rand(n,n);  grad=u*0;                % line 4
% Vectorisation/index for u(i,j) and the loop ---
I = 2:n-1; J = 2:n-1;                  % line 5
% ---- Time stepping ----------------------------
for step=1:time,                       % line 6
% Laplace gradient of the equation     % line 7
 grad(I,J)= u(I,J-1)+u(I,J+1)+u(I-1,J)+u(I+1,J);
 u =(1-4*D)*u+D*grad+gamma*u.*(1-u);   % line 8
% ----- Show results ----------------------------
 pcolor(u);  shading interp;           % line 9
% ----- Coloring and showing colorbar -----------
 colorbar; colormap jet;               % line 10
 drawnow;                              % line 11
end                                    % line 12
% ----- Topology of the final surface ----------
surf(u);                               % line 13
shading interp;                        % line 14
view([-25 70]);                        % line 15
```

As there are some differences in visualisation in Matlab and Octave, the octave version is given below:

**Pattern_octave.m**

```
% Pattern formation:  a 15 line matlab program
% PDE form: u_t=D*(u_{xx}+u_{yy})+gamma*q(u)
% where q(u)='u.*(1-u)';
% The solution of this PDE is obtained by the
% finite difference method, assuming dx=dy=dt=1.
% Written by X S Yang (Cambridge University)
% Usage: pattern(200)   or  simply >pattern
% ------------------------------------------------
function pattern_octave(n)                    % line 1
% Input number of time steps
if nargin<1, n=200; end                % line 2
% ------------------------------------------------
% Initialize parameters
% ---- time=100, D=0.2; gamma=0.5; --------------
time=100; D=0.2; gamma=0.5;            % line 3
% ---- Set initial values of u randomly ---------
u=rand(n,n);  grad=u*0;                % line 4
% Vectorisation/index for u(i,j) and the loop ---
I = 2:n-1; J = 2:n-1;                  % line 5
```

```
% ---- Time stepping ----------------------------
for step=1:time,                          % line 6
% Laplace gradient of the equation        % line 7
 grad(I,J)= u(I,J-1)+u(I,J+1)+u(I-1,J)+u(I+1,J);
 u =(1-4*D)*u+D*grad+gamma*u.*(1-u);      % line 8
end
% ----- Show results in colors ------------------
colormap (jet);                           % line 9
pcolor(u);                                % line 10
```

## B.4 Wave Equation

The finite element method for the wave equation in a 1-D case (in Section 12.4.3) is implemented in the following program (Matlab and Octave):

**Fem_wave.m**

```
% Solving the 1-D wave equation using the
% finite element method, implemented in Matlab
% written by X S Yang (Cambridge University)
% PDE form: u_{tt}-c^2 u_{xx}=0;   c=speed=1;
% n=number of nodes, N=time-step
n=100;                                    % line 1
% ----- Initializing various parameters ---------
L=1.0;              % length of domain
speed=1.0;          % wave speed
m=n-1;              % number of elements
time=1;             % total time of simulations
% ----- Time steps and element size -------------
dt=L/(n*speed);  hh=L/m;
N=time/dt;          % Number of time steps
% ----- Preprocessing ---------------------------
% Split the domain into regularly-spaced n nodes
for i=1:n,
           x(i)=(i-1)*L/m;
end                                       % line 10
x(1)=0; x(n)=L;
% ----- Finding the element connectivity --------
% Simple 1-D 2-node elements E(1,:) and E(2,:)
for i=1:m,
   E(1,i)=i;  E(2,i)=i+1;
   h(i)=abs(x(E(2,i))-x(E(1,i)));
end                                       % line 15
% ----- Initialization of arrays/matrices --------
u=zeros(1,n)'; f=zeros(1,n)';
```

```
K=zeros(n,n);  M=zeros(n,n);
% ----- Element-by-element assembly --------------
% M d^2U/dt^2+KU=f;
for i=1:m,
   K(i,i)=K(i,i)+1/h(i);
   K(i,i+1)=K(i,i+1)-1/h(i);           % line 20
   K(i+1,i)=K(i+1,i)-1/h(i);
   K(i+1,i+1)=K(i+1,i+1)+1/h(i);
end
% ----- Application of boundary conditions ------
% Fixed boundary  at both ends: u(0)=u(1)=0
K(n,n)=1; K(n,n-1)=0;
K(1,1)=1; K(1,2)=0;                    % line 25
% ----- General mass matrix M -------------------
for i=2:n-1,
    M(i,i)=h(i);
end;
M(1,1)=hh/2;
M(n,n)=hh/2;                           % line 30
Minv=inv(M);
% ----- Preparing time stepping -----------------
D=2*eye(n,n)-dt*dt*Minv*K;
% ----- Initial waveforms with two peaks --------
u0=exp(-(40*(x-L/2)).^2)+0.5*exp(-(40*(x-L/4)).^2);
v=u0'; U=v;
% ---- Solving the matrix equation --------------
% ---- Start time-stepping ----------------------
for t=1:N,
        u=D*U-v;                       % line 35
        v=U;         % stored to be used later
        U=u;         % stored as previous values
% ----- Display the travelling wave -------------
    plot(x,u,x,u0); axis([0 L -1 1]);
    drawnow;
end                                    % line 40
```

It works in both Matlab and Octave and generates an animation of travelling waves.

# References

Matlab info, http://www.mathworks.com

Octave info, http://www.octave.org

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

# Bibliography


[1] Abramowitz M. and Stegun I. A., *Handbook of Mathematical Functions*, Dover Publication, (1965).

[2] Adamatzky A. and Teuscher C., *From Utopian to Genuine Unconventional Computers*, Luniver Press, (2006).

[3] Arfken G., *Mathematical Methods for Physicists*, Academic Press, (1985).

[4] Armstrong M., *Basic Linear Geostatistics*, Springer (1998).

[5] Ashby M. F. and Jones D. R., *Engineering Materials*, Pergamon Press, (1980).

[6] Atluri S. N., *Methods of Computer Modeling in Engineering and the Sciences*, Vol. I, Tech Science Press, (2005).

[7] Bathe K. J., *Finite Element Procedures in Engineering Analysis*, Prentice-Hall, (1982).

[8] Berger A. L., Long term variations of the Earth's orbital elements, *Celestial Mechanics*, **15**, 53-74 (1977).

[9] Biot M. A., General theory of three-dimensional consolidation, *J. Appl. Phys.*, **12**, 155-164 (1941).

[10] Carrrier G. F. and Pearson C. E., *Partial Differential Equations: Theory and Technique*, 2nd Edition, Academic Press, (1988).

[11] Carslaw H. S. and Jaeger J. C., *Conduction of Heat in Solids*, 2nd Ed, Oxford University Press, (1986).

[12] Courant R. and Hilbert, D., *Methods of Mathematical Physics*, 2 volumes, Wiley-Interscience, New York, (1962).

[13] Crank J., *Mathematics of Diffusion*, Clarendon Press, Oxford, (1970).

[14] Cook R. D., *Finite Element Modelling For Stress Analysis*, Wiley & Sons, (1995).

[15] Das B. M., *Advanced Soil Mechanics*, MicGraw-Hill, New York, (1983).

[16] Devaney R. L., *An Introduction to Chaotic Dynamical Systems*, Redwood, (1989).

[17] Drew D. A., Mathematical modelling of two-phase flow, *A. Rev. Fluid Mech.*, **15**, 261-291 (1983).

[18] Fenner R. T., *Engineering Elasticity*, Ellis Horwood Ltd, (1986).

[19] Farlow S. J., *Partial Differential Equations for Scientists and Engineers*, Dover Publications, (1993).

[20] Fletcher, C. A. J. and Fletcher C. A., *Computational Techniques for Fluid Dynamics*, Vol. I, Springer-Verlag, GmbH, (1997).

[21] Forsyth A. R., *Calculus of Variations*, New York, Dover (1960).

[22] Fowler A. C., A mathematical model of magma transport in the asthenosphere, *Geophys. Astrophys. Fluid Dyn.*, **33**, 63-96 (1985).

[23] Fowler A. C., *Mathematical Models in the Applied Sciences*, Cambridge University Press, (1997).

[24] Fowler A. C. and Yang X. S., Fast and slow compaction in sedimentary basins, *SIAM J. Appl. Math.*, **59**, 365-385 (1998).

[25] Fowler A. C. and Yang X. S., Pressure solution and viscous compaction in sedimentary basins, *J. Geophys. Res.*, **104**, 12989-12997 (1999).

[26] Fowler A. C. and Yang X. S., Dissolution-precipitation mechanism for diagenesis in sedimentary basins, *J. Geophys. Res.*, **108**, 10.1029/2002JB002269 (2003).

[27] Gardiner C. W., *Handbook of Stochastic Methods*, Springer, (2004).

[28] Gershenfeld N., *The Nature of Mathematical Modeling*, Cambridge University Press, (1998).

[29] Gill P. E., Murray W., and Wright M. H., *Practical optimisation*, Academic Press Inc, (1981).

[30] Gleick J., *Chaos: Making a New Science*, Penguin, (1988).

[31] Goodman R., *Teach Yourself Statistics*, London, (1957).

[32] Greenkorn R. A., *Flow phenomena in porous media*, Marcel Dekker Inc, (1983).

[33] Hinch E. J., *Perturbation Methods*, Cambridge Univ. Press, (1991).

[34] Happel J. and Brenner H., *Low Reynolds Number Hydrodynamics*, Nijhoff, Dordrecht, (1983).

[35] Jeffrey A., *Advanced Engineering Mathematics*, Academic Press, (2002).

[36] John F., *Partial Differential Equations*, Springer, New York, (1971).

[37] Kant T., *Finite Elements in Computational Mechanics*, Vols. I/II, Pergamon Press, Oxford, (1985).

[38] Kardestruncer H. and Norrie D. H., *Finite Element Handbook*, McGraw-Hill, (1987).

[39] Keener J. and Sneyd J., *A Mathematical Physiology*, Springer-Verlag, New York, (2001).

[40] Kitanidis P. K., *Introduction to Geostatistics*, Cambridge University Press, (1997).

[41] Korn G. A. and Korn T. M., *Mathematical Handbook for Scientists and Engineers*, Dover Publication, (1961).

[42] Korn R. W., Pattern formation in the leaf of zebra grass, *J. Theor. Biol.*, **187**, 449-451 (1997).

[43] Kreyszig E., *Advanced Engineering Mathematics*, 6th Edition, Wiley & Sons, New York, (1988).

[44] Krige D. G., A statistical approach to some basic mine valuation problems on the Witwatersrand, *J. of the Chem., Metal. and Mining Soc. of South Africa*, **52**, 119-139 (1951).

[45] Kuhn H. W. and Tucker A. W., Nonlinear programming, *Proc. 2nd Berkeley Symposium*, pp. 481-492, University of California Press, (1951).

[46] Langtangen H. P., *Computational Partial Differential Equations: Numerical Methods and Diffpack Programming*, Springer, (1999).

[47] LeVeque R. J., *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, (2002).

[48] Lewis R. W., Morgan K., Thomas H. and Seetharamu S. K., *The Finite Element Method in Heat Transfer Analysis*, Wiley & Sons, (1996).

[49] Lewis R. W., Gethin D. T., Yang X. S. and Rowe R. C., A combined finite-discrete element method for simulating pharmaceutical powder tableting, *Int. J. Num. Meth. Eng.*, **62**, 853-869 (2005).

[50] Matlab info, http://www.mathworks.com

[51] Meinhardt H., *The Algorithmic Beauty of Sea Shells*, Springer-Verlag, New York, (1995).

[52] Milankovitch M., Canon of isolation and the ice-age problem, *K. Serb. Adad. Geogr.*, special publication, **132**, 484 (1941).

[53] Mitchell A. R. and Griffiths D. F., *Finite Difference Method in Partial Differential Equations*, Wiley & Sons, New York, (1980).

[54] Moler C. B., *Numerical Computing with MATLAB*, SIAM, (2004).

[55] Murch B. W. and Skinner B. J., *Geology Today - Understanding Our Planet*, John Wiley & Sons, (2001).

[56] Murray J. D., *Mathematical Biology*, Springer-Verlag, New York, (1998).

[57] Ockendon J., Howison S., Lacey A. and Movchan A., *Applied Partial Differential Equations*, Oxford University Press, (2003).

[58] Octave info, http://www.octave.org

[59] Pallour J. D. and Meadows D. S., *Complex Variables for Scientists and Engineers*, Macmillan Publishing Co., (1990).

[60] Papoulis A., *Probability and statistics*, Englewood Cliffs, (1990).

[61] Pearson C. E., *Handbook of Applied Mathematics*, 2nd Ed, Van Nostrand Reinhold, New York, (1983).

[62] Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P., *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, (2002).

[63] Puckett E. G. and Colella P., *Finite Difference Methods for Computational Fluid Dynamics*, Cambridge University Press, (2005).

[64] Revil A., Pervasive pressure-solution transfer: a poro-visco-plastic model, *Geophys. Res. Lett.*, **26**, 255-258 (1999).

[65] Revil A., Pervasive pressure solutoin transfer in a quartz sand, *J. Geophys. Res.*, **106**, 8665-8690 (2001).

[66] Riley K. F., Hobson M. P. and Bence S. J., *Mathematical Methods for Physics and Engineering*, 3rd Edition, Cambridge University Press (2006).

[67] Ross S., *A first Course in Probability*, 5th Edition, Prentice-Hall, (1998).

[68] Ryan M. P., *Magma Transport and Storage*, Wiley & Sons, (1991).

[69] Sawaragi Y., Nakayama H. and Tanino T., *Theory of Multiobjective Optimisation*, Academic Press, (1985).

[70] Selby S. M., *Standard Mathematical Tables*, CRC Press, (1974).

[71] Shi G. R., *Numerical Methods for Petroliferous Basin Modeling*, Petroleum Industry Press, Beijing (2000).

[72] Smith D. R., *Variation Methods in Optimization*, New York, Dover, (1998).

[73] Smith G. D., *Numerical Solutions of Partial Differential Equations: Finite Differerence Methods*, 3rd ed., Clarendon Press, Oxford, (1985).

[74] Smith G. N., *Elements of Soil Mechanics for Civil and Mining Engineers*, 5th Edition, Granada Publishing,(1982).

[75] Strang G. and Fix G. J., *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, (1973).

[76] Thomee V., *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, (1997).

[77] Turcotte D. L. and Schubert G., *Geodynamics: Application of Continuum Physics to Geological Problems*, Wiley, New York (1982).

[78] Wang H. F., *Theory of Linear Poroelasticity: with applications to geomechanics and hydrogeology*, Princeton University Press, (2000).

[79] Weisstein E. W., http://mathworld.wolfram.com

[80] Weinstock R., *Calculus of Variations: with applications to Physics and Engineering*, New York, Dover, (1974).

[81] Wikipedia, http://en.wikipedia.com

[82] Wylie C. R., *Advanced Engineering Mathematics*, Tokyo, (1972).

[83] Versteeg H. K. and Malalasekra W., *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*, Prentice Hall, (1995).

[84] Yang X. S., Nonlinear viscoelastic compaction in sedimentary basins, *Nonlinear Processes in Geophysics*, **7**, 1-7 (2000).

[85] Yang X. S., A mathematical model for Voigt poro-visco-plastic deformation, *Geophys. Res. Lett.*, **29**, 1066-1070 (2002).

[86] Yang X. S., *Theoretical Basin Modelling*, Exposure Publishing, (2006).

[87] Yang X. S. and Young Y., Cellular automata, PDEs and pattern formation (Chapter 18), in *Handbook of Bioinspired Algorithms*, edited by Olarius S. and Zomaya A., Chapman & Hall/CRC, (2005).

[88] Yang X. S., *Introduction to Mathematical Optimization: From Linear Programming to Metaheuristics*, Cambridge International Science Publishing (CISP), (2008).

[89] Zienkiewicz O. C. and Taylor R. L., *The Finite Element Method*, vol. I/II, McGraw-Hill, 4th Edition, (1991).

# Index

# Mathematical Modelling for Earth Sciences

**Dr Xin-She Yang**

Mathematical modelling and computer simulations are an essential part of the analytical toolset used by earth scientists. Computer simulations based on mathematical models are routinely used to study geophysical, environmental and geological processes in many areas of work and research from geophysics to petroleum engineering and from hydrology to environmental fluid dynamics.

Dr Yang has carefully selected topics which will be of most value to students and has recognised the need to be careful in his examples whilst being comprehensive enough to include important topics and popular algorithms. The book is designed to be 'theorem-free' and yet to balance formality and practicality. Using worked examples and tackling each problem in a step-by-step manner the text is especially suitable for non-mathematicians approaching this aspect of earth sciences for the first time, The coverage and level, for instance in the calculus of variation and pattern formation, that even mathematicians will find the examples interesting.

Topics covered include: vector and matrix analysis ● ordinary differential equations ● partial differential equations ● calculus of variations ● integral equations ● probability ● geostatistics ● numerical integration ● optimisation ● finite difference methods ● finite volume methods ● finite element methods ● reaction-diffusion system ● elasticity ● fracture mechanics ● poroelasticity, and flows in porous media.

*Mathematical Modelling for Earth Sciences* introduces a wide range of mathematical modelling and numerical techniques, and is written for undergraduates and graduate students.

Xin-She Yang received his DPhil in applied mathematics from the University of Oxford, he is currently a research fellow at the University of Cambridge. He is in the Structure Group in the Civil, Structural and Environmental Engineering Division.

ISBN 9781903765920

DUNEDIN