Mathematical Modeling 數學建模

Ching-hsiao Arthur Cheng 鄭經斅

Chapter 0

Introduction

What is mathematical modeling (or simply modeling)? **Modeling is a process that uses math to represent, analyze, make predictions, or otherwise provide insight into real-world phenomena.**

1. Define the problem statement:
   (a) A concise statement of the problem will tell you what your model will measure or predict.
   (b) Focus and define subjective words (so that they are quantifiable)
   (c) Explore with research and brainstorming.
   (d) Brainstorm like you have access to any and all data.
   (e) Assign a team member to record every idea.
   (f) Visual diagrams can be a powerful tool to help structure.
   (g) Keep an open mind.
2. Making assumptions: After defining the problem statement, you probably will find that your problem is still too complicated. Sharpen your focus by making assumptions. These basic conjectures allow you to reduce the number of factors affecting your model helping you decide what is important.
   (a) Assumptions come from brainstorming.
   (b) Preliminary research will help you make assumptions.
   (c) In the absence of relevant data, it is reasonable to make (and justify) your assumptions.
   (d) Assumptions develop as you move through the modeling process.
3. Defining variables: The variables you need to develop your solution come from the perspective of the problem statement. Dependent variables are often called outputs that represent the information you seek. Independent variables, also known as inputs,

represent quantities you know the value of but may change. Fixed model parameters represent constants that remain the same.

   (a) Your problem statement will define the output.

   (b) Initial brainstorming should give clues to independent, dependent, and fixed model parameters.

   (c) Keep track of the units of measurement you are using (because they can reveal relationship between variables - dimensional analysis)

   (d) You may need to do additional research or make new assumptions to find values of parameters.

   (e) Sub models or multiple models may be needed to reveal certain model input.

4. Getting a solution: use any math tools and softwares to find a answer to the model proposed in the previous steps.

5. Analysis: When one gets a solution of a proposed model, one needs to check the following:

   (a) Is the magnitude of the answer reasonable?

   (b) Does the model behave as expected?

   (c) Can one validate the model?

   You may also determine if the model is acceptable by doing the following:

   (a) List the model's strengths and weaknesses/limitations.

   (b) Determine your model's sensitivity to parameters and assumptions.

   (c) Consider potential improvements.

Chapter 1

Dimensional Analysis（量綱/因次分析）

One of the basic techniques useful in the early stage of modeling problems is the analysis of the relevant quantities and how they relate to each other in a *dimensional way*. The relationship among the variables must have ***dimensional homogeneity*** which simply says that variables with different dimensions cannot be identical (or in short, apples cannot equal oranges). These observations form the basis of the subject known as ***dimensional analysis***.

物理量的量綱可以用來分析或檢核幾個物理量之間的關係，這方法稱為量綱分析（dimensional analysis）。通常，一個物理量的量綱是由像質量、長度、時間、電荷量、溫度一類的基礎物理量綱結合而成。例如，速度的量綱為長度每單位時間，而計量單位為公尺每秒、英哩每小時或其它單位。量綱分析所根據的重要原理是，物理定律必需跟其計量物理量的單位無關。任何有意義的方程式，其左手邊與右手邊的量綱必需相同。檢查有否遵循這規則是做量綱分析最基本的步驟。

Remark 1.1. We distinguish the word ***unit***（單位）from the word ***dimension***（量綱/因次）. By units we mean specific physical units like seconds, hours, days, and years; all of these units have dimensions of time. Similarly, grams, kilograms, pounds, and so on are units of the dimension mass.

注意到術語「量綱」比尺度「單位」更抽象：質量是一種量綱，而公斤是量綱為質量的一種尺度單位。對於每一種量綱，不同的標準制會規定不同的單位。物理量速度的量綱是長度／時間（$\frac{L}{T}$ 或 LT^{-1}），物理量作用力的量綱是質量 × 長度／(時間的平方)（$\frac{ML}{T^2}$ or MLT^{-2}）。原則而言，其它種物理量的量綱也可以定義為基礎量綱，可以替換上述幾個量綱。例如，動量、能量或電流都可以選為基礎量綱。

有些物理學者不認為溫度是基礎量綱，因為溫度表達為粒子的能量每自由度，這可以以能量（或質量、長度、時間）來表達。有些物理學者不認為電荷量是基礎量綱；在厘米-克-秒（cgs）制內，電荷量可以以質量、長度、時間共同結合在一起來表達。另外，還有一些物理學者懷疑，大自然存在著具有不相容基礎量綱的物理量。

For a given physical quantity q, we use $[q]$ to denote the dimension of q, and use L, M, T to denote the dimension of length, mass, and time, respectively. A quantity which does not change after changing unit of every fundamental dimension is called dimensionless.

1.1 Dimensional Methods

The cornerstone result in dimensional analysis is known as the ***Pi theorem*** which states that if there is a physical law which provides a relation among several dimensioned physical quantities, then there is an equivalent law that can be expressed as a relation among certain dimensionless quantities.

Question: What does it mean by a relation among several dimensioned physical quantities?

Example 1.2. The air resistance F a biker encounters appears to be related to the speed v and the cross-sectional area A, as well as the air density ρ. Therefore,

$$F = \phi(\rho, A, v)$$

or equivalently,

$$f(F, \rho, A, v) = F - \phi(\rho, A, v) = 0\,.$$

Example 1.3. Suppose that we want to compute the yield of the first atomic explosion after viewing photographs of the spread of the fireball. In such an explosion a large amount of energy E is released in a short time in a region small enough to be considered a point. From the center of the explosion a strong shock wave spreads outwards; the pressure behind the shock is on the order of hundreds of thousands of atmospheres, far greater than the ambient air pressure whose magnitude can be accordingly neglected in the early stages of the explosion. It is plausible that there is a relation between the radius of the blast wave front r, time t, the initial air density ρ, and the energy released E. Hence, we assume there is a physical law

$$f(t, r, \rho, E) = 0$$

which provides a relationship among these quantities.

Suppose that m quantities $q_1, q_2, \cdots, q_m$ are dimensioned quantities that are expressed in terms of certain selected fundamental dimensions $L_1, L_2, \cdots, L_n$, where $n < m$. The dimensions of q_i, denoted by $[q_i]$, can be written in terms of the fundamental dimensions as

$$[q_i] = L_1^{a_{1i}} L_2^{a_{2i}} \cdots L_n^{a_{ni}}$$

for some exponents $a_{1i}, a_{2i}, \cdots, a_{ni}$. If $[q_i] = 1$, then q_i is said to be ***dimensionless***. The $n \times m$ matrix

$$\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \tag{1.1}$$

containing the exponents is called the ***dimension matrix***. The entries in the i-th column give the exponents for q_i in terms of the powers of $L_1, \cdots, L_n$.

Any fundamental dimension L_i has the property that its units can be changed upon multiplication by the appropriate conversion factor $\lambda_i > 0$ to obtain $\bar{L}_i$ in a new system of units. We write $\bar{L}_i = \lambda_i L_i$. The units of derived quantities q can be changed in a similar fashion. If

$$[q] = L_1^{b_1} L_2^{b_2} \cdots L_n^{b_n} , \tag{1.2}$$

then

$$\bar{q} = \lambda_1^{b_1} \lambda_2^{b_2} \cdots \lambda_n^{b_n} q \tag{1.3}$$

gives q in the new system of units.

Definition 1.4. Let $q_1, q_2, \cdots, q_m$ be dimensioned quantities. The physical law

$$f(q_1, q_2, \cdots, q_m) = 0 \tag{1.4}$$

is said to be ***unit free*** if for all choices of real numbers $\lambda_1, \cdots, \lambda_n$ with $\lambda_i > 0$ for all $1 \leqslant i \leqslant n$, we have $f(q_1, \cdots, q_m) = 0$ if and only if $f(\bar{q}_1, \cdots, \bar{q}_m) = 0$, where q_i and $\bar{q}_i$ are related by (1.3) if q_i obeys (1.2).

Theorem 1.5 (Pi Theorem)**.** *Let*

$$f(q_1, q_2, \cdots, q_m) = 0 \tag{1.5}$$

be a unit free physical law that relates the dimensioned quantities $q_1, q_2, \cdots, q_m$. Let L_1, L_2, $\cdots$, L_n, where $n < m$, be fundamental dimensions with

$$[q_i] = L_1^{a_{1i}} L_2^{a_{2i}} \cdots L_n^{a_{ni}} , \qquad i = 1, \cdots, m ,$$

and let $r = rank(D)$, where D is the dimension matrix given by (1.1)*. Then there exist $(m-r)$ independent dimensionless quantities $\pi_1, \pi_2, \cdots, \pi_{m-r}$ that can be formed from $q_1, \cdots, q_m$ and the physical law* (1.5) *is equivalent to an equation*

$$F(\pi_1, \cdots, \pi_{m-r}) = 0$$

expressed only in terms of the dimensionless quantities.

Proof. Let $D = [a_{ij}]_{n \times m}$ be the dimension matrix and $\pi = q_1^{\alpha_1} q_2^{\alpha_2} \cdots q_m^{\alpha_m}$ be a dimensionless quantities. Then with $\boldsymbol{\alpha}$ denoting the vector $(\alpha_1, \cdots, \alpha_m)^{\mathrm{T}}$, we have

$$D\boldsymbol{\alpha} = \mathbf{0} ,$$

where $\mathbf{0}$ denotes the zero vector in $\mathbb{R}^n$. Since $\mathrm{rank}(D) = r$, without loss of generality we can assume that the first r column of D is linearly independent; thus $\alpha_1, \cdots, \alpha_r$ can be expressed in terms of $(\alpha_{r+1}, \alpha_{r+2}, \cdots, \alpha_m)$. In fact,

$$D(:, 1:r)\boldsymbol{\alpha}(1:r) = -D(:, r+1:m)\boldsymbol{\alpha}(r+1:m) ,$$

where $D(:, i:j)$ denotes the matrix formed by the i-th to j-th columns of D, and $\boldsymbol{\alpha}(i:j)$ denotes the (column) vector formed by the i-th to j-th components of $\boldsymbol{\alpha}$. Assume that the vector $\boldsymbol{\alpha}(1:r)$ is given by

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1(m-r)} \\ b_{21} & b_{22} & \cdots & b_{2(m-r)} \\ \vdots & & & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{r(m-r)} \end{bmatrix} \begin{bmatrix} \alpha_{r+1} \\ \alpha_{r+2} \\ \vdots \\ \alpha_m \end{bmatrix}.$$

Then π_j, $1 \leqslant j \leqslant m-r$, defined by (with $\alpha_{r+\ell} = \delta_{\ell j}$ for $1 \leqslant \ell \leqslant m-r$)

$$\pi_j = q_1^{b_{1j}} q_2^{b_{2j}} \cdots q_r^{b_{rj}} q_{r+j}$$

are dimensionless quantities (so change of units will not change the value of π_j). Define

$$\begin{aligned} &G(q_1, \cdots, q_r, \pi_1, \cdots, \pi_{m-r}) \\ &\quad = f(q_1, q_2, \cdots, q_r, \pi_1 q_1^{-b_{11}} \cdots q_r^{-b_{r1}}, \pi_2 q_1^{-b_{12}} \cdots q_r^{-b_{r2}}, \cdots, \pi_{m-r} q_1^{-b_{1(m-r)}} \cdots q_r^{-b_{r(m-r)}}). \end{aligned}$$

Then $G(q_1, \cdots, q_r, \pi_1, \cdots, \pi_{m-r}) = 0$ if and only if $f(q_1, \cdots, q_m) = 0$. Moreover, since $f(q_1, q_2, \cdots, q_m) = 0$ is unit free, $G(q_1, \cdots, q_r, \pi_1, \cdots, \pi_{m-r}) = 0$ is unit free.

Now, since $G(q_1, \cdots, q_r, \pi_1, \cdots, \pi_{m-r}) = 0$ is unit free, for any choice of conversion factors $\lambda_1, \cdots, \lambda_n > 0$ and

$$\bar{q}_j = \lambda_1^{a_{1j}} \lambda_2^{a_{2j}} \cdots \lambda_n^{a_{nj}} q_j, \quad 1 \leqslant j \leqslant r,$$

we must have $G(\bar{q}_1, \cdots, \bar{q}_r, \pi_1, \cdots, \pi_{m-r}) = 0$. Since $D(:, 1:r)$ consists of r linearly independent column vectors and $n \geqslant r$, there exist $\lambda_1, \cdots, \lambda_n$ (might not be unique if $n > r$) such that

$$\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{nr} \end{bmatrix} \begin{bmatrix} \log \lambda_1 \\ \log \lambda_2 \\ \vdots \\ \log \lambda_n \end{bmatrix} = \begin{bmatrix} -\log q_1 \\ -\log q_2 \\ \vdots \\ -\log q_r \end{bmatrix} \tag{1.6}$$

Choosing $\lambda_1, \cdots, \lambda_n$ satisfying (1.6). Then in the new system of units $\bar{q}_j = 1$; thus in the new system of units,

$$F(\pi_1, \cdots, \pi_{m-r}) \equiv G(1, \cdots, 1, \pi_1, \cdots, \pi_{m-r}) = 0. \qquad \square$$

Example 1.6 (Example 1.2 - revisit)**.** Since

$$[F] = MLT^{-2}, \quad [\rho] = ML^{-3}, \quad [A] = L^2, \quad v = LT^{-1},$$

the dimension matrix (with the order of dimension T, L, M) is

$$\begin{bmatrix} -2 & 0 & 0 & -1 \\ 1 & -3 & 2 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

The rank of the dimension matrix above is 3; thus there is only one dimensionless quantity that can be formed from F, ρ, A, v. Suppose that $\pi = F^{\alpha_1}\rho^{\alpha_2}A^{\alpha_3}v^{\alpha_4}$ is a dimensionless quantity. Then

$$\begin{bmatrix} -2 & 0 & 0 & -1 \\ 1 & -3 & 2 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

which gives a dimensionless quantity $\pi = F\rho^{-1}A^{-1}v^{-2}$. Therefore, an equivalent physical law is given by $g(\pi) = 0$ which shows that $\pi = k$ (or equivalently, $F = k\rho A v^2$) for some (dimensionless) constant k.

Example 1.7 (Example 1.3 - revisit)**.** Since

$$[t] = T\,, \quad [r] = L\,, \quad [\rho] = ML^{-3}\,, \quad E = ML^2T^{-2}\,,$$

the dimension matrix (with the order of dimension T, L, M) is

$$\begin{bmatrix} 1 & 0 & 0 & -2 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

The rank of the dimension matrix above is clearly 3; thus there is only one dimensionless quantity that can be formed from r, r, ρ, E. Suppose that $\pi = t^{\alpha_1}r^{\alpha_2}\rho^{\alpha_3}E^{\alpha_4}$ is a dimensionless quantity. Then

$$\begin{bmatrix} 1 & 0 & 0 & -2 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

which gives a dimensionless quantity $\pi = t^2r^{-5}\rho^{-1}E$. Therefore, an equivalent physical law is given by $F(\pi) = 0$ which shows that $\pi = k$ (or equivalently, $t^2E = k\rho r^5$) for some (dimensionless) constant k.

Example 1.8. At time $t = 0$ an amount of heat energy e, concentrated at a point in space, is allowed to diffuse outward into a region with temperature zero. If r denotes the radial distance from the source and t is time, the problem is to determine the temperature u as a function of r and t.

Clearly the temperature u depends on t, r and e. Moreover, it is "reasonable" that the "thermal diffusivity" k with dimension length-squared per time and the "heat capacity" c of the region, with dimension energy per degree per volume, play a role. Therefore, the physical law is given by

$$f(t, r, u, e, k, c) = 0\,.$$

This physical law has 6 dimensioned quantities

$$[t] = T, \quad [r] = L, \quad [u] = \Theta, \quad [e] = E, \quad [k] = L^2T^{-1}, \quad [c] = E\Theta^{-1}L^{-3}\,.$$

The dimension matrix (with the order of dimension T, L, Θ, E) is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 2 & -3 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

It is easy to see that the dimension matrix has rank 4; thus by the Pi theorem there are 2 dimensionless quantities that can be formed from t, r, u, e, c, k. To see how we form dimensionless quantities, we assume that the combination

$$\left[t^{\alpha_1} r^{\alpha_2} u^{\alpha_3} e^{\alpha_4} k^{\alpha_5} c^{\alpha_6}\right] = 1\,.$$

In other words,

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 2 & -3 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

which shows that $\alpha_1 = \alpha_5$, $\alpha_3 = -\alpha_4 = \alpha_6$, and $\alpha_2 = -2\alpha_5 + 3\alpha_6$. Therefore, two dimensionless quantities can be formed $\big($using $(\alpha_5, \alpha_6) = \big(-\frac{1}{2}, 0\big)$ or $\big(\frac{3}{2}, 1\big)\big)$ as

$$\pi_1 = \frac{r}{\sqrt{kt}} \qquad \text{and} \qquad \pi_2 = \frac{uc}{e}(kt)^{\frac{3}{2}}$$

and an equivalent physical law is given by $F(\pi_1, \pi_2) = 0$ which "implies" that $\pi_2 = g(\pi_1)$ for some function g. Therefore, the temperature u can be expressed by

$$u = \frac{e}{c(kt)^{\frac{3}{2}}} g\big(\frac{r}{\sqrt{kt}}\big)\,.$$

Example 1.9. Suppose that at time $t = 0$ an object of mass m is given a vertical upward velocity V from the surface of a spherical planet (with mass M and radius R). The height h of the object is a function of t that obeys

$$m\frac{d^2h}{dt^2} = -\frac{GMm}{(R+h)^2}\,.$$

The gravitational acceleration g on the surface of the planet is given by $g = \dfrac{GM}{R^2}$; thus including the *initial data*,

$$\frac{d^2h}{dt^2} = -\frac{R^2 g}{(R+h)^2}\,, \qquad h(0) = 0\,, \quad h'(0) = V\,. \tag{1.7}$$

The physical law of the system above can be written as

$$f(t, h, R, V, g) = 0\,,$$

where the five dimensioned quantities have dimension

$$[t] = T\,, \quad [h] = L\,, \quad [R] = L\,, \quad [V] = LT^{-1} \quad \text{and} \quad [g] = LT^{-2}\,,$$

and the dimension matrix (with the order of dimension T, L) is given by

$$\begin{bmatrix} 1 & 0 & 0 & -1 & -2 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

If $\pi = t^{\alpha_1} h^{\alpha_2} R^{\alpha_3} V^{\alpha_4} g^{\alpha_5}$ is a dimensionless quantity, then

$$\begin{bmatrix} 1 & 0 & 0 & -1 & -2 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

or equivalently, $\alpha_1 = \alpha_4 + 2\alpha_5$ and $\alpha_2 = -(\alpha_3 + \alpha_4 + \alpha_5)$. Since the rank of the dimension matrix is 2 there are three dimensionless quantities that can be formed: we choose $(\alpha_3, \alpha_4, \alpha_5) = (-1, 0, 0), (-1, 1, 0)$ and $\left(-\frac{1}{2}, 1, -\frac{1}{2}\right)$ to form

$$\pi_1 = \frac{h}{R}\,, \qquad \pi_2 = \frac{tV}{R}\,, \qquad \pi_3 = \frac{V}{\sqrt{gR}}\,.$$

Therefore, the Pi theorem "implies" that there exists a function F such that $\pi_1 = F(\pi_2, \pi_3)$ or

$$\frac{h}{R} = F\big(\frac{tV}{R}, \frac{V}{\sqrt{gR}}\big)\,.$$

Suppose that at $t = t_{\max}$ the object reaches its maximum height. Intuitively $t_{\max}$ should depends on three dimensional quantities g, R, V. On the other hand, we have $h'(t_{\max}) = 0$; thus

$$0 = h'(t_{\max}) = R\frac{\partial}{\partial \pi_2}\Big|_{t=t_{\max}} F\big(\frac{tV}{R}, \frac{V}{\sqrt{gR}}\big) = V\frac{\partial F}{\partial \pi_2}\big(\frac{t_{\max}V}{R}, \frac{V}{\sqrt{gR}}\big)\,.$$

The above relation "implies" that $\dfrac{t_{\max}V}{R}$ is a function of $\dfrac{V}{\sqrt{gR}}$; thus

$$\frac{t_{\max}V}{R} = G\big(\frac{V}{\sqrt{gR}}\big)\,.$$

1.2 Characteristic Scales and Scaling

The use of "characteristic scales" helps us reduce mathematical model into dimensionless form.

Example 1.10. Let $p = p(t)$ denote the population of an animal species located in a fixed region at time t. The simplest model of population growth is the classic **Malthus model** which states that the growth rate $\dfrac{dp}{dt}$ is proportional to the population p, or equivalently

$$\frac{dp}{dt} = rp.$$

where r is the growth rate, given in dimensions of inverse-time. A more reasonable model, called the **logistics model**, is given by

$$\frac{dp}{dt} = rp\big(1 - \frac{p}{K}\big)\,,$$

where $K > 0$ is called the *carring capacity* (with dimension of population). Let $\tau = rt$, and $\frac{p}{K} = P$. Then τ and P are dimensionless variables that satisfy

$$\frac{dP}{d\tau} = P(1-P)\,.$$

The above ODE is a relation between two dimensionless quantities.

Suppose that an initial condition $p(0) = p_0$ is imposed on this ODE. Then using P and τ we have the following dimensionless model

$$\frac{dP}{d\tau} = P(1-P)\,, \qquad P(0) = \epsilon\,, \tag{1.8}$$

where $\epsilon = \frac{p_0}{K}$. On the other hand, there is another way of rewriting

$$\frac{dp}{dt} = rp\big(1 - \frac{p}{K}\big)\,, \qquad p(0) = p_0$$

into dimensionless form. Let $\widetilde{P} = \dfrac{P}{p_0}$ and $\tau = rt$. Then we have

$$\frac{d\widetilde{P}}{d\tau} = \widetilde{P}\big(1 - \epsilon\widetilde{P}\big)\,, \qquad \widetilde{P}(0) = 1\,. \tag{1.9}$$

We note that if $\epsilon \ll 1$, we tend to let $\epsilon = 0$ and find that (1.9) provides a more reasonable approximation.

Example 1.11 (Example 1.9 - revisit)**.** In this example we choose characteristic time scale t_c and length scale ℓ_c to recast the ODE (1.7)

$$\frac{d^2h}{dt^2} = -\frac{R^2 g}{(R+h)^2}\,, \qquad h(0) = 0\,, \quad h'(0) = V\,. \tag{1.7}$$

We note that with dimensionless time $\bar{t} = t/t_c$ and dimensionless height $\bar{h} = h/\ell_c$ $\big($so that $\bar{h}(\bar{t}) = \dfrac{h(t_c\bar{t})}{\ell_c}\big)$, ODE (1.7) is equivalent to the dimensionless ODE

$$\frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{t_c^2 g}{\ell_c}\frac{1}{(1+\frac{\ell_c}{R}\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = \frac{t_c V}{\ell_c}\,. \tag{1.10}$$

Three dimensioned quantities in (1.7) are

$$[R] = L\,, \qquad [g] = LT^{-2} \qquad \text{and} \qquad [V] = LT^{-1}\,.$$

Therefore, three relevant time scales are $t_c = R/V$, $t_c = \sqrt{R/g}$ or $t_c = V/g$, and two relevant length scales are $\ell_c = R$ or $\ell_c = V^2/g$.

Define a dimensionless quantity $\epsilon = \dfrac{V^2}{gR}$. Using these characteristic scales, we reach at the following dimensionless problems:

1. Let $t_c = R/V$ and $\ell_c = R$. Then (1.10) implies that

$$\epsilon \frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{1}{(1+\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = 1\,.$$

2. Let $t_c = R/V$ and $\ell_c = V^2/g$. Then (1.10) implies that

$$\epsilon^2 \frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{1}{(1+\epsilon\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = \frac{1}{\epsilon}\,.$$

3. Let $t_c = \sqrt{R/g}$ and $\ell_c = R$. Then (1.10) implies that

$$\frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{1}{(1+\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = \sqrt{\epsilon}\,.$$

4. Let $t_c = \sqrt{R/g}$ and $\ell_c = V^2/g$. Then (1.10) implies that

$$\frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{1}{\epsilon}\frac{1}{(1+\epsilon\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = \frac{1}{\sqrt{\epsilon}}\,.$$

5. Let $t_c = V/g$ and $\ell_c = R$. Then (1.10) implies that

$$\frac{d^2\bar{h}}{d\bar{t}^2} = -\epsilon\frac{1}{(1+\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = \epsilon\,.$$

6. Let $t_c = V/g$ and $\ell_c = V^2/g$. Then (1.10) implies that

$$\frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{1}{(1+\epsilon\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = 1\,.$$

Suppose that $\epsilon \ll 1$; that is, V^2 is much smaller than gR. In such a case, we are tempted to delete the terms involving ϵ (or simply setting $\epsilon = 0$) in the scaled problem. Then only case 3, 5, 6 provide meaningful models; however, only case 6 can provide a reasonable interpretation of the real phenomena. Therefore, one needs to be very careful about choosing characteristic scales.

The reason why $t_c = V/g$ and $\ell_c = V^2/g$ is the correct characteristic scale when $\epsilon \ll 1$?

When the gravity acceleration is always g $\big($instead of $\dfrac{GM}{(R+h)^2}\big)$, the rocket takes V/g time to reach its maximum height $\dfrac{V^2}{2g}$; thus $t_c = \dfrac{V}{g}$ is a good choice of the characteristic time scale and $\ell_c = \dfrac{V^2}{g}$ is a good choice of the characteristic length scale.

Example 1.12. The Navier-Stokes equation (which we will derive much later) is used to described the dynamics of fluids such as the air or liquids. Consider incompressible fluids (which means the density of the fluid under consideration is a constant). Let $\boldsymbol{u}(x_1, x_2, x_3, t) = \big(u_1(x_1, x_2, x_3, t), u_2(x_1, x_2, x_3, t), u_3(x_1, x_2, x_3, t)\big)$ and $p(x_1, x_2, x_3, t)$ denote the velocity and

the pressure of the fluid at point (x_1, x_2, x_3) and time t, respectively. Then $\boldsymbol{u}$ and p obeys a system of PDEs, called the incompressible Navier-Stokes equations:

$$\rho(\boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla_x \boldsymbol{u}) + \nabla_x p = \mu \Delta_x \boldsymbol{u}\,, \tag{1.11a}$$

$$\operatorname{div} \boldsymbol{u} = 0\,, \tag{1.11b}$$

where ρ is the density of the fluid, $\boldsymbol{u}_t$ denotes the partial derivative of $\boldsymbol{u}$ w.r.t. t, $\nabla_x p$ is the gradient of the pressure function p, μ is the dynamical viscosity with dimension of mass per length per time , and

$$\begin{aligned}
\boldsymbol{u} \cdot \nabla_x \boldsymbol{u} &= \sum_{j=1}^{3} u_j \frac{\partial \boldsymbol{u}}{\partial x_j} = u_1 \frac{\partial \boldsymbol{u}}{\partial x_1} + u_2 \frac{\partial \boldsymbol{u}}{\partial x_2} + u_3 \frac{\partial \boldsymbol{u}}{\partial x_3}\,, \\
\Delta_x \boldsymbol{u} &\equiv \sum_{j=1}^{3} \frac{\partial^2 \boldsymbol{u}}{\partial x_j^2} = \frac{\partial^2 \boldsymbol{u}}{\partial x_1^2} + \frac{\partial^2 \boldsymbol{u}}{\partial x_2^2} + \frac{\partial^2 \boldsymbol{u}}{\partial x_3^2}\,, \\
\operatorname{div} \boldsymbol{u} &\equiv \sum_{j=1}^{3} \frac{\partial u_j}{\partial x_j} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}\,.
\end{aligned}$$

Let ℓ_c denote the characteristic length, and u_c denote the characteristic speed (which implies that $t_c = \ell_c / u_c$ is the characteristic time). Define $\tau = \dfrac{t}{t_c}$, $y = \dfrac{x}{\ell_c}$, and

$$\begin{aligned}
\boldsymbol{v}(y_1, y_2, y_3, \tau) &= \frac{\boldsymbol{u}}{u_c}(\ell_c y_1, \ell_c y_2, \ell_c y_3, t_c \tau)\,, \\
q(y_1, y_2, y_3, \tau) &= \frac{p}{u_c^2 \rho}(\ell_c y_1, \ell_c y_2, \ell_c y_3, t_c \tau)\,.
\end{aligned}$$

Then with $\nu = \dfrac{\mu}{\rho}$ denoting the kinetic viscosity, we have

$$\begin{aligned}
\boldsymbol{v}_\tau + \boldsymbol{v} \cdot \nabla_y \boldsymbol{v} + \nabla_y q &= \frac{\nu}{\ell_c u_c} \Delta_y \boldsymbol{v}\,, \\
\operatorname{div}_y \boldsymbol{v} &= 0\,,
\end{aligned}$$

where $\boldsymbol{v} \cdot \nabla_y \boldsymbol{v}$, $\Delta_y \boldsymbol{v}$ and $\operatorname{div}_y \boldsymbol{v}$ are defined similarly. The dimensionless number $\mathrm{Re} \equiv \dfrac{\ell_c u_c}{\nu}$ is called the Reynolds number, and the equations above read

$$\begin{aligned}
\boldsymbol{v}_\tau + \boldsymbol{v} \cdot \nabla_y \boldsymbol{v} + \nabla_y q &= \frac{1}{\mathrm{Re}} \Delta_y \boldsymbol{v}\,, \\
\operatorname{div}_y \boldsymbol{v} &= 0\,.
\end{aligned}$$

1.3 Scaling Arguments

In mathematics there are lots of inequalities that involve comparison of integrals of functions and their derivatives. For example, let $\mathscr{C}_0^1(\mathbb{R})$ denote the collection of all continuously differentiable functions defined on $\mathbb{R}$ that vanish at infinity; that is, if $f \in \mathscr{C}_0^1(\mathbb{R})$, then $f \in \mathscr{C}^1(\mathbb{R})$ and $\lim\limits_{x \to \pm\infty} f(x) = 0$. Then if $f \in \mathscr{C}_0^1(\mathbb{R})$ and $x \in \mathbb{R}$,

$$\int_{-\infty}^{x} f'(t)\, dt = f(x) \qquad \text{and} \qquad \int_{x}^{\infty} f'(t)\, dt = -f(x)\,.$$

Therefore,

$$2|f(x)| \leqslant \int_{-\infty}^{x} \big|f'(x)\big|\,dt + \int_{x}^{\infty} \big|f'(t)\big|\,dt = \int_{-\infty}^{\infty} \big|f'(t)\big|\,dt \qquad \forall\, f \in \mathscr{C}_0^1(\mathbb{R})\,, x \in \mathbb{R}\,.$$

The above inequality then shows that

$$\max_{x\in\mathbb{R}} \big|f(x)\big| \leqslant \frac{1}{2}\int_{-\infty}^{\infty} \big|f'(x)\big|\,dx \qquad \forall\, f \in \mathscr{C}_0^1(\mathbb{R})\,. \tag{1.12}$$

The scaling arguments sometimes is useful to determined what kind of integrals can be compared.

Example 1.13. Suppose that we have the following inequality (which can be thought as a generalization of (1.12))

$$\max_{x\in\mathbb{R}} \big|f(x)\big| \leqslant C\Big(\int_{-\infty}^{\infty} \big|f'(x)\big|^p\,dx\Big)^r \qquad \forall\, f \in \mathscr{C}_0^1(\mathbb{R})\,, \tag{1.13}$$

where C is a constant independent of the choice of f. Find the relation between p, q, r, s.

Let $f \in \mathscr{C}_0^1(\mathbb{R})$ be given. For given constants $M, L > 0$, define

$$u(x) = Mf(Lx)\,.$$

Then clearly $u \in \mathscr{C}_0^1(\mathbb{R})$; thus (1.13) (which is assumed to be valid) implies that

$$\max_{x\in\mathbb{R}} \big|u(x)\big| \leqslant C\Big(\int_{-\infty}^{\infty} \big|u'(x)\big|^p\,dx\Big)^r.$$

Since $\max\limits_{x\in\mathbb{R}} \big|u(x)\big| = M \max\limits_{x\in\mathbb{R}} \big|f(x)\big|$ and the substitution of variables implies that

$$\int_{-\infty}^{\infty} \big|u'(x)\big|^p\,dx = \int_{-\infty}^{\infty} \big|MLf'(Lx)\big|^p\,dx = M^pL^{p-1}\int_{-\infty}^{\infty} \big|f'(x)\big|^p\,dx\,,$$

we have

$$\max_{x\in\mathbb{R}} \big|f(x)\big| \leqslant CM^{pr-1}L^{(p-1)r}\Big(\int_{-\infty}^{\infty} \big|f'(x)\big|^p\,dx\Big)^r.$$

If $pr \neq 1$ or $(p-1)r \neq 0$, we can let M, L approach 0 or ∞ to make the right-hand side approach zero which shows $f \equiv 0$, an impossible situation. Therefore, we must have $pr = 1$ and $(p-1)r = 0$ which implies that $p = r = 1$ is the only possible case for (1.13) to hold.

Example 1.14 (Hölder's inequality)**.** Suppose that one knows that for some $p, q, r, s \in \mathbb{R}$, we have the following inequality

$$\begin{aligned}&\int_{\mathbb{R}^n} \big|f(x_1,\cdots,x_n)g(x_1,\cdots,x_n)\big|\,d(x_1,\cdots,x_n)\\ &\quad\leqslant \Big(\int_{\mathbb{R}^n} \big|f(x_1,\cdots,x_n)\big|^p\,d(x_1,\cdots,x_n)\Big)^r \Big(\int_{\mathbb{R}^n} \big|g(x_1,\cdots,x_n)\big|^q\,d(x_1,\cdots,x_n)\Big)^s\end{aligned} \tag{1.14}$$

for all $f \in L^p(\mathbb{R}^n)$ and $g \in L^q(\mathbb{R}^n)$, where that a function h belongs to class $L^r(\mathbb{R}^n)$ means that $h : \mathbb{R}^n \to \mathbb{R}$ and

$$\int_{\mathbb{R}^n} \big|h(x_1,\cdots,x_n)\big|^r\,d(x_1,\cdots,x_n) < \infty\,.$$

We would like to know the relation between p, q, r, s.

Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be such that $f \in L^p(\mathbb{R}^n)$ and $g \in L^q(\mathbb{R}^n)$. For $M_1, M_2, L > 0$, define

$$u(x_1, \cdots, x_n) = M_1 f(Lx_1, \cdots, Lx_n) \quad \text{and} \quad v(x_1, \cdots, x_n) = M_2 g(Lx_1, \cdots, Lx_n)\,.$$

Then $u, v : \mathbb{R}^n \to \mathbb{R}$. Moreover, the change of variables formula implies that

$$\begin{aligned}
\int_{\mathbb{R}^n} \big|u(x_1, \cdots, x_n)\big|^p \, d(x_1, \cdots, x_n) &= M_1^p L^{-n} \int_{\mathbb{R}^n} \big|f(x_1, \cdots, x_n)\big|^p \, d(x_1, \cdots, x_n)\,, \\
\int_{\mathbb{R}^n} \big|v(x_1, \cdots, x_n)\big|^q \, d(x_1, \cdots, x_n) &= M_2^q L^{-n} \int_{\mathbb{R}^n} \big|g(x_1, \cdots, x_n)\big|^q \, d(x_1, \cdots, x_n)\,;
\end{aligned} \tag{1.15}$$

thus $u \in L^p(\mathbb{R}^n)$ and $v \in L^q(\mathbb{R}^n)$. Since (1.14) is assumed to be known, we must have

$$\begin{aligned}
&\int_{\mathbb{R}^n} \big|u(x_1, \cdots, x_n) v(x_1, \cdots, x_n)\big| \, d(x_1, \cdots, x_n) \\
&\qquad \leqslant \Big(\int_{\mathbb{R}^n} \big|u(x_1, \cdots, x_n)\big|^p \, d(x_1, \cdots, x_n) \Big)^r \Big(\int_{\mathbb{R}^n} \big|v(x_1, \cdots, x_n)\big|^q \, d(x_1, \cdots, x_n) \Big)^s .
\end{aligned}$$

By the fact that

$$\begin{aligned}
&\int_{\mathbb{R}^n} \big|u(x_1, \cdots, x_n) v(x_1, \cdots, x_n)\big| \, d(x_1, \cdots, x_n) \\
&\qquad = M_1 M_2 L^{-n} \int_{\mathbb{R}^n} \big|f(x_1, \cdots, x_n) g(x_1, \cdots, x_n)\big| \, d(x_1, \cdots, x_n)\,,
\end{aligned}$$

(1.15) further implies that

$$\begin{aligned}
&M_1 M_2 L^{-n} \int_{\mathbb{R}^n} \big|f(x_1, \cdots, x_n) g(x_1, \cdots, x_n)\big| \, d(x_1, \cdots, x_n) \\
&\quad = \int_{\mathbb{R}^n} \big|u(x_1, \cdots, x_n) v(x_1, \cdots, x_n)\big| \, d(x_1, \cdots, x_n) \\
&\quad \leqslant \Big(\int_{\mathbb{R}^n} \big|u(x_1, \cdots, x_n)\big|^p \, d(x_1, \cdots, x_n) \Big)^r \Big(\int_{\mathbb{R}^n} \big|v(x_1, \cdots, x_n)\big|^q \, d(x_1, \cdots, x_n) \Big)^s \\
&\quad \leqslant M_1^{pr} M_2^{qs} L^{-nr-ns} \Big(\int_{\mathbb{R}^n} \big|f(x_1, \cdots, x_n)\big|^p \, d(x_1, \cdots, x_n) \Big)^r \times \\
&\qquad\qquad \times \Big(\int_{\mathbb{R}^n} \big|g(x_1, \cdots, x_n)\big|^q \, d(x_1, \cdots, x_n) \Big)^s .
\end{aligned}$$

Therefore, the same reason in Example 1.13 shows that $pr = 1$, $qs = 1$ and $-n = -nr - ns$; thus $r = \dfrac{1}{p}$, $s = \dfrac{1}{q}$ and we have

$$\begin{aligned}
&\int_{\mathbb{R}^n} \big|f(x_1, \cdots, x_n) g(x_1, \cdots, x_n)\big| \, d(x_1, \cdots, x_n) \\
&\qquad \leqslant \int_{\mathbb{R}^n} \big|f(x_1, \cdots, x_n)\big|^p \, d(x_1, \cdots, x_n) \Big)^{\frac{1}{p}} \Big(\int_{\mathbb{R}^n} \big|g(x_1, \cdots, x_n)\big|^q \, d(x_1, \cdots, x_n) \Big)^{\frac{1}{q}} ,
\end{aligned} \tag{1.16}$$

where $\dfrac{1}{p} + \dfrac{1}{q} = 1$.

Remark 1.15. Later on we will simply write $\int_{\mathbb{R}^n} f(x_1,\cdots,x_n)\,d(x_1,\cdots,x_n)$ as $\int_{\mathbb{R}^n} f(x)\,dx$ with $x=(x_1,\cdots,x_n)$ in mind.

Remark 1.16. Inequality (1.16) in fact holds for $1<p,q<\infty$ and $\frac{1}{p}+\frac{1}{q}=1$. In general, suppose that $\Omega\subseteq\mathbb{R}^n$ is a region on which two functions u,v are defined so that $u\in L^p(\Omega)$ and $v\in L^q(\Omega)$ for some $1<p,q<\infty$ and $\frac{1}{p}+\frac{1}{q}=1$, where that a function h belongs to class $L^r(\Omega)$ means that $h:\Omega\to\mathbb{R}$ and

$$\int_\Omega \big|h(x)\big|^r\,dx \equiv \int_\Omega \big|h(x_1,\cdots,x_n)\big|^r\,d(x_1,\cdots,x_n)<\infty\,.$$

Letting $f=\mathbf{1}_\Omega u$ and $g=\mathbf{1}_\Omega v$ in (1.14), where $\mathbf{1}_\Omega$ is the indicator function of Ω given by

$$\mathbf{1}_\Omega(x)=\begin{cases}1 & \text{if } x\in\Omega\,,\\ 0 & \text{otherwise}\,,\end{cases}$$

we find that

$$\int_\Omega \big|u(x)v(x)\big|\,dx \leqslant \Big(\int_\Omega \big|u(x)\big|^p\,dx\Big)^{\frac{1}{p}}\Big(\int_\Omega \big|v(x)\big|^q\,dx\Big)^{\frac{1}{q}}\,. \tag{1.17}$$

The inequality above is called the (general) Hölder inequality.

Example 1.17 (Sobolev's inequalities)**.** The simplest Sobolev's inequalities is of the form

$$\Big(\int_{\mathbb{R}^n}\big|f(x)\big|^q\,dx\Big)^s \leqslant C\Big(\int_{\mathbb{R}^n}\big|(\nabla f)(x)\big|^p\,dx\Big)^r \qquad \forall\, f\in\mathscr{C}^1_c(\mathbb{R}^n)\,, \tag{1.18}$$

where C is a generic constant independent of f, and $\mathscr{C}^1_c(\mathbb{R}^n)$ denotes the collection of continuously differentiable functions that vanish outside certain balls. In this example we determine the relation among n,p,q,r,s.

Let $f:\mathbb{R}^n\to\mathbb{R}$ be such that $f\in\mathscr{C}^1_c(\mathbb{R}^n)$. For given constants $M,L>0$, define $u(x)=Mf(Lx)$. Then $u\in\mathscr{C}^1_c(\mathbb{R}^n)$; thus u also satisfies

$$\Big(\int_{\mathbb{R}^n}\big|u(x)\big|^q\,dx\Big)^s \leqslant C\Big(\int_{\mathbb{R}^n}\big|(\nabla u)(x)\big|^p\,dx\Big)^r\,. \tag{1.19}$$

On the other hand, the change of variables formula implies that

$$\int_{\mathbb{R}^n}\big|u(x)\big|^q\,dx = M^qL^{-n}\int_{\mathbb{R}^n}\big|f(x)\big|^q\,dx\,,\quad \int_{\mathbb{R}^n}\big|(\nabla u)(x)\big|^p\,dx = M^pL^{p-n}\int_{\mathbb{R}^n}\big|(\nabla f)(x)\big|^p\,dx\,;$$

thus (1.19) implies that

$$M^{qs}L^{-ns}\Big(\int_{\mathbb{R}^n}\big|f(x)\big|^q\,dx\Big)^s \leqslant CM^{pr}L^{(p-n)r}\Big(\int_{\mathbb{R}^n}\big|(\nabla f)(x)\big|^p\,dx\Big)^r\,.$$

Since (1.18) holds for all $M,L>0$, we must have $pr=qs$ and $(p-n)r=-ns$. If $pr=qs=\alpha$, we find that (1.19) becomes

$$\Big(\int_{\mathbb{R}^n}\big|u(x)\big|^q\,dx\Big)^{\frac{\alpha}{q}} \leqslant C\Big(\int_{\mathbb{R}^n}\big|(\nabla u)(x)\big|^p\,dx\Big)^{\frac{\alpha}{p}}$$

and n,p,q must satisfy

$$\frac{n}{q}+\frac{p-n}{p}=0 \qquad \Big(\text{or}\quad \frac{1}{q}=\frac{1}{p}-\frac{1}{n}\Big)\,.$$

Chapter 2

Ordinary Differential Equations

Definition 2.1. A differential equation is a mathematical equation that relates some unknown function with its derivatives. The unknown functions in a differential equations are sometimes called ***dependent variables***, and the variables which the derivatives of the unknown functions are taken with respect to are sometimes called the ***independent variables***. A differential equation is called an ***ordinary differential equation*** (ODE) if it contains an unknown function of one independent variable and its derivatives. A differential equation is called a ***partial differential equation*** (PDE) if it contains unknown multi-variable functions and their partial derivatives.

We note that in most of the mathematical ODE models, the independent variable is the time variable t or the spatial variable x.

Definition 2.2. The ***order*** of a differential equation is the order of the highest-order derivatives present in the equation. A differential equation of order 1 is called first order, order 2 second order, etc.

Remark 2.3. It is commonly assumed that an ordinary differential equation of order n

$$F(t, y, y', \cdots, y^{(n-1)}, y^{(n)}) = 0 \qquad \text{(if the independent variable is } t\text{)}$$

can be written as

$$y^{(n)}(t) = f(t, y, y', \cdots, y^{(n-2)}, y^{(n-1)}) .$$

Moreover, given a differential equation above, we can define a vector-valued function $\boldsymbol{z} = \left(y, y', y'', \cdots, y^{(n-1)}\right)^{\mathrm{T}}$ and write the ODE above as

$$\boldsymbol{z}'(t) = \frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \\ f(t, z_1, z_2, \cdots, z_n) \end{bmatrix} = \boldsymbol{f}(t, \boldsymbol{z}) \tag{2.1}$$

which is a first order ODE with a vector-valued unknown.

Definition 2.4. The ordinary differential equation

$$F(t, y, y', \cdots, y^{(n-1)}, y^{(n)}) = y^{(n)}(t) - f(t, y, y', \cdots, y^{(n-2)}, y^{(n-1)}) = 0 \tag{2.2}$$

is said to be ***linear*** if

$$\begin{aligned} &F(t, cy, cy', \cdots, cy^{(n-1)}, cy^{(n)}) - F(t, 0, 0, \cdots, 0) \\ &\quad = c\big[F(t, y, y', \cdots, y^{(n-1)}, y^{(n)}) - F(t, 0, 0, \cdots, 0)\big] \end{aligned} \qquad \forall\, c \in \mathbb{R}\,. \tag{2.3}$$

The ODE (2.2) is said to be ***nonlinear*** if it is not linear.

2.1 Initial Value Problems

Definition 2.5. An ***initial value problem*** (**IVP**) is a (system of) differential equation

$$y^{(n)}(t) = f(t, y, y', \cdots, y^{(n-2)}, y^{(n-1)})\,. \tag{2.4a}$$

equipped with an initial condition

$$y(t_0) = y_0, \quad y'(t_0) = y_1, \quad y''(t_0) = y_2, \quad \cdots \quad y^{(n-1)}(t_0) = y_{n-1}\,, \tag{2.4b}$$

where t_0 is a given point/time, and $y_0, y_1, \cdots, y_{n-1}$ are given numbers. A solution to the IVP (2.4) is a function y defined on an open interval I so that $t_0 \in I$ and (2.4) is satisfied.

Example 2.6. In Example 1.10 we have talked about the Malthus model

$$\frac{dp}{dt} = rp\,, \qquad p(0) = p_0$$

for the growth of population. In this model, the growth rate is assumed to be positive. However, the same differential equation can be used to model the decay of radioactive substance such as plutonium (鈽). If $p(t)$ is the total amount of such kind of substance at time t, the "growth" rate $\dfrac{dp}{dt}$ is proportional to the total amount p, except that the "growth" rate r is negative. In such a case, r is called the decay rate.

The model has linear ODE and usually is called linear model.

Example 2.7 (Spring-mass system with or without Friction)**.** Consider an object of mass m attached to a spring with Hook's constant k. Let $x(t)$ denote the signed distance between the object and the equilibrium position at time t. If there is no friction, by the Newton second law of motion we find that x obeys the ODE

$$m\ddot{x} = -kx\,.$$

When the friction is under consideration, by the fact that the friction is proportional to the velocity, we find that

$$m\ddot{x} = -kx - r\dot{x}\,.$$

If in addition some external force $f(t)$ are exerted on the mass, the model becomes

$$m\ddot{x} = -kx - r\dot{x} + f\,.$$

We note that the ODE above is linear since the function

$$F(t, x, \dot{x}, \ddot{x}) = m\ddot{x} + r\dot{x} + kx - f(t)$$

satisfies (2.3).

If the initial position and the initial velocity of the object is $x(0) = x_0$ and $x'(0) = x_1$, then $x(t)$ satisfies the IVP

$$m\ddot{x} = -kx - r\dot{x} + f\,, \qquad x(0) = x_0\,, \quad x'(0) = v_0\,. \tag{2.5}$$

The ODE in (2.5) is linear.

Example 2.8 (Oscillating pendulum)**.** A simple pendulum consists of a mass m hanging from a string of length L and fixed at a pivot point P. When displaced to an initial angle and released, the pendulum will swing back and forth with periodic motion.

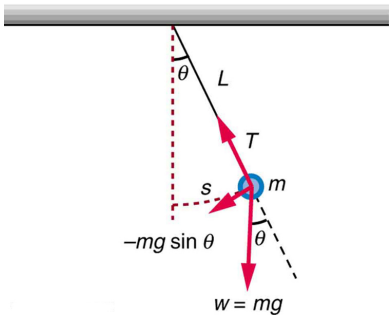


Figure 2.1: A simple pendulum system

Let $\theta(t)$ denote the angle, measured from the vertical dashed line (see figure 2.1), at time t. By Newton's second law,

$$mL\ddot{\theta} = -mg\sin\theta\,, \qquad \theta(0) = \theta_0\,, \quad \theta'(0) = \omega_0\,.$$

The ODE above is a nonlinear ODE.

Example 2.9 (Lotka-Volterra or Prey-Predator model)**.** Suppose that two different species of animals interact within the same environment or ecosystem, and suppose further that the first species eats only vegetation and the second eats only the first species. In other words, one species is a predator (掠食者) and the other is a prey (獵物).

Let $p(t)$ and $q(t)$ denote, respectively, the populations of the prey and the predator. If there is no prey, then the population of the predator should decrease/decay and follows

$$\frac{dq}{dt} = -\beta q\,, \qquad \beta > 0\,.$$

When preys are present in the environment, it seems reasonable that the number of encounters or interactions between these two species per unit time is jointly proportional to their populations p and q; that is, proportional to the product pq. Thus when preys are present, the predator are added to the system at a rate bpq, $b > 0$. In other words, the population of q should follows

$$\frac{dq}{dt} = -\beta q + \delta pq\,, \qquad \beta, \delta > 0\,.$$

On the other hand, if there is no predator, the population of the prey should follow the Malthus model (assuming that the supply of food is always sufficient); however, the population of the prey will decrease by the rate at which the preys are consumed during their encounters with the predator; thus

$$\frac{dp}{dt} = \alpha p - \gamma pq\,, \qquad \alpha, \gamma > 0\,.$$

Therefore, we reach at the ***predator-prey model*** (or the ***Lotka-Volterra model***):

$$\frac{dp}{dt} = \alpha p - \gamma pq = p(\alpha - \gamma q)\,, \tag{2.6a}$$

$$\frac{dq}{dt} = -\beta q + \delta pq = q(-\beta + \delta p)\,. \tag{2.6b}$$

An initial condition $p(0) = p_0$, $q(0) = q_0$ can be imposed so that it becomes an IVP.

The ODE (2.6) is nonlinear since by letting $\boldsymbol{z} = [p, q]^{\mathrm{T}}$, we can write (2.6) as

$$\dot{\boldsymbol{z}} = \boldsymbol{f}(t, \boldsymbol{z}) = \begin{bmatrix} \alpha & 0 \\ 0 & -\beta \end{bmatrix} \boldsymbol{z} + \begin{bmatrix} -\gamma z_1 z_2 \\ \delta z_1 z_2 \end{bmatrix}$$

which shows that $F(t, c\boldsymbol{z}, c\dot{z}) - F(t, 0, 0) \neq c\big[F(t, \boldsymbol{z}, \dot{\boldsymbol{z}}) - F(t, 0, 0)\big]$ if $c \neq 1$.

Example 2.10. Now we consider another spring-mass system in which there are two objects, of mass m_1 and m_2, moving on a frictionless surface under the influence of external forces $F_1(t)$ and $F_2(t)$, and they are also constrained by the three springs whose Hooke's constants are k_1, k_2 and k_3, respectively (see figure 2.2).

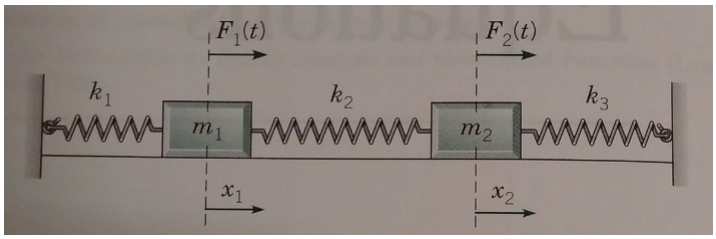


Figure 2.2: A two-mass, three-spring system

Then the equations for the coordinate x_1 and x_2, measured from the equilibrium positions of mass m_1 and m_2 respectively, are given by

$$m_1 \frac{d^2 x_1}{dt^2} = -k_1 x_1 + k_2(x_2 - x_1) + F_1\,, \tag{2.7a}$$

$$m_2 \frac{d^2 x_2}{dt^2} = -k_2(x_2 - x_1) - k_3 x_2 + F_2\,. \tag{2.7b}$$

Reason: Let L_1, L_2, L_3 be the length of the unconstrained springs, and ℓ_1, ℓ_2, ℓ_3 be the increment of the springs in equilibrium. Then

$$k_1\ell_1 = k_2\ell_2 = k_3\ell_3\,. \tag{2.8}$$

Let $x(t)$ and $y(t)$ be the position of mass m_1 and m_2, measured from the left end, respectively. Then $x(t)$ and $y(t)$ satisfy

$$m_1\frac{d^2x}{dt^2} = -k_1(x - L_1) + k_2(y - x - L_2) + F_1\,, \tag{2.9a}$$

$$\begin{aligned} m_2\frac{d^2y}{dt^2} &= -k_2(y - x - L_2) + k_3(L_1 + L_2 + L_3 + \ell_1 + \ell_2 + \ell_3 - y - L_3) + F_2 \\ &= -k_2(y - x - L_2) + k_3(L_1 + L_2 + \ell_1 + \ell_2 + \ell_3 - y) + F_2\,. \end{aligned} \tag{2.9b}$$

Let x_1, x_2 be the position of masses m_1 and m_2 measured from the equilibrium position; that is, $x_1 = x - L_1 - \ell_1$ and $x_2 = y - L_1 - \ell_1 - L_2 - \ell_2$. Then (2.7) follows from using (2.8) in (2.9).

Example 2.11 (Kepler's laws of planetary motion)**.** Kepler's laws of planetary motion describe the motion of planets around the Sun and state that

1. The orbit of a planet is an ellipse with the Sun at one of the two foci.
2. A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.
3. The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

Suppose that planet under consideration is Earth. Since Earth moves on the plane of the ecliptic（黃道面）, we can treat the orbit of Earth as a plane curve. Now we introduce a polar coordinate system and a Cartesian coordinate system on this plane as follows:

1. Let the sun be the pole of the polar coordinate system, and fixed a polar axis on this plane.
2. Let $\boldsymbol{i}$ be the unit vector in the direction of the polar axis, and $\boldsymbol{j}$ be the corresponding unit vector obtained by rotating $\boldsymbol{i}$ counterclockwise by $\dfrac{\pi}{2}$.

Suppose the position of the planet on the planet at time $t \in I$ is given by $\boldsymbol{r}(t) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j}$. For each $t \in I$, let $(r(t), \theta(t))$ be the polar representation of $(x(t), y(t))$ in the trajectory. We would like to determine the relation that $r(t)$ and $\theta(t)$ satisfy.

Define two vectors $\widehat{r}(t) = \cos\theta(t)\boldsymbol{i} + \sin\theta(t)\boldsymbol{j}$ and $\widehat{\theta}(t) = -\sin\theta(t)\boldsymbol{i} + \cos\theta(t)\boldsymbol{j}$. Then $\boldsymbol{r} = r\widehat{r}$. Moreover, let M and m be the mass of the sun and the planet, respectively. Then Newton's second law of motion implies that

$$-\frac{GMm}{r^2}\widehat{r} = m\boldsymbol{r}''\,. \tag{2.10}$$

By the fact that

$$\widehat{r}' = (-\sin\theta, \cos\theta)\theta' = \theta'\widehat{\theta} \qquad \text{and} \qquad \widehat{\theta}' = -(\cos\theta, \sin\theta)\theta' = -\theta'\widehat{r}\,,$$

we find that

$$\begin{aligned}\boldsymbol{r}'' &= \frac{d}{dt}\big(r'\widehat{r} + r\theta'\widehat{\theta}\big) = r''\widehat{r} + r'\theta'\widehat{\theta} + r'\theta'\widehat{\theta} + r\theta''\widehat{\theta} - r(\theta')^2\widehat{r} \\ &= \big[r'' - r(\theta')^2\big]\widehat{r} + \big[2r'\theta' + r\theta''\big]\widehat{\theta}\,.\end{aligned}$$

Therefore, (2.10) implies that

$$-\frac{GM}{r^2}\widehat{r} = \big[r'' - r(\theta')^2\big]\widehat{r} + \big[2r'\theta' + r\theta''\big]\widehat{\theta}\,.$$

Since $\widehat{r}$ and $\widehat{\theta}$ are linearly independent, we must have

$$-\frac{GM}{r^2} = r'' - r(\theta')^2\,, \tag{2.11a}$$

$$2r'\theta' + r\theta'' = 0\,. \tag{2.11b}$$

Note that (2.11b) implies that $(r^2\theta')' = 0$; thus $r^2\theta'$ is a constant. Since $mr^2\theta'$ is the angular momentum, (2.11b) implies that the angular momentum is a constant, so-called the conservation of angular momentum（角動量守恆）.

Let ℓ be the constant angular momentum so that

$$\ell = mr^2\theta'\,. \tag{2.12}$$

Now assume that in each small time interval $J \subseteq I$ of interest, $\theta : J \to \mathbb{R}$ is one-to-one so that the inverse function of θ exists. Then $t = t(\theta)$, and every function of t can be viewed as a function of θ for $t \in J$.

For a function f of t, we let $\dot{f}(\theta)$ denote $\dfrac{d}{d\theta}f(t(\theta))$ and $\ddot{f}(\theta)$ denote $\dfrac{d^2}{d\theta^2}f(t(\theta))$. In other words, $\dot{f}$ denotes the derivative (in θ) of the composite function $f \circ t$. By the chain rule,

$$\frac{d}{dt} = \frac{d\theta}{dt}\frac{d}{d\theta} = \theta'\frac{d}{d\theta} = \frac{\ell}{mr^2}\frac{d}{d\theta} \quad \text{ or equivalently, } \quad f' = \frac{\ell}{mr^2}\dot{f}\,;$$

thus $r' = \dfrac{\ell}{m}\dfrac{\dot{r}}{r^2}$. Let $u = \dfrac{1}{r}$. Then $\dot{u} = -\dfrac{\dot{r}}{r^2}$ which implies that $r' = -\dfrac{\ell}{m}\dot{u}$. Therefore,

$$r'' = -\frac{\ell^2}{m^2r^2}\ddot{u} = -\frac{\ell^2}{m^2}\ddot{u}u^2\,;$$

thus (2.11a) and (2.12) together show that

$$-GMu^2 = -\frac{\ell^2}{m^2}\ddot{u}u^2 - r\Big(\frac{\ell}{mr^2}\Big)^2 = -\frac{\ell^2}{m^2}\ddot{u}u^2 - \frac{\ell^2}{m^2}u^3\,.$$

or equivalently,

$$\ddot{u} + u = \frac{GMm^2}{\ell^2}\,.$$

The general solution to the ODE above is

$$u = C_1 \cos\theta + C_2 \sin\theta + \frac{GMm^2}{\ell^2} = C\cos(\theta - \theta_0) + \frac{GMm^2}{\ell^2}\,.$$

Choose the polar axis so that $\theta_0 = 0$. Using that $u = \dfrac{1}{r}$, we find that

$$\frac{1}{r} = \frac{GMm^2}{\ell^2}(1 + e\cos\theta)\,, \tag{2.13}$$

where $e = \dfrac{C\ell^2}{GMm^2}$. (2.13) is the polar presentation of a conic section, and this proves Kepler's first law of planetary motion.

Example 2.12. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function. To find a relative minimum of f, we first look for critical points of f. In general, it may not be easy to solve for zeros of f'. In this example we provide a way to "find" possible local minimum of f.

Suppose that x_0 is given. If $f'(x_0) < 0$, we expect that the value of $f(x)$ will be smaller than $f(x_0)$ when x is close but on the right-hand side of x_0. Similarly, if $f'(x_0) > 0$, then the value of $f(x)$ will be smaller than $f(x_0)$ when x is close but on the left-hand side of x_0. Therefore, for a given point x_0, we can localize the position of the nearest critical point where f attains a local minimum by "moving" the position of x_0 to the right or to the left based on the sign of f'. This motivates the following IVP

$$x' = -f'(x)\,, \qquad x(0) = x_0\,.$$

In general, for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we use

$$\boldsymbol{x}' = -(\nabla f)(\boldsymbol{x})\,, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0\,,$$

where $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, to find a critical point near $\boldsymbol{x}_0$.

Theorem 2.13 (Existence and Uniqueness of Solution/Fundamental theorem of ODE)**.** *Consider the initial value problem*

$$y^{(n)} = f(t, y, y', \cdots, y^{(n-1)})\,, \quad y(t_0) = y_0\,, \quad y'(t_0) = y_1\,, \quad \cdots \quad y^{(n-1)}(t_0) = y_{n-1}\,. \tag{2.14}$$

If f and the first partial derivatives of f with respect to all its variables, possibly except t, are continuous functions in some rectangular domain $R = [a, b] \times [c_0, d_0] \times [c_1, d_1] \times \cdots \times [c_{n-1}, d_{n-1}]$ that contains the point $(t_0, y_0, y_1, \cdots, y_{n-1})$ in the interior, then the initial value problem has a unique solution $\varphi(t)$ in some interval $I = (t_0 - h, t_0 + h)$ for some positive number h.

2.2 Boundary Value Problems

In this section we only consider ODE of the form

$$y'' + p(x)y' + q(x)y = g(x)\,, \tag{2.15}$$

where p, q and g are given functions, and $y = y(x)$ is the unknown function. Instead of imposing the initial condition $y(t_0) = y_0$ and $y'(t_0) = y_1$, sometimes the following four kinds of boundary condition can be imposed:

1. $y(\alpha) = y_0$, $y(\beta) = y_1$;
2. $y(\alpha) = y_0$, $y'(\beta) = y_1$;
3. $y'(\alpha) = y_0$, $y(\beta) = y_1$;
4. $y'(\alpha) = y_0$, $y'(\beta) = y_1$,

where α, β, y_0 and y_1 are given numbers. Such kind of combination of ODE and boundary condition is called a (two-point) ***boundary value problem*** (**BVP**), and a solution y to a BVP must be defined on the interval $I = [\alpha, \beta]$, as well as satisfy the ODE and the boundary condition.

Example 2.14. In this example we reconsider the ODE in the spring-mass system

$$m\ddot{x} = -kx - r\dot{x} + f(t)\,.$$

We explain the meaning of the different boundary condition as follows:

1. $x(0) = x_0$ and $x(T) = x_1$: the initial and the terminal position of the mass are given.
2. $x(0) = x_0$ and $x'(T) = v_1$: the initial position and the terminal velocity of the mass are given.
3. $x'(0) = v_0$ and $x(T) = x_1$: the initial velocity and the terminal position of the mass are given.
4. $x'(0) = v_0$ and $x'(T) = v_1$: the initial and the terminal velocity of the mass are given.

Example 2.15. Again we consider the ODE

$$m\frac{d^2h}{dt^2} = -\frac{GMm}{(R+h)^2}\,.$$

in Example 1.9. This time we do not require that initial height $h(0)$ and the initial velocity $h'(0)$ are given but instead we want the object to reach certain height H at time $t = T$. Then the BVP is written as

$$m\frac{d^2h}{dt^2} = -\frac{GMm}{(R+h)^2}\,, \qquad h(0) = 0\,, \quad h(T) = H.$$

Similarly, if we want the object to reach certain velocity V at time $t = T$, then we have the BVP

$$m\frac{d^2h}{dt^2} = -\frac{GMm}{(R+h)^2}\,, \qquad h(0) = 0\,, \quad h'(T) = V.$$

Consider the two-point boundary value problem

$$y'' + p(x)y' + q(x)y = g(x)\,, \qquad y(\alpha) = y_0, \quad y(\beta) = y_1\,. \tag{2.16}$$

Let $z(x) = y(x) - \dfrac{x-\alpha}{\beta-\alpha}y_1 - \dfrac{x-\beta}{\alpha-\beta}y_0$. Then z satisfies

$$z'' + p(x)z' + q(x)z = G(x)\,, \qquad z(\alpha) = z(\beta) = 0\,, \tag{2.17}$$

where $G(x) = g(x) - p(x)\dfrac{y_0 - y_1}{\alpha - \beta} - q(x)\big(\dfrac{x-\alpha}{\beta-\alpha}y_1 + \dfrac{x-\beta}{\alpha-\beta}y_0\big)$. Therefore, in general we can assume the homogeneous boundary condition $y_0 = y_1 = 0$ in (2.16). Similarly, ODE (2.15) with the other three kinds of boundary conditions can also be rewritten as a BVP with homogeneous boundary condition.

Remark 2.16. Even though the initial value problem

$$y'' + p(t)y' + q(t)y = g(t)\,, \qquad y(t_0) = y_0\,, \quad y'(t_0) = y_1 \tag{2.18}$$

looks quite similar to the boundary value problem (2.16), they actually differ in some very important ways. For example, if p, q, g are continuous, the initial value problem (2.18) always have a unique solution, while the boundary value problem (2.16) might have no solution or infinitely many solutions:

1. $y'' + y = 0$ with boundary condition $y(0) = y(\pi) = 0$ has infinite many solutions $y_c(x) = c\sin x$.
2. $y'' + y = \sin x$ with boundary condition $y(0) = y(\pi) = 0$ has no solution.

On the other hand, there are cases that (2.16) has a unique solution. For example, the general solution to the boundary value problem

$$y'' + 2y = 0$$

is given by

$$y(x) = C_1 \cos\sqrt{2}x + C_2 \sin\sqrt{2}x\,;$$

thus to validate the boundary condition $y(0) = 1$ and $y(\pi) = 0$, we must have $C_1 = 1$ and $C_2 = -\cot\sqrt{2}\pi$. In other words, the solution $y(x) = \cos\sqrt{2}x - \cot\sqrt{2}\pi\sin\sqrt{2}x$.

The existence theory of the solution to (2.16) requires a totally different functional framework, and will not be proved in this course. However, we will still state the existence theory.

Theorem 2.17. *Let α, β be real numbers and $\alpha < \beta$. Suppose that $p : [\alpha, \beta] \to \mathbb{R}$ is continuously differentiable, and $q : [\alpha, \beta] \to \mathbb{R}$ is continuous. Then* (2.16) (*with $y_0 = y_1 = 0$*) *has a solution if and only if $g : [\alpha, \beta] \to \mathbb{R}$ is integrable and*

$$\int_\alpha^\beta g(x)\varphi(x)\,dx = 0$$

for all φ satisfying $\varphi'' - p(x)\varphi' + \big(q(x) - p'(x)\big)\varphi = 0$ and $\varphi(\alpha) = \varphi(\beta) = 0$. The solution is unique if the ODE $y'' + p(x)y' + q(x)y = 0$ with $y(\alpha) = y(\beta) = 0$ has only trivial solution $y \equiv 0$.

2.3 Solving IVP Using Mablab

We can use the command "ode45" in Matlab to solve for the IVP (2.4). Suppose that we want to solve the IVP

$$y^{(n)} = f(t, y, y', \cdots, y^{(n-1)}), \qquad y(0) = y_0,\ y'(0) = y_1,\ \cdots,\ y^{(n-1)}(0) = y_{n-1}$$

numerically using matlab.

Step 1: Write the IVP in the vector form $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ (form (2.1)) with initial condition $\boldsymbol{y}(0) = \boldsymbol{y}_0$. Note that usually you need to write the IVP in a dimensionless form and then transform

Step 2: Write (and save) the function $\boldsymbol{f}$ in matlab.

Step 3: Once the function $\boldsymbol{f}$ is saved, use the command "ode45" (based on the **adaptive Runge-Kutta** method) to solve the IVP: the format is

```
[t,y] = ode45(@name of the function,[starting time, terminal time], initial data)
```

where the output of this command has two pieces t and y (whose names can also be changed and does not have to agree with the names you use in writing the function):

(a) t is a column vector whose components are the samples of time at which the numerical solution evaluates.

(b) y is a $m \times n$ matrix, where m is the total number of time samples, and n is the dimension of the vector y.

To illustrate how these steps are carried out, we look at the following example.

Example 2.18. In this example we solve for the IVP (from the Lotka-Volterra model)

$$\frac{dp}{dt} = -0.16p + 0.08pq\,, \tag{2.19a}$$

$$\frac{dq}{dt} = 4.5q - 0.9pq\,, \tag{2.19b}$$

$$p(0) = 5, \quad q(0) = 3\,. \tag{2.19c}$$

Let $\boldsymbol{y} = [p, q]^{\mathrm{T}}$, and $\boldsymbol{f}(t, \boldsymbol{y}) = \begin{bmatrix} -0.16p + 0.08pq \\ 4.5q - 0.9pq \end{bmatrix}$ numerically using matlab. In matlab, the function $\boldsymbol{f}$ can be given by the following m-file:

```
function yp = ODE_RHS(t,y)
yp(1,1) = -0.16*y(1,1) + 0.08*y(1,1)*y(2,1);
yp(2,1) = 4.5*y(2,1) - 0.9*y(1,1)*y(2,1);
```
(2.20)

where

1. the word "function" in the first line indicates that this m-file will be a function that you can use in matlab.

2. yp is the name of the output variable, and t, y are the names of the input variables (and the names can be changed); however, you should keep t (time) as the first input/variable and y (the unknowns in the ODE) as the second input/variable in order to use the built-in matlab ODE solver.

3. ODE_RHS is the name of the function (and also the name of the file so that matlab can see it) that will be used/recognized in matlab. The name can be changed but you need to have this name different from built-in functions such as sin, exp, and etc.

4. In this example, the input y is a 2-d column vector. y(1,1) and y(2,1) denote the first and second component of y, respectively. Similarly, the output *yp* is also a 2-d column vector, and yp(1,1) and yp(2,1) denote the first and second component of yp, respectively.

Once the function is saved, you can check if matlab is able to use this function by assigning the value of t and y (remember, y has to be a 2-d column vector) and see if it outputs the correct value. For example, in the main window of matlab you can type

```
ODE_RHS(1,[2;5])
```

where $[2;5]$ is the column vector $[2,5]^{\mathrm{T}}$, and it should output something like this

```
>> ODE_RHS(1,[2;5])

ans =

  0.4800
 13.5000
```

which means the first component of the output (in our code it is yp(1,1)) is 0.48 while the second component of the output (in our code it is yp(2,1)) is 13.5.

For the readability of codes, we recommender the reader to have (2.20) written, at least, as

```
function yp = ODE_RHS(t,y)
p = y(1,1);
q = y(2,1);
yp(1,1) = -0.16*p + 0.08*p*q;
yp(2,1) = 4.5*q - 0.9*p*q;
```

As long as the function $\boldsymbol{f}$ (named ODE_RHS) is saved, we can use the matlab built-in ODE solver "ode45" to solve for the IVP (2.19). In the main window of matlab, type

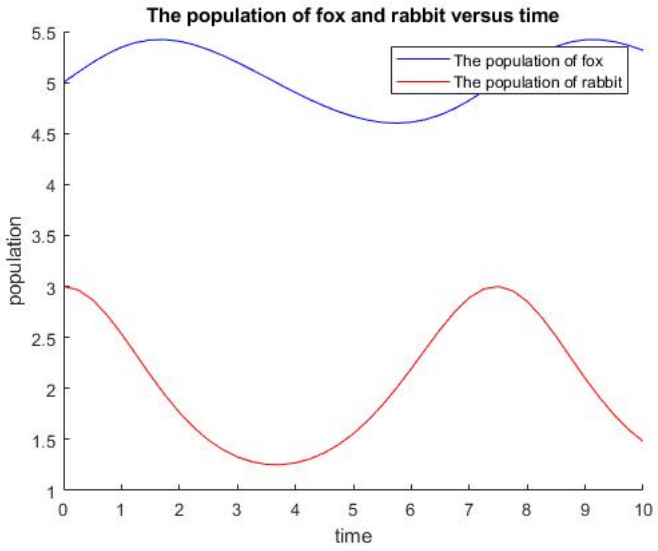
```
[t,y] = ode45(@ODE_RHS,[0,10],[5;3]);
```

to solve (numerically) for the IVP in the time interval $[0, 10]$ and initial data $[5, 3]^{\mathrm{T}}$. In this case, the solution y is an $m \times 2$ matrix: the first column is the value of p (at those sampled time t) and the second column is the value of q (at those sampled time t).

• **Visualization of the numerical solution**: In the following we provide two codes

```
figure(1)
title('The population of fox and rabbit versus time')
hold on;
plot(t,y(:,1),'b');
plot(t,y(:,2),'r');
legend('The population of fox','The population of rabbit')
xlabel('time')
ylabel('population')
```

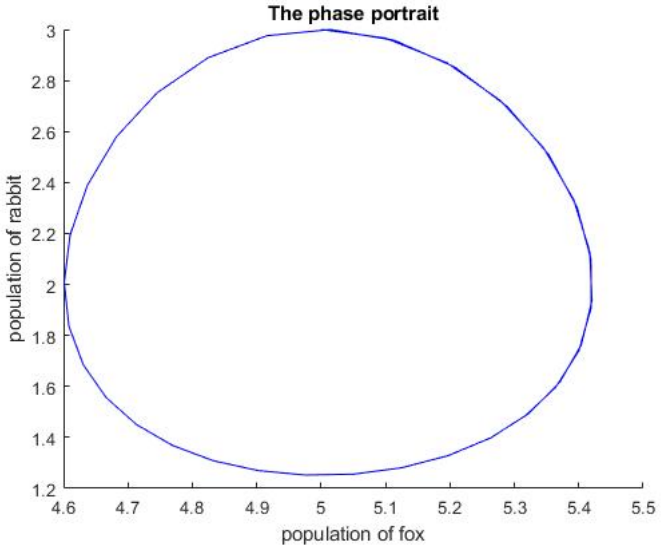
which outputs



and

```
figure(2)
title('The phase portrait')
hold on;
plot(y(:,1),y(:,2),'b');
xlabel('population of fox')
ylabel('population of rabbit')
```

which outputs



for the visualization of the numerical solution. The figures themselves should explain the codes clearly.

Example 2.19. In this example we look for the minimum of the function $f(x, y) = xe^{-x^2-y^2}$ using gradient flows. First we provide the graph of f so that we have some information about this function. To do this, do the following:

```
[x,y] = meshgrid(–2:0.1:2,–2:0.1:2);
z = x.*exp(–x.∧2–y.∧2);
surf(z);
```

and this will produce the following figure

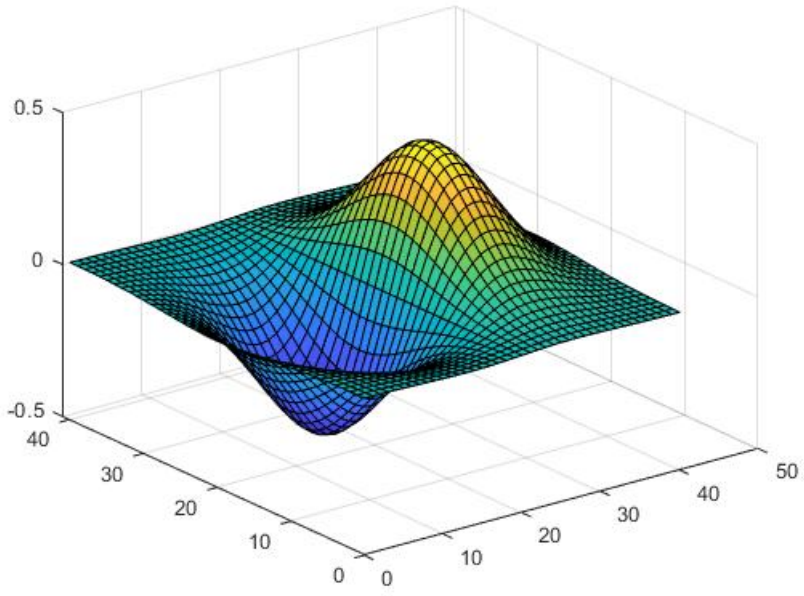


Figure 2.3: The graph of the function $f(x, y) = xe^{-x^2-y^2}$.

From the graph of f, we find that there is a minimum and a maximum for f.

Now we try to find the minimum using the gradient flow. We compute the first partial derivative of f and obtain that

$$f_x(x,y) = (1-2x^2)e^{-x^2-y^2} \qquad \text{and} \qquad f_y(x,y) = -2xye^{-x^2-y^2}\,.$$

Therefore, we will focus on the following ODE

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = -(\nabla f)(x,y) = \begin{bmatrix} (2x^2-1)e^{-x^2-y^2} \\ 2xye^{-x^2-y^2} \end{bmatrix} \equiv \boldsymbol{F}\big(t,[x,y]'\big)\,.$$

As in the previous example, we first name (and save) the function $\boldsymbol{F}$ as ODE_RHS (again, the name of the function can be changed) as follows:

```
function yp = ODE_RHS(t,INPUT)
x = INPUT(1,1);
y = INPUT(2,1);
yp(1,1) = (2*x^2-1)*exp(-x^2-y^2);
yp(2,1) = 2*x*y*exp(-x^2-y^2);
```

Here we rename the second input of the function as "INPUT" in order to differentiate this input from the real variable y in the equation. Maybe it is much clearer if we rewrite the code as

```
function zp = ODE_RHS(t,z)
x = z(1,1);
y = z(2,1);
zp(1,1) = (2*x^2-1)*exp(-x^2-y^2);
zp(2,1) = 2*x*y*exp(-x^2-y^2);
```

Once we finish saving the function ODE_RHS, we can use

```
[t,y] = ode45(@ODE_RHS,[0,10],[0.5;0.5]);
```

or

```
[t,y] = ode45(@(t,y) ODE_RHS(t,y),[0,10],[0.5;0.5]);
```

↑ there is a space here

to find the numerical solution of the gradient flow with initial condition $[x(0),y(0)] = [0.5,0.5]$. We are only interested in the final destination of the flow; thus we use

```
y(end,:)
```

to find the last row of y (note that the unknown is a 2-d column vector, so the output y using "ode45" will be an $N \times 2$ matrix) and obtain that

```
>> y(end,:)

ans =

  -0.7071    0.0006
```

From the computation of the gradient of f, we find that the critical points of f should be $\left(\pm\frac{1}{\sqrt{2}}, 0\right)$. So, why does the gradient flow not produce the correct/approximated critical point? This is because the time interval is too small so that the flow has not reach its final destination yet. Let us replace the time interval as $[0, 20]$ and rerun the whole process again, one should obtain y(end,:) = [−0.7071 0.0000].

• **Geometric point of view**: The solution to the IVP

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = -(\nabla f)(x, y)\,, \tag{2.21a}$$

$$\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \tag{2.21b}$$

produces a curve $(x(t), y(t))$, where t belongs to some time interval (for example $[0, 10]$ or $[0, 20]$ in our previous tests). This curve is called an ***integral curve*** of the direction field $-(\nabla f)(x, y)$, and the initial data (x_0, y_0) is the point where the integral curve starts and is called the starting point of the curve (in the code above the starting point is $(0.5, 0.5)$). The ODE (2.21a) shows that the tangent direction of the integral curve should agree with the direction field.

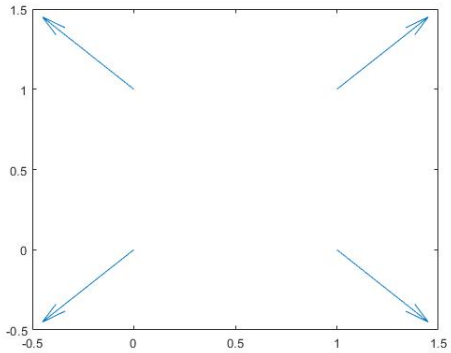
Let us visualize this by plotting first the vector field $-(\nabla f)$. To plots a vector $\boldsymbol{u}$ = (x component, y component) at the point p = (x coordinate, y coordinate), we use the command "quiver" in the following way:

```
quiver(x coordinate, y coordinate, x component, y component)
```

For example, if you want to plot 4 vectors $(1, 1)$, $(-1, -1)$, $(1, -1)$ and $(-1, 1)$ at 4 points $(1, 1)$, $(0, 0)$, $(1, 0)$ and $(0, 1)$, respectively, you can do the following:

```
L = [1,1;0,0;1,0;0,1];
V = [1,1;–1,–1;1,–1;–1,1];
quiver(L(:,1),L(:,2),V(:,1),V(:,2));
```

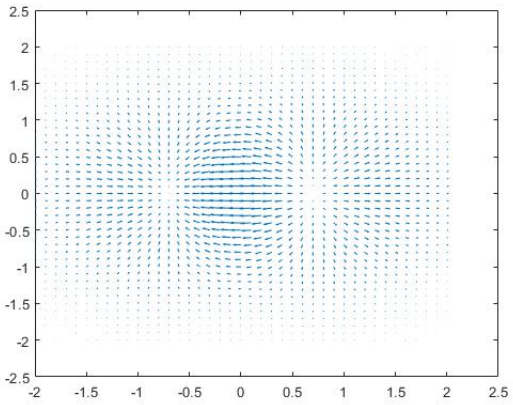
and the following figure will be produced:



Note that if you replace the last line of commands by "quiver(L,V)", it will produce garbages. You need to give "quiver" the x coordinate and y coordinate of base points, as well as the x component and y component of vectors, separately, in order to have the correct plot. Now, since we have build up a grid using "[x,y] = meshgrid(-2:.1:2,-2:.1:2);", we can simply use

```
quiver(x,y,(2*x.∧2–1).*exp(–x.∧2–y.∧2),2*x.*y.*exp(–x.∧2–y.∧2))
```

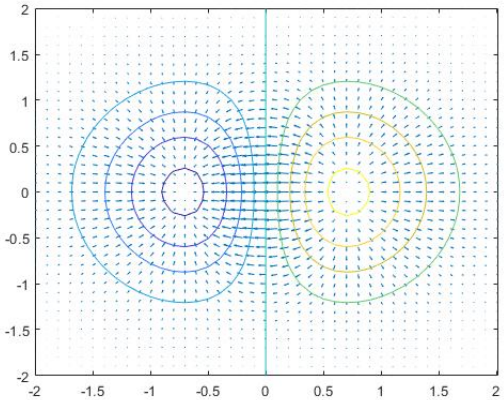
to produce the following figure of the vector field:



We can also add the level sets of f onto the plot by the following command

```
contour(x,y,z)
```

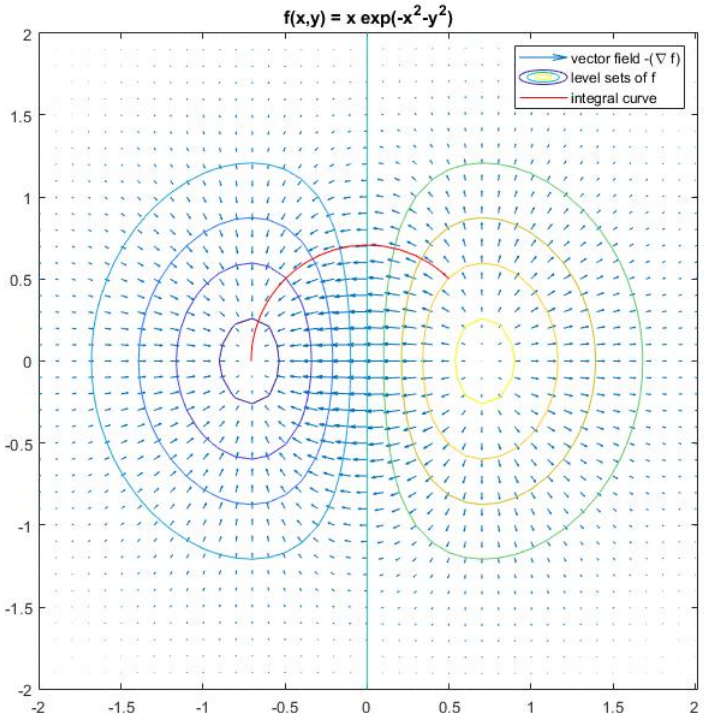
so that we obtain



Finally, we plot the integral curve (in red color) using

```
plot(y(:,1),y(:,2),'r')
```

after the ode solver "[t,y] = ode45(@ODE_RHS,[0,20],[0.5;0,5]);" is applied. You should be able to obtain the following figure:



We note that the tangent direction of the integral curve is indeed parallel to the vector field $-(\nabla f)$, and the integral curve is perpendicular to the level set of f (which agrees with what we learned in Calculus).

We summarize our codes in the following (in case you cannot reproduce the result):

```
[x,y] = meshgrid(–2:0.1:2,–2:0.1:2);
z = x.*exp(–x.∧2–y.∧2);
figure(1)
title('f(x,y) = x exp(–x∧2–y∧2)')
hold on;
quiver(x,y,(2*x.∧2–1).*exp(–x.∧2–y.∧2),2*x.*y.*exp(–x.∧2–y.∧2))
contour(x,y,z)
[t,y] = ode45(@ODE_RHS,[0,20],[0.5;0.5]);
plot(y(:,1),y(:,2),'r');
axis equal;
legend('vector field –(\nabla f)','level sets of f','integral curve')
```

Chapter 3

Partial Differential Equations

3.1 Models with One Temporal Variable and One Spatial Variable

3.1.1 The 1-dimensional conservation laws

Suppose that a substance of interest lives in a 1-dimensional space such as a tube. Let $u(x,t)$ be the density or concentration of the substance at position x and time t. Then

$$\int_x^{x+\Delta x} u(y,t)\,dt$$

is the total amount of the substance in the interval $I=[x,x+\Delta x]$ at time t; thus during the time period $[t,t+\Delta t]$, the change of the amount of the substance in the interval I in the time period $[t,t+\Delta t]$ is given by

$$\int_x^{x+\Delta x} u(y,t+\Delta t)\,dt - \int_x^{x+\Delta x} u(y,t)\,dt = \int_x^{x+\Delta x} \big[u(y,t+\Delta t)-u(y,t)\big]\,dy\,.$$

On the other hand, there are two sources of changing the amount of the substance in the interval I:

1. a flux that describes any effect that appears to pass or travel the substance through points.

2. a source that will release or absorb the substance in this interval.

Let f denote the flux and q denote the source. Then in the time interval $[t,t+\Delta t]$ the amount of the substance flowing into I from the point x is given by

$$\int_t^{t+\Delta t} f(x,t')\,dt'$$

while amount of the substance flowing out of I from the point $x+\Delta x$ is given by

$$\int_t^{t+\Delta t} f(x+\Delta x,t')\,dt'\,.$$

Moreover, the change of the amount of the substance in the interval I in the time period $[t, t+\Delta t]$ due to the source is given by

$$\int_t^{t+\Delta t} \int_x^{x+\Delta x} q(y,t')\,dydt'\,.$$

Therefore, the change of amount of the substance in the interval I in the time period $[t, t+\Delta t]$ is given by

$$\int_t^{t+\Delta t} \big[f(x,t') - f(x+\Delta x,t')\big]dt' + \int_t^{t+\Delta t} \int_x^{x+\Delta x} q(y,t')\,dydt'\,.$$

As a consequence,

$$\begin{aligned}&\int_x^{x+\Delta x} \big[u(y,t+\Delta t) - u(y,t)\big]\,dy\\ &\qquad = \int_t^{t+\Delta t} \big[f(x,t') - f(x+\Delta x,t')\big]dt' + \int_t^{t+\Delta t} \int_x^{x+\Delta x} q(y,t')\,dydt'\,.\end{aligned}$$

Dividing both sides through Δx and then passing to the limit as $\Delta x \to 0$, by the fundamental theorem of Calculus we find that (without any rigorous verification)

$$u(x,t+\Delta t) - u(x,t) = -\int_t^{t+\Delta t} \frac{\partial}{\partial x} f(x,t')\,dt' + \int_t^{t+\Delta t} q(x,t')\,dt'\,.$$

Similarly, dividing both sides of the equality above through Δt and then passing to the limit as $\Delta t \to 0$, the fundamental theorem of Calculus implies that

$$\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(x,t) = q(x,t)\,.$$

Example 3.1 (Traffic flows)**.** Consider the traffic on the highway (parameterized by $\mathbb{R}$). Let u denote the car density (given in the number of vehicles per unit length). Then the flux f is a function of u with the property that

(a) $f(u) = 0$ if $u = 0$ or $u > L$,

(b) $f'(u) > 0$ if $u \in (0, u_{\max})$, and $f'(u) < 0$ if $u \in (u_{\max}, L)$.

If f is differentiable, and $f'(u) = c(u)$. Then the equation of continuity reads

$$u_t(x,t) + c(u(x,t))u_x(x,t) = q(x,t) \qquad \forall\, x \in \mathbb{R}\,, t \in \mathbb{R}$$

which can be abbreviated as

$$u_t + c(u)u_x = q \qquad \text{in} \quad \mathbb{R} \times \mathbb{R}\,.$$

To complete the model, an initial condition

$$u(x,0) = u_0(x) \qquad \forall\, x \in \mathbb{R} \qquad (\text{or simply } u = u_0 \text{ on } \mathbb{R} \times \{t = 0\})$$

has to be imposed.

When the domain of interest (for example, the highway) has finite length, we can parameterize it as $[0, L]$; however, to complete the model we also have to impose the boundary condition which tells us what happened to u at the start and end of the highway. The boundary condition for 1-d conservation laws are usually given by one of the following three types:

1. $u(0,t) = u_L$ and $u(L,t) = u_R$ which says that the boundary value of u is prescribed (and can be time-dependent if u_L or u_R are time dependent).

2. $u_x(0,t) = 0$ and $u(L,t) = 0$ which says that the derivative of u on the boundary is zero.

3. Mixed boundary condition: on one end u is given and on the other end u_x is given.

3.1.2 The 1-dimensional heat equations

Consider the heat distribution on a rod of length L: Parameterize the rod by $[0, L]$, and let t be the time variable. Let $\rho(x)$, $s(x)$, $\kappa(x)$ denote the density, specific heat, and the thermal conductivity of the rod at position $x \in (0, L)$, respectively, and $\vartheta(x,t)$ denote the temperature at position x and time t. For $0 < x < L$, and $\Delta x, \Delta t \ll 1$,

$$\int_x^{x+\Delta x} \rho(y)s(y)\big[\vartheta(y,t+\Delta t) - \vartheta(y,t)\big]\,dy = \int_t^{t+\Delta t} \big[\kappa(x+\Delta x)\vartheta_x(x+\Delta x,t') - \kappa(x)\vartheta_x(x,t')\big]\,dt',$$

where the left-hand side denotes the change of the total heat in the small section $(x, x+\Delta x)$, and the right-hand side denotes the heat flows from outside. If there is a heat source Q, then the equation above can be modified as

$$\begin{aligned}\int_x^{x+\Delta x} &\rho(y)s(y)\big[\vartheta(y,t+\Delta t) - \vartheta(y,t)\big]\,dy \\ &= \int_t^{t+\Delta t} \big[\kappa(x+\Delta x)\vartheta_x(x+\Delta x,t') - \kappa(x)\vartheta_x(x,t')\big]\,dt' + \int_t^{t+\Delta t}\int_x^{x+\Delta x} Q(y,t')\,dydt'\,.\end{aligned}$$

Dividing both sides by Δt and passing to the limit as $\Delta t \to 0$, by the Fundamental Theorem of Calculus (assuming that all the functions appearing in the equation above are smooth enough) we obtain that

$$\int_x^{x+\Delta x} \rho(y)s(y)\vartheta_t(y,t)\,dy = \big[\kappa(x+\Delta x)\vartheta_x(x+\Delta x,t) - \kappa(x)\vartheta_x(x,t)\big] + \int_x^{x+\Delta x} Q(y,t)\,dy\,.$$

Dividing both sides of the equation above by Δx and then passing to the limit to $\Delta x \to 0$, we find that

$$\rho(x)s(x)\vartheta_t(x,t) = \big[\kappa(x)\vartheta_x(x,t)\big]_x + Q(x,t) \qquad 0 < x < L\,, \quad t > 0\,. \tag{3.1}$$

Assuming uniform rod; that is, ρ, s, κ are constant, then (3.1) reduces to that

$$\vartheta_t(x,t) = \alpha^2 \vartheta_{xx}(x,t) + q(x,t)\,, \qquad 0 < x < L\,, \quad t > 0\,, \tag{3.2a}$$

where $\alpha^2 = \dfrac{\kappa}{\rho s}$ is called the ***thermal diffusivity***.

To determine the state of the temperature, we need to impose that initial condition

$$\vartheta(x,0) = \vartheta_0(x) \qquad 0 < x < L \tag{3.2b}$$

and a boundary condition.

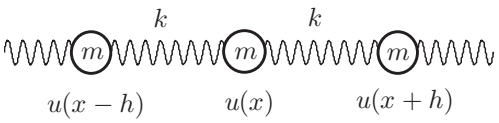
(a) Temperature on the end-points of the rod is fixed: $\vartheta(0,t) = T_1$ and $\vartheta(L,t) = T_2$.

(b) Insulation on the end-points of the rod: $\vartheta_x(0,t) = \vartheta_x(L,t) = 0$.

(c) Mixed boundary conditions: $\vartheta(0,t) = T_1$ and $\vartheta_x(L,t) = 0$, or $\vartheta(L,t) = T_2$ and $\vartheta_x(0,t) = 0$.

3.1.3 The 1-dimensional wave equations

1. From Hooke's law:



imagine an array of little weights of mass m interconnected with massless springs of length h, and the springs have a stiffness of k (see the figure). If $u(x)$ measures the distance from the equilibrium of the mass situated at x, then the forces exerted on the mass m at the location x are

$$\begin{aligned} F_{\text{Newton}} &= ma = m\frac{\partial^2 u}{\partial t^2}(x,t) \\ F_{\text{Hooke}} &= k\big[u(x+h,t) - u(x,t)\big] - k\big[u(x,t) - u(x-h,t)\big] \\ &= k\big[u(x+h,t) - 2u(x,t) + u(x-h,t)\big]\,. \end{aligned}$$

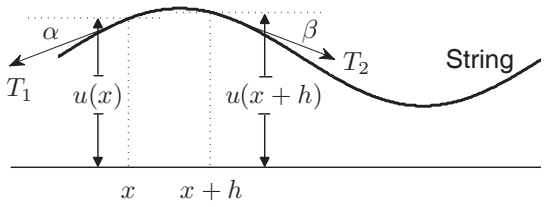
If the array of weights consists of N weights spaced evenly over the length $L = (N+1)h$ of total mass $M = Nm$, and the total stiffness of the array $K = k/N$, then

$$\frac{\partial^2 u}{\partial t^2}(x,t) = \frac{N}{N+1}\frac{KL^2}{M}\frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2}\,.$$

Taking the limit $N \to \infty$, $h \to 0$ (and assuming smoothness) we obtain

$$u_{tt}(x,t) = c^2 u_{xx}(x,t)\,. \tag{3.3}$$

2. Equation of vibrating string: let $u(x,t)$ measure the distance of a string from its equilibrium, and $T(x,t)$ denote the tension of the string at position x and time t.



Assuming only motion in the vertical direction, the horizontal component of tensions $T_1 = T(x,t)$ and $T_2 = T(x+h,t)$ have to be the same; thus

$$T_1 \cos\alpha = T_2 \cos\beta\,. \tag{3.4}$$

Noting that

$$\cos\alpha = \frac{1}{\sec\alpha} = \frac{1}{\sqrt{1+\tan^2\alpha}} = \frac{1}{\sqrt{1+u_x(x,t)^2}}\,,$$
$$\cos\beta = \frac{1}{\sec\beta} = \frac{1}{\sqrt{1+\tan^2\beta}} = \frac{1}{\sqrt{1+u_x(x+h,t)^2}}\,,$$

identity (3.4) implies that the function $\dfrac{T(x,t)}{\sqrt{1+u_x(x,t)^2}}$ is constant in x (but not necessary in t). Denote this constant as $\tau(t)$. Then by the fact that the difference of the vertical component of T_1 and T_2 induces the motion in the vertical direction, we obtain that

$$\begin{aligned} m\frac{\partial^2 u}{\partial t^2}(x+\theta h,t) &= T_2\sin\beta - T_1\sin\alpha = (T_2\cos\beta)\tan\beta - (T_1\cos\alpha)\tan\alpha \\ &= \tau(t)\big[u_x(x+h,t) - u_x(x,t)\big]\,, \end{aligned}$$

here we use $\dfrac{\partial^2 u}{\partial t^2}(x+\theta h,t)$, where $0<\theta<1$, to denote the average acceleration of the segment from x to $x+h$. If μ is the density of the string, then $m=\mu h$; hence

$$\mu\frac{\partial^2 u}{\partial t^2}(x,t) = \tau(t)\frac{u_x(x+h,t)-u_x(x,t)}{h}\,.$$

Passing to the limit as $h\to 0$, we obtain

$$\mu u_{tt}(x,t) = \tau(t)u_{xx}(x,t)\,. \tag{3.5}$$

If there is an external forcing f acting on the string, then (3.5) becomes

$$\mu u_{tt}(x,t) = \tau(t)u_{xx}(x,t) + f(x,t)\,. \tag{3.6}$$

If τ is constant in t (which is a reasonable assumption if the vibration of the string is very small and uniform), then (3.6) reduces to

$$u_{tt}(x,t) = c^2 u_{xx}(x,t) + \frac{1}{\mu}f(x,t)\,. \tag{3.7}$$

Initial conditions: $\begin{cases} u(x,0) = \varphi(x) \\ u_t(x,0) = \psi(x) \end{cases}$, where φ and ψ are given functions.

Boundary conditions:

(a) Vibration string with fixed ends: $u(0,t) = u(L,t) = 0$.

(b) Vibration string with free ends: $u_x(0,t) = u_x(L,t) = 0$.

(c) Mixed boundary conditions: $u(0,t) = u_x(L,t) = 0$ or $u(L,t) = u_x(0,t) = 0$.

3.2 Models with Several Spatial Variables

3.2.1 The Divergence Theorem

• The Surface Integrals

Definition 3.2. A subset $\Sigma \subseteq \mathbb{R}^3$ is called a surface if for each $p \in \Sigma$, there exist an open neighborhood $\mathcal{U} \subseteq \Sigma$ of p ($\mathcal{U}$ is the intersection of Σ and some open balls in $\mathbb{R}^3$), an open set $\mathcal{V} \subseteq \mathbb{R}^2$, and a continuous map $\varphi : \mathcal{U} \to \mathcal{V}$ such that $\varphi : \mathcal{U} \to \mathcal{V}$ is one-to-one, onto, and its inverse $\psi = \varphi^{-1}$ is also continuous. Such a pair $\{\mathcal{U}, \varphi\}$ is called a coordinate chart (or simply chart) at p, and $\{\mathcal{V}, \psi\}$ is called a (local) parametrization at p.

Remark 3.3. In some literatures the surface is defined in the following equivalent but reversed way: A subset $\Sigma \subseteq \mathbb{R}^3$ is a surface if for each $p \in \Sigma$, there exists a neighborhood $\mathcal{U} \subseteq \mathbb{R}^3$ of p and a map $\psi : \mathcal{V} \to \mathcal{U} \cap \Sigma$ of an open set $\mathcal{V} \subseteq \mathbb{R}^2$ onto $\mathcal{U} \cap \Sigma \subseteq \mathbb{R}^3$ such that ψ is a homeomorphism; that is, ψ has an inverse $\varphi = \psi^{-1} : \mathcal{U} \cap \Sigma \to \mathcal{V}$ which is continuous. The mapping ψ is called a parametrization or a system of (local) coordinates in (a neighborhood of) p.

Definition 3.4 (Regular surfaces)**.** A surface $\Sigma \subseteq \mathbb{R}^3$ is said to be regular if for each $p \in \Sigma$, there exists a differentiable local parametrization $\{\mathcal{V}, \psi\}$ of Σ at p such that $\psi_{,1}\left(\psi^{-1}(p)\right)$ and $\psi_{,2}\left(\psi^{-1}(p)\right)$ are linearly independent, where

$$\psi_{,1}\left(\psi^{-1}(p)\right) = \frac{\partial}{\partial u}\Big|_{(u,v)=(q_1,q_2)=\psi^{-1}(p)} \psi \quad \text{and} \quad \psi_{,2}\left(\psi^{-1}(p)\right) = \frac{\partial}{\partial v}\Big|_{(u,v)=(q_1,q_2)=\psi^{-1}(p)} \psi$$

denote, respectively, the first partial derivative of ψ with respect to its first and second variable at point $\psi^{-1}(p)$. The span of the two vectors $\psi_{,1}\left(\psi^{-1}(p)\right)$ and $\psi_{,2}\left(\psi^{-1}(p)\right)$ is called the ***tangent plane*** of Σ at p, and is denoted by $\mathbf{T}_p\Sigma$.

Remark 3.5. A vector-valued function $\psi : \mathcal{V} \to \mathbb{R}^3$ is differentiable if each component of ψ is differentiable, and the derivative of ψ, denoted by $D\psi$, is defined by

$$\left[D\psi(q)\right] = \begin{bmatrix} \dfrac{\partial \psi_1}{\partial u}(q) & \dfrac{\partial \psi_1}{\partial v}(q) \\ \dfrac{\partial \psi_2}{\partial u}(q) & \dfrac{\partial \psi_2}{\partial v}(q) \\ \dfrac{\partial \psi_3}{\partial u}(q) & \dfrac{\partial \psi_3}{\partial v}(q) \end{bmatrix}.$$

Therefore, Σ is regular if for each p there exists a local parametrization $\{\mathcal{V}, \psi\}$ at p such that $\big[D\psi\big]$ has full rank at $\psi^{-1}(p)$ (or equivalently, $\big[D\psi\big]$ is injective at $\psi^{-1}(p)$).

In the following, we always assume that the matrix $\big[D\psi(q)\big]$ has full rank for all $q \in \mathcal{V}$ if $\{\mathcal{V}, \psi\}$ is a local parametrization of a regular surface $\Sigma \subseteq \mathbb{R}^3$.

Remark 3.6. Let $p \in \Sigma$ and $q = \psi^{-1}(p)$. Since $D\psi(q)$ is injective, each $\boldsymbol{v} \in \mathbf{T}_p\Sigma$ corresponds a unique vector $(a, b) \in \mathbb{R}^2$ such that $\boldsymbol{v} = a\psi_{,1}(q) + b\psi_{,2}(q)$. This vector $(a, b) \in \mathbb{R}^2$ satisfies $[\boldsymbol{v}] = \big[D\psi(q)\big][a, b]^{\mathrm{T}}$, and can be computed by

$$\begin{bmatrix} a \\ b \end{bmatrix} = \Big(\big[D\psi(q)\big]^{\mathrm{T}}\big[D\psi(q)\big]\Big)^{-1}\big[D\psi(q)\big]^{\mathrm{T}}[\boldsymbol{v}]\,.$$

Example 3.7. Let $\mathbb{S}^2 = \big\{(x, y, z) \in \mathbb{R}^3 \,\big|\, x^2 + y^2 + z^2 = 1\big\}$ be the unit sphere in $\mathbb{R}^3$. If $p = (x_0, y_0, z_0) \in \mathbb{S}^2$, then either x_0, y_0 or z_0 is non-zero. Suppose that $z_0 \neq 0$. Let $r = 1 - \sqrt{x_0^2 + y_0^2} > 0$. Define

$$\psi(x, y) = \begin{cases} \big(x, y, \sqrt{1 - x^2 - y^2}\big) & \text{if } z_0 > 0\,, \\ \big(x, y, -\sqrt{1 - x^2 - y^2}\big) & \text{if } z_0 < 0\,, \end{cases}$$

$\mathcal{V} = B\big((x_0, y_0), r\big)$, and $\mathcal{U} = \psi(\mathcal{V})$. Then $\psi : \mathcal{V} \to \mathcal{U}$ is a bijection. Let $\varphi = \psi^{-1}$. Then $\{\mathcal{U}, \varphi\}$ is a coordinate chart at p; thus $\mathbb{S}^2$ is a surface.

There exists another coordinate chart. Let $\mathcal{U}_1 = \mathbb{S}^2\backslash(0, 0, -1)$ and $\mathcal{U}_2 = \mathbb{S}^2\backslash(0, 0, 1)$. Define the map $\varphi_1 : \mathcal{U}_1 \to \mathbb{R}^2$ by that $\varphi_1(p)$ is the unique point on $\mathbb{R}^2$ such that $(0, 0, -1)$, $\varphi_1(p)$ and $(x, y, 0)$ are on the same straight line. Similarly, define $\varphi_2 : \mathcal{U}_2 \to \mathbb{R}^2$ by that $\varphi_2(p)$ is the unique point on $\mathbb{R}^2$ such that $(0, 0, 1)$, $\varphi_2(p)$ and $(x, y, 0)$ are on the same straight line. It is easy to check that if $p \in \mathbb{S}^2$, then either $\{\mathcal{U}_1, \varphi_1\}$ or $\{\mathcal{U}_2, \varphi_2\}$ is a coordinate chart at p.

A third kind of coordinate chart is given as follows. Let $\mathcal{U} = (0, 2\pi) \times (0, \pi)$, and define

$$\psi(\theta, \phi) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi).$$

Then $\psi : \mathcal{U} \to \mathbb{S}^2\backslash\{(x, 0, z) \,|\, 0 \leqslant x \leqslant 1, x^2 + z^2 = 1\}$ is a continuous bijection with a continuous inverse. We note that for any $\mathcal{U} = (\theta_0, \theta_0 + 2\pi) \times (\phi_0, \phi_0 + \pi)$, ψ is a homeomorphism between $\mathcal{U}$ and an open subset of $\mathbb{S}^2$.

Using one of the three parametrizations above, we find that $\psi_{,1}$ and $\psi_{,2}$ must be linearly independent; thus we find that $\mathbb{S}^2$ is a regular surface.

•• The metric tensor and the first fundamental form

Definition 3.8 (Metric)**.** Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface. The metric tensor associated with the local parametrization $\{\mathcal{V}, \psi\}$ (at some point $p \in \Sigma$) is the matrix $g = [g_{\alpha\beta}]_{2\times 2}$ given by

$$g_{\alpha\beta} = \psi_{,\alpha} \cdot \psi_{,\beta} = \sum_{i=1}^{3} \frac{\partial \psi^i}{\partial y_\alpha}\frac{\partial \psi^i}{\partial y_\beta} \qquad \text{in} \quad \mathcal{V}$$

or equivalently, $g = [D\psi]^{\mathrm{T}}[D\psi]$.

Proposition 3.9. *Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface, and $g = [g_{\alpha\beta}]_{2\times 2}$ be the metric tensor associated with the local parametrization $\{\mathcal{V}, \psi\}$ (at $p \in \Sigma$). Then the metric tensor g is positive definite; that is,*

$$\sum_{\alpha,\beta=1}^{2} g_{\alpha\beta} v^\alpha v^\beta > 0 \qquad \forall\, \boldsymbol{v} = \sum_{\gamma=1}^{2} v^\gamma \frac{\partial \psi}{\partial y^\gamma} \neq \mathbf{0}\,.$$

Proof. Since $D\psi$ has full rank on $\mathcal{V}$, every tangent vector $\boldsymbol{v}$ can be expressed as the linear combination of $\left\{\dfrac{\partial \psi}{\partial y_1}, \dfrac{\partial \psi}{\partial y_2}\right\}$. Write $\boldsymbol{v} = \displaystyle\sum_{\gamma=1}^{2} v^\gamma \frac{\partial \psi}{\partial y^\gamma}$. Then if $\boldsymbol{v} \neq \mathbf{0}$,

$$0 < \|\boldsymbol{v}\|^2_{\mathbb{R}^3} = \sum_{i=1}^{3} \sum_{\alpha,\beta=1}^{2} v^\alpha \frac{\partial \psi^i}{\partial y_\alpha} v^\beta \frac{\partial \psi^i}{\partial \psi_\beta} = \sum_{\alpha,\beta=1}^{2} g_{\alpha\beta} v^\alpha v^\beta\,. \qquad \square$$

Definition 3.10 (The first fundamental form)**.** Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface, and $g = [g_{\alpha\beta}]_{2\times 2}$ be the metric tensor associated with the local parametrization $\{\mathcal{V}, \psi\}$ (at $p \in \Sigma$). The first fundamental form associated with the local parametrization $\{\mathcal{V}, \psi\}$ (at $p \in \Sigma$) is the scalar function $\mathrm{g} = \det(g)$.

Theorem 3.11. *Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface, and $\{\mathcal{V}, \psi\}$ be a local parametrization at $p \in \Sigma$. Then*

$$\sqrt{\mathrm{g}} = \|\psi_{,1} \times \psi_{,2}\|_{\mathbb{R}^3}\,. \tag{3.8}$$

Proof. Using the permutation symbol (given in the next remark) and Kronecker's delta, we have

$$\begin{aligned}
\|\psi_{,1} \times \psi_{,2}\|^2_{\mathbb{R}^3} &= \sum_{i=1}^{3} \Big(\sum_{j,k=1}^{3} \varepsilon_{ijk} \psi^j{}_{,1}\, \psi^k{}_{,2} \Big) \Big(\sum_{r,s=1}^{3} \varepsilon_{irs} \psi^r{}_{,1}\, \psi^s{}_{,2} \Big) \\
&= \sum_{j,k,r,s=1}^{3} \Big[\Big(\sum_{i=1}^{3} \varepsilon_{ijk}\varepsilon_{irs} \Big) \psi^j{}_{,1}\, \psi^k{}_{,2}\, \psi^r{}_{,1}\, \psi^s{}_{,2} \Big] \\
&= \sum_{j,k,r,s=1}^{3} \big(\delta_{jr}\delta_{ks} - \delta_{js}\delta_{kr} \big) \psi^j{}_{,1}\, \psi^k{}_{,2}\, \psi^r{}_{,1}\, \psi^s{}_{,2}\,,
\end{aligned}$$

where we use the identity

$$\sum_{i=1}^{3} \varepsilon_{ijk}\varepsilon_{irs} = \delta_{jr}\delta_{ks} - \delta_{js}\delta_{kr} \tag{3.9}$$

to conclude the last equality. Therefore,

$$\begin{aligned}
\|\psi_{,1} \times \psi_{,2}\|^2_{\mathbb{R}^3} &= \sum_{j,k=1}^{3} \big(\psi^j{}_{,1}\, \psi^k{}_{,2}\, \psi^j{}_{,1}\, \psi^k{}_{,2} - \psi^j{}_{,1}\, \psi^k{}_{,2}\, \psi^j{}_{,2}\, \psi^k{}_{,1} \big) \\
&= g_{11}g_{22} - g_{12}g_{21} = \det(g) = \mathrm{g}\,.
\end{aligned}$$

Finally, (3.8) is concluded from the fact that g is positive definite. $\square$

Remark 3.12. A sequence $(k_1, k_2, \cdots, k_n)$ of positive integers not exceeding n, with the property that no two of the k_i are equal, is called a ***permutation of degree*** n. The

collection of all permutations of degree n is denoted by $\mathbb{P}(n)$. For $1 \leqslant i, j \leqslant n$ and $i \neq j$, the operator $\tau_{(i,j)}$ interchange the i-th and j-th elements of a sequence in $\mathbb{P}(n)$. For example, if n = 3, the permutation $(3,1,2)$ can be obtained by interchanging pairs of $(1,2,3)$ twice:

$$(1,2,3) \xrightarrow{\tau_{(1,3)}} (3,2,1) \xrightarrow{\tau_{(2,3)}} (3,1,2);$$

thus $(3,1,2)$ is called an even permutation of $(1,2,3)$. On the other hand, $(1,3,2)$ is obtained by interchanging pairs of $(1,2,3)$ once:

$$(1,2,3) \xrightarrow{\tau_{(2,3)}} (1,3,2);$$

thus $(1,3,2)$ is an odd permutation of $(1,2,3)$.

For $n = 3$, the even and odd permutations can also be viewed as the orientation of the permutation (k_1, k_2, k_3). To be more precise, if $(1,2,3)$ is arranged in a counter-clockwise orientation (see Figure 3.1), then an even permutation of degree 3 is a permutation in the counter-clockwise orientation, while an odd permutation of degree 3 is a permutation in the clockwise orientation. From figure 3.1, it is easy to see that $(3,1,2)$ is an even permutation of degree 3 and $(1,3,2)$ is an odd permutation of degree 3.

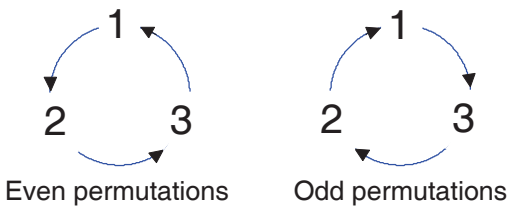


Figure 3.1: Even and odd permutations of degree 3

The permutation symbol is a function on $\mathbb{P}(n)$ defined by

$$\varepsilon_{k_1 k_2 \cdots k_n} = \begin{cases} 1 & \text{if } (k_1, k_2, \cdots, k_n) \text{ is an even permutation of } (1, 2, \cdots, n), \\ -1 & \text{if } (k_1, k_2, \cdots, k_n) \text{ is an odd permutation of } (1, 2, \cdots, n). \end{cases}$$

Example 3.13. Let Σ be the sphere centered at the origin with radius R. Consider the local parametrization $\psi(\theta, \phi) = (R\cos\theta\sin\phi, R\sin\theta\sin\phi, R\cos\phi)$ with $(\theta, \phi) \in \mathcal{V} \equiv (0, 2\pi) \times (0, \pi)$. Then

$$\begin{aligned} \psi_{,1}(\theta, \phi) \equiv \psi_\theta(\theta, \phi) &= (-R\sin\theta\sin\phi, R\cos\theta\sin\phi, 0), \\ \psi_{,2}(\theta, \phi) \equiv \psi_\phi(\theta, \phi) &= (R\cos\theta\cos\phi, R\sin\theta\cos\phi, -R\sin\phi); \end{aligned}$$

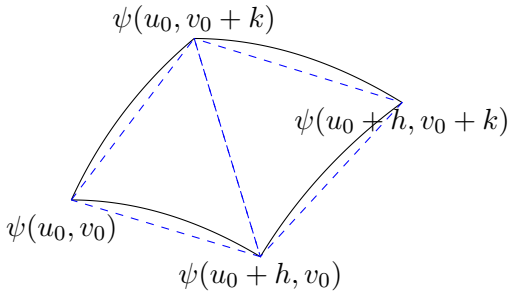
thus the metric tensor and the first fundamental form associated with the parametrization $\{\mathcal{V}, \psi\}$ are

$$g(\theta, \phi) = \big[D\psi\big]^{\mathrm{T}}\big[D\psi\big](\theta, \phi) = \begin{bmatrix} R^2\sin^2\phi & 0 \\ 0 & R^2 \end{bmatrix}$$

and $\mathrm{g} = \det(g) = R^4\sin^2\phi$.

•• What does the first fundamental form do for us?

Let $p = \psi(u_0, v_0)$ be a point in Σ. Then the surface area of the region $\psi\big([u_0, u_0 + h] \times [v_0, v_0 + k]\big)$, where h, k are very small, can be approximated by the sum of the area of two triangles, one with vertices $\psi(u_0, v_0)$, $\psi(u_0 + h, v_0)$, $\psi(u_0, v_0 + k)$ and the other with vertices $\psi(u_0 + h, v_0)$, $\psi(u_0, v_0 + k)$, $\psi(u_0 + h, v_0 + k)$.



Here we remark that the approximation of the surface area of a regular $\mathscr{C}^1$-surface obeys

$$\lim_{(h,k)\to(0,0)} \frac{\text{the surface area of } \psi\big([u_0, u_0 + h] \times [v_0, v_0 + k]\big)}{\text{the sum of area of the two triangles given in the context}} = 1\,. \tag{3.10}$$

The area of the triangle with vertices $\psi(u_0, v_0)$, $\psi(u_0 + h, v_0)$, $\psi(u_0, v_0 + k)$ is

$$A_1 = \frac{1}{2}\big\|\big(\psi(u_0 + h, v_0) - \psi(u_0, v_0)\big) \times \big(\psi(u_0, v_0 + k) - \psi(u_0, v_0)\big)\big\|_{\mathbb{R}^3}\,.$$

By the mean value theorem, for each component $j \in \{1, 2, 3\}$, we have

$$\begin{aligned}
\psi^j(u_0 + h, v_0) - \psi^j(u_0, v_0) &= \psi_{,1}\,(u_0 + \theta_1^j h, v_0)h\,,\\
\psi^j(u_0, v_0 + k) - \psi^j(u_0, v_0) &= \psi_{,2}\,(u_0, v_0 + \theta_2^j k)k
\end{aligned}$$

for some $\theta_i^j \in (0, 1)$; thus if ψ is of class $\mathscr{C}^1$,

$$\begin{aligned}
\psi(u_0 + h, v_0) - \psi(u_0, v_0) &= \psi_{,1}\,(u_0, v_0)h + \boldsymbol{E}_1(u_0, v_0; h)h\,,\\
\psi(u_0, v_0 + k) - \psi(u_0, v_0) &= \psi_{,2}\,(u_0, v_0)k + \boldsymbol{E}_2(u_0, v_0; k)k\,,
\end{aligned}$$

where $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are bounded vector-valued functions satisfying that $\lim\limits_{h\to 0} \boldsymbol{E}_1(u_0, v_0; h) = 0$ and $\lim\limits_{k\to 0} \boldsymbol{E}_2(u_0, v_0; k) = 0$. Therefore,

$$\lim_{(h,k)\to(0,0)} \frac{\big(\psi(u_0+h, v_0) - \psi(u_0, v_0)\big) \times \big(\psi(u_0, v_0+k) - \psi(u_0, v_0)\big)}{hk} - \psi_{,1}\,(u_0, v_0) \times \psi_{,2}\,(u_0, v_0) = \mathbf{0}\,.$$

Since $\sqrt{\mathrm{g}} = \|\psi_{,1} \times \psi_{,2}\|_{\mathbb{R}^3}$, we have

$$A_1 = \frac{1}{2}\sqrt{\mathrm{g}(u_0, v_0)}hk + f_1(u_0, v_0; h, k)hk$$

for some function f_1 which converges to 0 as $(h, k) \to (0, 0)$ and is bounded since $\nabla\psi$ is bounded. Similarly, the area of the triangle with vertices $\psi(u_0 + h, v_0)$, $\psi(u_0, v_0 + k)$, $\psi(u_0 + h, v_0 + k)$ is

$$A_2 = \frac{1}{2}\sqrt{\mathrm{g}(u_0, v_0)}hk + f_2(u_0, v_0; h, k)hk\,.$$

Taking (3.10) into account, we find that

$$\text{the surface area of } \psi\big([u_0,u_0+h]\times[v_0,v_0+k]\big) = \sqrt{\mathrm{g}(u_0,v_0)}hk + f(u_0,v_0;h,k)hk \quad (3.11)$$

for some bounded function $f(\cdot,\cdot;\cdot,\cdot)$ which converges to 0 as the last two variables h,k approach 0.

Now consider the surface area of $\psi([a,a+L]\times[b,b+W])$. Let $\varepsilon>0$ be given. Choose $N>0$ such that

$$\big|f(u,v;h,k)\big| < \frac{\varepsilon}{2LW} \quad \forall\, 0<h<\frac{L}{N}, 0<k<\frac{W}{N} \text{ and } (u,v)\in[a,a+L]\times[b,b+W]\,,$$

and

$$\Big|\sum_{j=1}^{m}\sum_{i=1}^{n}\sqrt{\mathrm{g}\big(a+\tfrac{i-1}{n}L, b+\tfrac{j-1}{m}M\big)}\frac{L}{n}\frac{W}{m} - \int_{[a,a+L]\times[b,b+W]}\sqrt{\mathrm{g}}\,d\mathbb{A}\Big| < \frac{\varepsilon}{2} \quad \text{if } n,m\geqslant N\,.$$

Then for $n,m\geqslant N$, with (h,k) denoting $\big(\frac{L}{n},\frac{W}{m}\big)$ (3.11) implies that

$$\begin{aligned}
&\Big|\text{the surface area of } \psi([a,a+L]\times[b,b+W]) - \int_{[a,a+L]\times[b,b+W]}\sqrt{\mathrm{g}}\,d\mathbb{A}\Big| \\
&\quad= \Big|\sum_{j=1}^{m}\sum_{i=1}^{n}\text{the surface area of } \psi([a+(i-1)h,a+ih]\times[b+(j-1)k,b+jk]) \\
&\qquad - \int_{[a,a+L]\times[b,b+W]}\sqrt{\mathrm{g}}\,d\mathbb{A}\Big| \\
&\quad\leqslant \Big|\sum_{j=1}^{m}\sum_{i=1}^{n}\sqrt{\mathrm{g}\big(a+(i-1)h,b+(j-1)k\big)}hk - \int_{[a,a+L]\times[b,b+W]}\sqrt{\mathrm{g}}\,d\mathbb{A}\Big| \\
&\qquad+\Big|\sum_{j=1}^{m}\sum_{i=1}^{n}f(a+(i-1)h,b+(j-1)k;h,k)hk\Big| \\
&\quad< \frac{\varepsilon}{2}+\frac{\varepsilon}{2LW}\sum_{j=1}^{m}\sum_{i=1}^{n}hk=\varepsilon\,.
\end{aligned}$$

The discussion above verifies the following

Theorem 3.14. *Let $\Sigma\subseteq\mathbb{R}^3$ be a regular $\mathscr{C}^1$-surface, $\{\mathcal{V},\psi\}$ be a local $\mathscr{C}^1$-parametrization of Σ at p, and g be the first fundamental form associated with $\{\mathcal{V},\psi\}$. Then*

$$\text{the surface area of } \psi(\mathcal{V}) = \int_{\mathcal{V}}\sqrt{\mathrm{g}}\,d\mathbb{A}\,.$$

Example 3.15. Recall from Example 3.13 that the first fundamental form g of the parametrization $\{\mathcal{V},\psi\}$ of the 2-sphere centered at the origin with radius R, where

$$\psi(\theta,\phi) = (R\cos\theta\sin\phi, R\sin\theta\sin\phi, R\cos\phi)$$

and $\mathcal{V}=(0,2\pi)\times(0,\pi)$, is given by $\mathrm{g}(\theta,\phi)=R^4\sin^2\phi$. Therefore,

$$\begin{aligned}
\text{the surface area of } \psi\big((0,2\pi)\times(0,\pi)\big) &= \int_{(0,2\pi)\times(0,\pi)}R^2\sin\phi\,d(\theta,\phi) \\
&= R^2\int_0^{2\pi}\int_0^{\pi}\sin\phi\,d\phi d\theta = 4\pi R^2\,.
\end{aligned}$$

Since the difference of the 2-sphere and $\psi\big((0,2\pi)\times(0,\pi)\big)$ has zero area, we find that the surface area of the 2-sphere with radius R is $4\pi R^2$.

Example 3.16. Let $\Sigma\subseteq\mathbb{R}^3$ be the upper half sphere; that is, $\Sigma=\big\{(x,y,z)\in\mathbb{R}^3\,\big|\,x^2+y^2+z^2=R^2, z>0\big\}$, and $\{\mathcal{V},\psi\}$ be a global parametrization of Σ given by

$$\psi(u,v)=(u,v,\sqrt{R^2-u^2-v^2})\,,\quad (u,v)\in\mathcal{V}=\big\{(u,v)\in\mathbb{R}^2\,\big|\,u^2+v^2\leqslant R^2\big\}\,.$$

To find the surface area using this parametrization, we first compute $\{\psi_{,1}\,,\psi_{,2}\}$ as follows:

$$\psi_{,1}\,(u,v)=\big(1,0,\frac{-u}{\sqrt{R^2-u^2-v^2}}\big)\quad\text{and}\quad \psi_{,2}\,(u,v)=\big(0,1,\frac{-v}{\sqrt{R^2-u^2-v^2}}\big)\,,$$

thus the first fundamental form associated with the parametrization $\{\mathcal{V},\psi\}$ is

$$\begin{aligned}\mathrm{g}(u,v)&=\|\psi_{,1}\,(u,v)\times\psi_{,2}\,(u,v)\|^2_{\mathbb{R}^3}=\Big\|\big(\frac{u}{\sqrt{R^2-u^2-v^2}},\frac{v}{\sqrt{R^2-u^2-v^2}},1\big)\Big\|^2_{\mathbb{R}^3}\\&=\frac{R^2}{R^2-u^2-v^2}\,.\end{aligned}$$

Therefore, the surface area of Σ is

$$\begin{aligned}\int_\Sigma dS&=\int_{\mathcal{V}}\frac{R}{\sqrt{R^2-u^2-v^2}}\,d\mathbb{A}=\int_{-R}^{R}\int_{-\sqrt{R^2-u^2}}^{\sqrt{R^2-u^2}}\frac{R}{\sqrt{R^2-u^2-v^2}}\,dvdu\\&=R\int_{-R}^{R}\arcsin\frac{v}{\sqrt{R^2-u^2}}\Big|_{v=-\sqrt{R^2-u^2}}^{v=\sqrt{R^2-u^2}}du=R\int_{-R}^{R}\pi\,du=2\pi R^2\,.\end{aligned}$$

Note the the computation above also shows that the surface area of the sphere in $\mathbb{R}^3$ with radius R is $4\pi R^2$ which is the same as what we have conclude in Example 3.15.

Remark 3.17. The example above provides one specific way of evaluating the surface integrals: if the surface Σ is in fact a subset of the graph of a function $f:\mathcal{D}\subseteq\mathbb{R}^2\to\mathbb{R}$; that is, $\Sigma\subseteq\big\{x,y,f(x,y))\,\big|\,(x,y)\in\mathcal{D}\big\}$, then Σ has a global parametrization

$$\psi(x,y)=\big(x,y,f(x,y)\big)\,,\qquad (x,y)\in\mathcal{V}\,,$$

where $\mathcal{V}$ is the projection of Σ onto the xy-plane along the z-direction. Then the first fundamental form associated to this parametrization is

$$\mathrm{g}(x,y)=\|\psi_{,1}\,(x,y)\times\psi_{,2}\,(x,y)\|^2_{\mathbb{R}^3}=1+\Big|\frac{\partial f}{\partial x}(x,y)\Big|^2+\Big|\frac{\partial f}{\partial y}(x,y)\Big|^2;$$

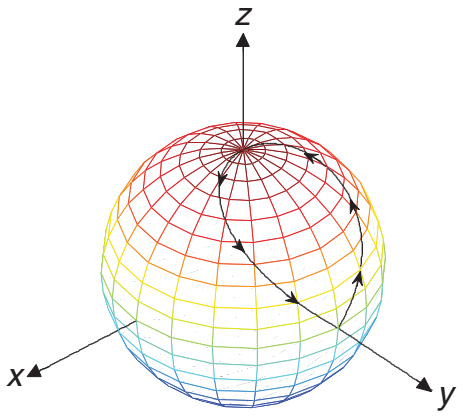
thus the surface area of Σ is

$$\int_\Sigma dS=\int_{\mathcal{V}}\sqrt{1+\Big|\frac{\partial f}{\partial x}(x,y)\Big|^2+\Big|\frac{\partial f}{\partial y}(x,y)\Big|^2}\,d(x,y)\,.$$

Example 3.18. Let C be a smooth curve parameterized by

$$\boldsymbol{r}(t)=(\cos t\sin t,\sin t\sin t,\cos t)\,,\qquad t\in\Big[-\frac{\pi}{2},\frac{\pi}{2}\Big]\,.$$

The clearly C is on the unit sphere $\mathbb{S}^2$ since $\|\boldsymbol{r}(t)\|_{\mathbb{R}^3} = 1$ for all $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Since C is a closed curve, C divides $\mathbb{S}^2$ into two parts. Let Σ denote the part with smaller area (see the following figure), and we are interested in finding the surface area of Σ.



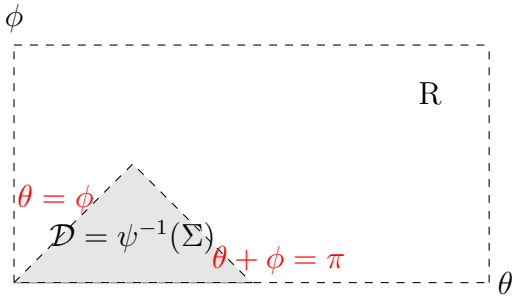
To compute the surface area of Σ, we need to find a way to parameterize Σ. Naturally we try to parameterize Σ using the spherical coordinate. In other words, let $\mathrm{R} = (0, 2\pi) \times (0, \pi)$ and $\psi : \mathrm{R} \to \mathbb{R}^3$ be defined by

$$\psi(\theta, \phi) = (\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)\,,$$

and we would like to find a region $\mathcal{D} \subseteq \mathrm{R}$ such that $\psi(\mathcal{D}) = \Sigma$.

Suppose that $\gamma(t) = \big(\theta(t), \varphi(t)\big)$, $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, is a curve in R such that $(\psi \circ \gamma)(t) = \boldsymbol{r}(t)$. Then for $t \in \left[0, \frac{\pi}{2}\right]$, the identity $\cos t = \cos\phi(t)$ implies that $\phi(t) = t$; thus the identities $\cos t \sin t = \cos\theta(t) \sin\phi(t)$ and $\sin t \sin t = \sin\theta(t) \sin\phi(t)$ further imply that $\theta(t) = t$.

On the other hand, for $t \in \left[-\frac{\pi}{2}, 0\right]$, the identity $\cos t = \cos\phi(t)$, where $\phi(t) \in (0, \pi)$, implies that $\phi(t) = -t$; thus the identities $\cos t \sin t = \cos\theta(t) \sin\phi(t)$ and $\sin t \sin t = \sin\theta(t) \sin\phi(t)$ further imply that $\theta(t) = \pi + t$.



Since the first fundamental form associate with $\{\mathrm{R}, \psi\}$ is the first fundamental form associated with $\{\mathrm{R}, \psi\}$ is

$$\begin{aligned}
\mathrm{g}(u, v) &= \big\|(\psi_\theta \times \psi_\phi)(u, v)\big\|_{\mathbb{R}^3}^2 \\
&= \big\|(-\sin\theta \sin\phi, \cos\theta \sin\phi, 0) \times (\cos\theta \cos\phi, \sin\theta \cos\phi, -\sin\phi)\big\|_{\mathbb{R}^3}^2 \\
&= \big\|(-\cos\theta \sin^2\phi, -\sin\theta \sin^2\phi, -(\sin^2\theta + \cos^2\theta) \sin\phi \cos\phi)\big\|_{\mathbb{R}^3}^2 \\
&= (\cos^2\theta + \sin^2\theta) \sin^4\phi + \sin^2\phi \cos^2\phi = \sin^2\phi\,,
\end{aligned}$$

the area of the desired surface can be computed by

$$\begin{aligned}\int_\Sigma dS &= \int_{\psi^{-1}(\Sigma)} \sqrt{\mathrm{g}}\, d\mathbb{A} = \int_0^{\frac{\pi}{2}} \int_\phi^{\pi-\phi} \sin\phi \, d\theta d\phi = \int_0^{\frac{\pi}{2}} (\pi - 2\phi)\sin\phi \, d\phi \\ &= \big(-\pi\cos\phi + 2\phi\cos\phi - 2\sin\phi\big)\Big|_{\phi=0}^{\phi=\frac{\pi}{2}} = \pi - 2\,.\end{aligned}$$

Another way to parameterize Σ is to view Σ as the graph of function $z = \sqrt{1-x^2-y^2}$ over $\mathcal{D}$, where $\mathcal{D}$ is the projection of Σ along z-axis onto xy-plane. We note that the boundary of $\mathcal{D}$ can be parameterized by

$$\widetilde{\boldsymbol{r}}(t) = (\cos t \sin t, \sin t \sin t)\,, \qquad t \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big]\,.$$

Let $(x,y) \in \partial\mathcal{D}$. Then $x^2+y^2 = y$; thus Σ can also be parameterized by $\psi : \mathcal{D} \to \mathbb{R}^3$, where

$$\psi(x,y) = \big(x, y, \sqrt{1-x^2-y^2}\big) \quad \text{and} \quad \mathcal{D} = \big\{(x,y) \,\big|\, x^2+y^2 \leqslant y\big\}\,.$$

Therefore, with f denoting the function $f(x,y) = \sqrt{1-x^2-y^2}$, Remark 3.17 implies that the surface area of Σ can be computed by

$$\begin{aligned}\int_{\mathcal{D}} \sqrt{1+f_x^2+f_y^2}\, d\mathbb{A} &= \int_0^1 \int_{-\sqrt{y-y^2}}^{\sqrt{y-y^2}} \frac{1}{\sqrt{1-x^2-y^2}}\, dxdy \\ &= \int_0^1 \arcsin\frac{x}{\sqrt{1-y^2}}\Big|_{x=-\sqrt{y-y^2}}^{x=\sqrt{y-y^2}} dy = 2\int_0^1 \arcsin\frac{\sqrt{y}}{\sqrt{1+y}}\, dy\,;\end{aligned}$$

thus making a change of variable $y = \tan^2\theta$ we conclude that

$$\begin{aligned}\text{the surface area of } \Sigma &= 2\int_0^{\frac{\pi}{4}} \arcsin\frac{\tan\theta}{\sec\theta}\, d(\tan^2\theta) = 2\int_0^{\frac{\pi}{4}} \theta\, d\big(\tan^2\theta\big) \\ &= 2\Big[\theta\tan^2\theta\Big|_{\theta=0}^{\theta=\frac{\pi}{4}} - \int_0^{\frac{\pi}{4}} \tan^2\theta d\theta\Big] \\ &= 2\Big[\frac{\pi}{4} - \int_0^{\frac{\pi}{4}} (\sec^2\theta - 1)\, d\theta\Big] = 2\Big[\frac{\pi}{4} - (\tan\theta - \theta)\Big|_{\theta=0}^{\theta=\frac{\pi}{4}}\Big] \\ &= 2\Big[\frac{\pi}{4} - \Big(1 - \frac{\pi}{4}\Big)\Big] = \pi - 2\,.\end{aligned}$$

Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface, and $\{\mathcal{V}, \psi\}$ be a parametrization of Σ such that $\psi(\mathcal{V}) = \Sigma$. If $f : \Sigma \to \mathbb{R}$ is a bounded continuous function, the surface integral of f over Σ, denoted by $\displaystyle\int_\Sigma f\, dS$, is defined by

$$\int_\Sigma f\, dS = \int_{\mathcal{V}} (f \circ \psi)\sqrt{\mathrm{g}}\, d\mathbb{A}\,. \tag{3.12}$$

In particular, if $f \equiv 1$, the number $\displaystyle\int_\Sigma dS \equiv \int_\Sigma 1\, dS$ is the surface area of Σ.

Since the surface integrals defined by (3.12) seems to depend on a given parametrization, before proceeding we show that the surface integral is indeed independent of the choice of the

parameterizations. Suppose that $\{\mathcal{V}_1, \psi_1\}$ and $\{\mathcal{V}_2, \psi_2\}$ are two local $\mathscr{C}^1$-parameterizations of a regular surface Σ at p, g_1, g_2 denote the metric tensors associated with the parameterizations $\{\mathcal{V}_1, \psi_1\}$, $\{\mathcal{V}_2, \psi_2\}$, respectively, and $\mathrm{g}_1 = \det(g_1)$, $\mathrm{g}_2 = \det(g_2)$ are corresponding first fundamental forms. Let $\Psi = \psi_2^{-1} \circ \psi_1$. Then the change of variables formula implies that

$$\int_{\mathcal{V}_2} (f \circ \psi_2)\sqrt{\mathrm{g}_2}\, d\mathbb{A} = \int_{\mathcal{V}_1} (f \circ \psi_2 \circ \Psi)\big(\sqrt{\mathrm{g}_2} \circ \Psi\big)|J_\Psi|\, d\mathbb{A} = \int_{\mathcal{V}_1} (f \circ \psi_1)\big(\sqrt{\mathrm{g}_2} \circ \Psi\big)|J_\Psi|\, d\mathbb{A}\,,$$

where J_Ψ is the Jacobian of the map Ψ. By the chain rule, we find that

$$\big[D\Psi\big]^{\mathrm{T}}\big[(D\psi_2) \circ \Psi\big]^{\mathrm{T}}\big[(D\psi_2) \circ \Psi\big]\big[D\Psi\big] = \big[D\psi_1\big]^{\mathrm{T}}\big[D\psi_1\big]\,;$$

thus by the fact that $\mathrm{g}_1 = \det\big([D\psi_1]^{\mathrm{T}}[D\psi_1]\big)$ and $\mathrm{g}_2 = \det\big([D\psi_2]^{\mathrm{T}}[D\psi_2]\big)$, we obtain that

$$\det\big([D\Psi]\big)^2(\mathrm{g}_2 \circ \Psi) = \mathrm{g}_1\,.$$

Since $J_\Psi = \det\big([D\Psi]\big)$, the identity above implies that $|J_\Psi|(\sqrt{\mathrm{g}_2} \circ \Psi) = \sqrt{\mathrm{g}_1}$, so we conclude that

$$\int_{\mathcal{V}_1} (f \circ \psi_1)\sqrt{\mathrm{g}_1}\, d\mathbb{A} = \int_{\mathcal{V}_2} (f \circ \psi_2)\sqrt{\mathrm{g}_2}\, d\mathbb{A}\,. \tag{3.13}$$

Therefore, the surface integral of f over Σ is independent of the choice of parameterizations of Σ. In particular, the surface area of a regular $\mathscr{C}^1$-surface which can be parameterized by a global parametrization is also independent of the choice of parameterizations.

Next, we study the surface area of general regular surfaces that cannot be parameterized using a single pair $\{\mathcal{V}, \psi\}$. Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface, and $\{\mathcal{V}_i, \psi_i\}_{i\in\mathcal{I}}$ be a collection of local parameterizations satisfying that for each $p \in \Sigma$ there exists $i \in \mathcal{I}$ such that $\{\mathcal{V}_i, \psi_i\}$ is a local parametrization of Σ at p. If there exists a countable collection of non-negative functions $\{\zeta_j\}_{j\in\mathcal{J}}$ defined on Σ such that

1. For each $j \in \mathcal{J}$, $\mathrm{supp}(\zeta_j) \equiv$ the closure of $\big\{x \in \Sigma \,\big|\, \zeta_j(x) \neq 0\big\} \subseteq \mathcal{V}_i$ for some $i \in \mathcal{I}$;
2. $\sum_{j\in\mathcal{J}} \zeta_j(x) = 1$ for all $x \in \Sigma$,

then intuitively we can compute the surface area by

$$\int_\Sigma dS = \sum_{j\in\mathcal{J}} \int_\Sigma \zeta_j\, dS\,, \tag{3.14}$$

where the surface integral of ζ_j over Σ is defined by (3.12) since $\mathrm{supp}(\zeta_j) \subseteq \psi(\mathcal{V}_i)$ and $\zeta_j = 0$ outside $\mathrm{supp}(\zeta_j)$. In other words, each term on the right-hand side of (3.14) can be evaluated by

$$\int_\Sigma \zeta_j\, dS = \int_{\mathcal{V}_i} (\zeta_j \circ \psi_i)\sqrt{\mathrm{g}_i}\, dS\,.$$

if $\mathrm{supp}(\zeta_j) \subseteq \psi_i(\mathcal{V}_i)$. Similarly, for a bounded continuous function f defined on Σ, the surface integral of f over Σ can be defined by

$$\int_\Sigma f\, dS = \sum_{j\in\mathcal{J}} \int_\Sigma (\zeta_j f)\, dS = \sum_{j\in\mathcal{J}} \sum_{\substack{\text{choose one } i \text{ such that} \\ \mathrm{supp}(\zeta_j) \subseteq \psi_i(\mathcal{V}_i)}} \int_{\mathcal{V}_i} (\zeta_j f) \circ \psi_i \sqrt{\mathrm{g}_i}\, dS\,. \tag{3.15}$$

Remark 3.19. Defining the surface integrals of a function as above, a question arises naturally: is the surface integral given by (3.15) independent of the choice of the parametrization and the partition-of-unity? In other words, if a regular $\mathscr{C}^k$-surface Σ admits two collections of local parametrization $\{\mathcal{U}_i, \varphi_i\}_{i\in\mathcal{I}}$ and $\{\mathcal{V}_j, \psi_j\}_{j\in\mathcal{J}}$, and $\{\zeta_i\}_{i\in\mathcal{I}}$ and $\{\lambda_j\}_{j\in\mathcal{J}}$ are $\mathscr{C}^k$-partition-of-unity subordinate to $\{\mathcal{U}_i\}_{i\in\mathcal{I}}$ and $\{\mathcal{V}_j\}_{j\in\mathcal{J}}$, respectively. Is it true that

$$\sum_{i\in\mathcal{I}} \sum_{\substack{\text{choose one } i \text{ such that} \\ \operatorname{supp}(\zeta_j)\subseteq\varphi_i(\mathcal{U}_i)}} \int_{\mathcal{U}_i} (\zeta_i f)\circ\varphi_i\sqrt{\mathrm{g}_i}\,dS = \sum_{j\in\mathcal{J}} \sum_{\substack{\text{choose one } j \text{ such that} \\ \operatorname{supp}(\lambda_k)\subseteq\psi_j(\mathcal{V}_j)}} \int_{\mathcal{V}_j} (\lambda_j f)\circ\psi_j\sqrt{g_j}\,dS\,,$$

where g_i and g_j are the first fundamental form associated with the parametrization $\{\mathcal{U}_i, \varphi_i\}$ and $\{\mathcal{V}_j, \psi_j\}$, respectively.

The answer to the question above is affirmative, and the surface integral given by (3.15) is indeed independent of the choice of parametrization of the surface and the partition-of-unity; however, we will not prove this and only treat this as a known fact.

Now we focus on the existence of a collection of functions $\{\zeta_j\}_{j\in\mathcal{J}}$ discussed above.

Definition 3.20. A collection of subsets of $\mathbb{R}^n$ is said to be ***locally finite*** if for every point $x\in\mathbb{R}^n$ there exists $r>0$ such that $B(x,r)$, the ball centered at x with radius r, intersects at most finitely many sets in this collection.

Definition 3.21 (Partition of Unity)**.** Let $A\subseteq\mathbb{R}^n$ be a subset. A collection of functions $\{\zeta_j\}_{j\in\mathcal{J}}$ is said to be a ***partition-of-unity*** of A if

1. $0\leqslant\zeta_j\leqslant 1$ for all $j\in\mathcal{J}$.

2. The collection of sets $\big\{\operatorname{supp}(\zeta_j)\big\}_{j\in\mathcal{J}}$ is locally finite.

3. $\sum\limits_{j\in\mathcal{J}}\zeta_j(x)=1$ for all $x\in A$.

Let $\{\mathcal{U}_j\}_{j\in\mathcal{J}}$ be an open cover of A; that is, $\mathcal{U}_j$ is open for all $j\in\mathcal{J}$ and $A\subseteq\bigcup_{j\in\mathcal{J}}\mathcal{U}_j$. A partition-of-unity $\{\zeta_j\}_{j\in\mathcal{J}}$ of A is said to be ***subordinate*** to $\{\mathcal{U}_j\}_{j\in\mathcal{J}}$ (or $\{\mathcal{U}_j\}_{j\in\mathcal{J}}$ has a subordinate partition-of-unity of A) if $\operatorname{supp}(\zeta_j)\subseteq\mathcal{U}_j$ for all $j\in\mathcal{J}$.

We note the if $\{\zeta_j\}_{j\in\mathcal{J}}$ is a partition-of-unity of A, then the property of local finiteness of $\{\operatorname{supp}(\zeta_j)\}_{j\in\mathcal{J}}$ ensures that for each point $x\in A$ has a neighborhood on which all but finitely many λ_j's are zero.

Lemma 3.22. *Let $A\subseteq\mathbb{R}^n$ be a subset, $\{\mathcal{U}_i\}_{i\in\mathcal{I}}$ be an open cover of A, and $\{\mathcal{V}_j\}_{j\in\mathcal{J}}$ be a collection of open sets such that each $\mathcal{V}_j$ is a subset of some $\mathcal{U}_i$; that is, for each $j\in\mathcal{J}$, $\mathcal{V}_j\subseteq\mathcal{U}_i$ for some $i\in\mathcal{I}$. If $\{\mathcal{V}_j\}_{j\in\mathcal{J}}$ has a subordinate $\mathscr{C}^k$-partition-of-unity of A, so has $\{\mathcal{U}_i\}_{i\in\mathcal{I}}$.*

Proof. Let $\{\zeta_j\}_{j\in\mathcal{J}}$ be a partition-of-unity of A subordinate to $\{\mathcal{V}_j\}_{j\in\mathcal{J}}$, and $f : \mathcal{J} \to \mathcal{I}$ be a map such that $\mathcal{V}_j \subseteq \mathcal{U}_{f(j)}$ (we note that such f in general is not unique). Define $\chi_i : \mathbb{R}^n \to [0,1]$ by

$$\chi_i(x) = \sum_{j\in f^{-1}(i)} \zeta_j(x)\,. \tag{3.16}$$

Then clearly $\operatorname{supp}(\chi_i) \subseteq \mathcal{U}_i$ and $\sum\limits_{i\in\mathcal{I}} \chi_i(x) = 1$ for all $x \in A$. Moreover, since the sum (3.16) is a finite sum, χ_i is of class $\mathscr{C}^k$ for all $i \in \mathcal{I}$ since ζ_j if of class $\mathscr{C}^k$ for all $j \in \mathcal{J}$. Now we show that $\big\{\operatorname{supp}(\chi_i)\big\}_{i\in\mathcal{I}}$ is locally finite. Let $x \in \mathbb{R}^n$ be given. By the local finiteness of $\big\{\operatorname{supp}(\zeta_j)\big\}_{j\in\mathcal{J}}$ there exists $r > 0$ such that $\#\big\{j \in \mathcal{J} \,\big|\, B(x,r) \cap \operatorname{supp}(\zeta_j) \neq \varnothing\big\} < \infty$. By the fact that $f^{-1}(i_1) \cap f^{-1}(i_2) = \varnothing$ if $i_1 \neq i_2$ (that is, each $j \in \mathcal{J}$ belongs to $f^{-1}(i)$ for exactly one $i \in \mathcal{I}$) and that

$$y \in B(x,r) \cap \operatorname{supp}(\chi_i) \quad \Leftrightarrow \quad y \in B(x,r) \cap \operatorname{supp}(\zeta_j) \text{ for some } j \in f^{-1}(i)\,,$$

we must have

$$\#\big\{i \in \mathcal{I} \,\big|\, B(x,r) \cap \operatorname{supp}(\chi_i) \neq \varnothing\big\} \leqslant \#\big\{j \in \mathcal{J} \,\big|\, B(x,r) \cap \operatorname{supp}(\zeta_j) \neq \varnothing\big\} < \infty\,. \qquad \square$$

Theorem 3.23. *Let $\Sigma \subseteq \mathbb{R}^3$ be a regular $\mathscr{C}^k$-surface. Then every open cover of Σ has a subordinate $\mathscr{C}^k$-partition-of-unity of Σ.*

Proof. Let $\{\mathcal{O}_i\}_{i\in\mathcal{I}}$ be a given open cover of Σ. Let $\{\mathcal{U}_j, \varphi_j\}_{j\in\mathcal{J}}$ be a collection of $\mathscr{C}^k$-charts of Σ such that $\{\mathcal{U}_j\}_{j\in\mathcal{J}}$ is a locally finite open cover of Σ and for each $j \in \mathcal{J}$, $\bar{\mathcal{U}}_j \subseteq \mathcal{O}_i$ for some $i \in \mathcal{I}$. By Lemma 3.22, it suffices to find a $\mathscr{C}^k$-partition-of-unity of Σ subordinate to $\{\mathcal{U}_j\}_{j\in\mathcal{J}}$.

W.L.O.G., we can assume that $\mathcal{U}_j$ and $\mathcal{V}_j \equiv \varphi(\mathcal{U}_j)$ is bounded for all $j \in \mathcal{J}$. Define $\psi_j = \varphi_j^{-1}$. Then $\{\mathcal{V}_j, \psi_j\}_{j\in\mathcal{J}}$ is a collection of local parametrization of Σ. Choose a collection of open sets $\{\mathcal{W}_j\}_{j\in\mathcal{J}}$ such that $\overline{\mathcal{W}}_j \subseteq \mathcal{V}_j$ for all $j \in \mathcal{J}$ and $\big\{\psi_j(\mathcal{W}_j)\big\}_{j\in\mathcal{J}}$ is still an open cover of Σ. For each $j \in \mathcal{J}$, let $\big\{B_k^{(j)}\big\}_{k=1}^{N_j}$ be a collection of open balls satisfying $\overline{\mathcal{W}}_j \subseteq \bigcup\limits_{k=1}^{N_j} B_k^{(j)}$ and $\operatorname{cl}(B_k^{(j)}) \subseteq \mathcal{V}_j$ for all $k \in \{1, \cdots, N_j\}$. For $j \in \mathcal{J}$ and $k \in \{1, \cdots, N_j\}$, with $c_{j,k}$ and $r_{j,k}$ denoting the center and the radius of $B_k^{(j)}$, respectively, let

$$\mu_{(j,k)}(x) = \begin{cases} \exp\Big(\dfrac{1}{\|x - c_{j,k}\|_{\mathbb{R}^2}^2 - r_{j,k}^2}\Big) & \text{if } x \in B_k^{(j)}\,, \\ 0 & \text{if } x \notin B_k^{(j)}\,, \end{cases}$$

and then define $\chi_j : \mathbb{R}^2 \to \mathbb{R}$ by $\chi_j(x) = \sum\limits_{k=1}^{N_j} \mu_{(j,k)}(x)$. Then $\chi_j > 0$ in $\overline{\mathcal{W}}_j$, and $\chi_j = 0$ outside $\bigcup\limits_{k=1}^{N_j} B_k^{(j)}$. Further define

$$\lambda_j(x) = \begin{cases} (\chi_j \circ \varphi_j)(x) & \text{if } x \in \mathcal{U}_j\,, \\ 0 & \text{if } x \in \mathcal{U}_j^{\complement}\,. \end{cases}$$

Then $\lambda_j > 0$ on $\psi_j(W_j)$ which implies that $\sum\limits_{j\in\mathcal{J}} \lambda_j > 0$. Finally, we define $\zeta_j = \dfrac{\lambda_j}{\sum_{j\in\mathcal{J}} \lambda_j}$. Then $\{\zeta_j\}_{j\in\mathcal{J}}$ is a $\mathscr{C}^k$-partition-of-unity subordinate to the open cover $\{\mathcal{U}_j\}_{j\in\mathcal{J}}$. $\square$

Definition 3.24 (Piecewise Regular Surface)**.** A surface $\Sigma \subseteq \mathbb{R}^3$ is said to be piecewise regular if there are finite many curves $C_1, \cdots, C_k$ such that $\Sigma \backslash \bigcup_{i=1}^{k} C_i$ is a disjoint union of regular surfaces.

Definition 3.25. Let $\Sigma \subseteq \mathbb{R}^3$ be a piecewise regular surface such that Σ is the disjoint union of regular surfaces Σ_i, where $i \in \mathcal{I}$ for some finite index set $\mathcal{I}$. For a continuous function $f : \Sigma \to \mathbb{R}$, the surface integral of f over Σ, still denoted by $\int_\Sigma f\, dS$, is defined by

$$\int_\Sigma f\, dS = \sum_{i \in \mathcal{I}} \int_{\Sigma_i} f\, dS\,.$$

Definition 3.26. Let $\mathscr{R}_\Sigma$ be the collection of piecewise regular surfaces in $\mathbb{R}^3$. The surface element is a set function $\mathscr{S} : \mathscr{R}_\Sigma \to \mathbb{R}$ that satisfies the following properties:

1. $\mathscr{S}(\Sigma) > 0$ for all $\Sigma \in \mathscr{R}_\Sigma$.
2. If Σ is the union of finitely many regular surfaces $\Sigma_1, \cdots, \Sigma_k$ that do not overlap except at their boundaries, then
$$\mathscr{S}(\Sigma) = \mathscr{S}(\Sigma_1) + \cdots + \mathscr{S}(\Sigma_k)\,.$$
3. The value of $\mathscr{S}$ agrees with the area on planar surfaces; that is,
$$\mathscr{S}(\mathcal{P}) = \mathbb{A}(\mathcal{P}) \qquad \text{for all planar surfaces } \mathcal{P}.$$

• The flux integral

Let $\Sigma \subseteq \mathbb{R}^3$ be a regular $\mathscr{C}^1$-surface with a continuous normal vector field $\mathbf{N} : \Sigma \to \mathbb{R}^3$, and $\boldsymbol{u} : \Sigma \to \mathbb{R}^3$ be a bounded continuous vector-valued function. The flux integral of $\boldsymbol{u}$ over Σ with given orientation $\mathbf{N}$ is the surface integral of $\boldsymbol{u} \cdot \mathbf{N}$ over Σ.

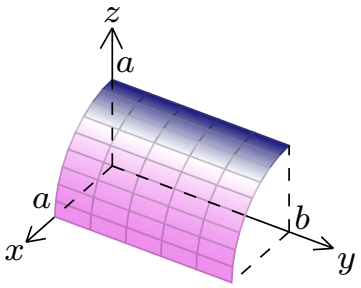
•• Physical interpretation

Let $\Omega \subseteq \mathbb{R}^3$ be an open set which stands for a fluid container and fully contains some liquid such as water, and $\boldsymbol{u} : \Omega \to \mathbb{R}^3$ be a vector-field which stands for the fluid velocity; that is, $\boldsymbol{u}(x)$ is the fluid velocity at point $x \in \Omega$. Furthermore, let $\Sigma \subseteq \Omega$ be a surface immersed in the fluid with given orientation $\mathbf{N}$, and $c : \Omega \to \mathbb{R}$ be the concentration of certain material dissolving in the liquid. Then the amount of the material carried across the surface in the direction $\mathbf{N}$ by the fluid in a time period of Δt is

$$\Delta t \cdot \int_\Sigma c\boldsymbol{u} \cdot \mathbf{N}\, dS\,.$$

Therefore, $\int_\Sigma c\boldsymbol{u} \cdot \mathbf{N}\, dS$ is the instantaneous amount of the material carried across the surface in the direction $\mathbf{N}$ by the fluid.

Example 3.27. Find the flux integral of the vector field $\boldsymbol{F}(x,y,z) = (x, y^2, z)$ upward through the first octant part Σ of the cylindrical surface $x^2 + z^2 = a^2$, $0 < y < b$.



Fist, we parameterize Σ by

$$\psi(u,v) = (u, v, \sqrt{a^2 - u^2}), \quad (u,v) \in \mathcal{V} = (0,a) \times (0,b)\,.$$

Since the first fundamental form g associated with $\{\mathcal{V}, \psi\}$ is $\mathrm{g} = \|\psi_{,1} \times \psi_{,2}\|^2_{\mathbb{R}^3} = \dfrac{a^2}{a^2 - u^2}$, and the upward-pointing unit normal is $\mathbf{N}(x,y,z) = (\dfrac{x}{a}, 0, \dfrac{z}{a})$, we have

$$\begin{aligned}\int_\Sigma \boldsymbol{F} \cdot \mathbf{N}\, dS &= \int_{\mathcal{V}} \frac{1}{a}(u^2 + a^2 - u^2)\frac{a}{\sqrt{a^2-u^2}}\, d(u,v) = a^2 \int_{\mathcal{V}} \frac{1}{\sqrt{a^2-u^2}}\, d(u,v) \\ &= a^2 \int_0^b \int_0^a \frac{1}{\sqrt{a^2-u^2}}\, du dv = a^2 b \arcsin\frac{u}{a}\Big|_{u=0}^{u=a} = \frac{\pi a^2 b}{2}\,.\end{aligned}$$

•• Measurements of the flux - the divergence operator

Let $\Omega \subseteq \mathbb{R}^3$ be an open set, and $\boldsymbol{u} = (u^1, u^2, u^3) : \Omega \to \mathbb{R}^3$ be a $\mathscr{C}^1$ vector field. Suppose that $\mathcal{O}$ is a bounded open set of class $\mathscr{C}^1$ such that $\bar{\mathcal{O}} \subseteq \Omega$ with outward-pointing unit normal vector field $\mathbf{N} = (\mathrm{N}_1, \mathrm{N}_2, \mathrm{N}_3)$. Then the flux integral of $\boldsymbol{u}$ over $\partial\mathcal{O}$ in the direction $\mathbf{N}$ is

$$\int_{\partial\mathcal{O}} \boldsymbol{u} \cdot \mathbf{N}\, dS\,.$$

Consider a special case that $\mathcal{O} = B(a,r)$ for some ball in $\mathbb{R}^3$ centered at $a = (a_1, a_2, a_3)$ with radius $r > 0$. We first compute $\displaystyle\int_{\partial B(a,r)} u^3 \mathrm{N}_3\, dS$. Consider

$$\begin{aligned}\psi_+(x_1,x_2) &= \big(x_1, x_2, a_3 + \sqrt{r^2 - (x_1-a)^2 - (x_2-a_2)^2}\big)\,, \qquad (x_1,x_2) \in D(a,r)\,, \\ \psi_-(x_2,x_2) &= \big(x_1, x_2, a_3 - \sqrt{r^2 - (x_1-a)^2 - (x_2-a_2)^2}\big)\,, \qquad (x_1,x_2) \in D(a,r)\,,\end{aligned}$$

where $D(a,r)$ is the disk in $\mathbb{R}^2$ given by $\big\{(x_1,x_2) \in \mathbb{R}^2 \,\big|\, (x_1-a_1)^2 + (x_2-a_2)^2 \leqslant r^2\big\}$. Since $\partial B(a,r) \backslash \big(\psi_+(D(a,r)) \cup \psi_-(D(a,r))$ is the equator of the sphere $\partial B(a,r)$ which has zero area, we must have

$$\int_{\partial B(a,r)} u^3 \mathrm{N}_3\, dS = \int_{\psi_+(D(a,r))} u^3 \mathrm{N}_3\, dS + \int_{\psi_-(D(a,r))} u^3 \mathrm{N}_3\, dS\,.$$

Note that $(\mathbf{N}\circ\psi_\pm)(x_1,x_2)=\dfrac{1}{r}\big(\psi_\pm(x_1,x_2)-a\big)$. In view of Example 3.16, we have

$$\begin{aligned}
&\int_{\psi_+(D(a,r))} u^3\mathrm{N}_3\,dS\\
&\quad=\int_{D(a,r)} u^3(\psi_+(x_1,x_2))\frac{\sqrt{r^2-(x_1-a_1)^2-(x_2-a_2)^2}}{r}\frac{r}{\sqrt{r^2-(x_1-a_1)^2-(x_2-a_2)^2}}\,d\mathbb{A}\\
&\quad=\int_{D(a,r)} u^3(\psi_+(x_1,x_2))\,d\mathbb{A}\,,
\end{aligned}$$

and similarly,

$$\int_{\psi_+(D(a,r))} u^3\mathrm{N}_3\,dS=-\int_{D(a,r)} u^3(\psi_-(x_1,x_2))\,d\mathbb{A}\,.$$

Therefore,

$$\begin{aligned}
\int_{\partial B(a,r)} u^3\mathrm{N}_3\,dS&=\int_{D(a,r)}\big[u^3(\psi_+(x_1,x_2))-u^3(\psi_-(x_1,x_2))\big]\,d\mathbb{A}\\
&=\int_{D(a,r)}\Big(\int_{a_3-\sqrt{r^2-(x_1-a_1)^2-(x_2-a_2)^2}}^{a_3+\sqrt{r^2-(x_1-a_1)^2-(x_2-a_2)^2}}\frac{\partial u^3}{\partial x_3}(x_1,x_2,x_3)\,dx_3\Big)d\,\mathbb{A}\\
&=\int_{B(a,r)}\frac{\partial u^3}{\partial x_3}\,dx\,.
\end{aligned}$$

Similarly,

$$\int_{\partial B(a,r)} u^1\mathrm{N}_1\,dS=\int_{B(a,r)}\frac{\partial u^1}{\partial x_1}\,dx\quad\text{and}\quad\int_{\partial B(a,r)} u^2\mathrm{N}_2\,dS=\int_{B(a,r)}\frac{\partial u^2}{\partial x_2}\,dx\,;$$

thus we conclude that

$$\int_{\partial B(a,r)}\boldsymbol{u}\cdot\mathbf{N}\,dS=\int_{B(a,r)}\sum_{i=1}^{3}\frac{\partial u^i}{\partial x_i}\,dx\,. \tag{3.17}$$

The computation above motivates the following

Definition 3.28 (The divergence operator)**.** Let $\boldsymbol{u}:\Omega\subseteq\mathbb{R}^{\mathrm{n}}\to\mathbb{R}^{\mathrm{n}}$ be a $\mathscr{C}^1$ vector field. The divergence of $\boldsymbol{u}$ is a scalar function defined by

$$\mathrm{div}\boldsymbol{u}=\sum_{i=1}^{\mathrm{n}}\frac{\partial u^i}{\partial x_i}\,.$$

Definition 3.29. A vector field $\boldsymbol{u}:\Omega\subseteq\mathbb{R}^n\to\mathbb{R}^n$ is called *solenoidal* or *divergence-free* if $\mathrm{div}\boldsymbol{u}=0$ in Ω.

Remark 3.30. Let $\Omega\subseteq\mathbb{R}^3$ be an open set, and $\boldsymbol{u}:\Omega\to\mathbb{R}^3$ be a $\mathscr{C}^1$ vector field. Using (3.17), by the continuity of $\mathrm{div}\boldsymbol{u}$ we conclude that

$$\lim_{r\to0}\frac{1}{|B(a,r)|}\int_{\partial\,B(a,r)}\boldsymbol{u}\cdot\mathbf{N}\,dS=(\mathrm{div}\boldsymbol{u})(a)\qquad\forall\,a\in\Omega\,,$$

where $|B(a,r)|=\dfrac{4\pi r^3}{3}$ is the volume of $B(a,r)$. In other words, $\mathrm{div}\boldsymbol{u}$ at a point x is the instantaneous amount (per volume) of material (with concentration 1) carried outside an infinitesimal ball centered at x.

• **The divergence theorem**

Theorem 3.31 (The Divergence Theorem). *Let $\Omega \subseteq \mathbb{R}^n$ be a bounded domain such that $\partial\Omega$ is piecewise smooth, and $\boldsymbol{w} = (w_1, \cdots, w_n) \in \mathscr{C}^1(\overline{\Omega})$ with outward pointing normal $\mathbf{N}$. Then*

$$\int_\Omega \operatorname{div}\boldsymbol{w}\, dx = \int_{\partial\Omega} \boldsymbol{w}\cdot\mathbf{N}\, dS\,.$$

3.2.2 Equation of continuity

Let u be the concentration of some physical quantity ($u = u(x,t)$) in a domain $\Omega \subseteq \mathbb{R}^n$, and let $\boldsymbol{F}$ be the flux of the quantity; that is, $\boldsymbol{F}\cdot\mathbf{N}\, dS$ is the flow rate of the quantity that passes through an area dS in the direction $\mathbf{N}$ (outward pointing) normal to dS. Then

$$\frac{d}{dt}\int_{\mathcal{U}} u\, dx = -\int_{\partial\mathcal{U}} \boldsymbol{F}\cdot\mathbf{N}\, dS + \int_{\mathcal{U}} q\, dx \quad \text{for all } \mathcal{U}\subseteq\Omega\,,$$

where q is the strength of sources for the quantity. If u is smooth, by the divergence theorem,

$$\int_{\mathcal{U}} u_t\, dx = \int_{\mathcal{U}} (q - \operatorname{div}\boldsymbol{F})\, dx \;\Rightarrow\; \int_{\mathcal{U}} \big[u_t + \operatorname{div}\boldsymbol{F} - q\big]\, dx = 0$$

for all open domains $\mathcal{U}$ with piecewise smooth boundary $\partial\mathcal{U}$. We then obtain *the equation of continuity* $u_t + \operatorname{div}\boldsymbol{F} = q$.

• **The conservation of mass**

Let $\varrho(x,t)$ and $\boldsymbol{u}(x,t)$ denote the density and the velocity of a fluid at point x at time t. Then the density flux $\boldsymbol{F} = \rho\boldsymbol{u}$, and the equation of continuity reads

$$\varrho_t + \operatorname{div}(\varrho\boldsymbol{u}) = 0 \qquad \forall\, x\in\Omega\,, t\in\mathbb{R}\,. \tag{3.18}$$

In particular, if the density of a fluid is constant (incompressible fluid), then the velocity $\boldsymbol{u}$ of this fluid must satisfy

$$\operatorname{div}\boldsymbol{u} = 0 \qquad \text{in} \quad \Omega\,. \tag{3.19}$$

3.2.3 The heat equations

Let $\vartheta(x,t)$ defined on $\Omega\times(0,T]$ be the temperature of a material body at point $x\in\Omega$ at time $t\in(0,T]$, and $c(x)$, $\varrho(x)$, $k(x)$ be the specific heat, density, and the inner thermal conductivity of the material body at x. Then by the conservation of heat, for any open set $\mathcal{U}\subseteq\Omega$,

$$\frac{d}{dt}\int_{\mathcal{U}} c(x)\varrho(x)\vartheta(x,t)\, dx = \int_{\partial\mathcal{U}} k(x)\nabla\vartheta(x,t)\cdot\mathbf{N}(x)\, dS\,, \tag{3.20}$$

where $\mathbf{N}$ denotes the outward-pointing unit normal of $\mathcal{U}$. Assume that u is smooth, and $\mathcal{U}$ is a domain with piecewise smooth boundary. By the divergence theorem, (3.20) implies

$$\int_{\mathcal{U}} c(x)\varrho(x)\vartheta_t(x,t)\, dx = \int_{\mathcal{U}} \operatorname{div}\big(k(x)\nabla\vartheta(x,t)\big)\, dx\,.$$

Since $\mathcal{U}$ is arbitrary, the equation above implies

$$c(x)\varrho(x)\vartheta_t(x,t) - \operatorname{div}(k(x)\nabla\vartheta(x,t)) = 0 \quad \forall\, x \in \Omega\,, t \in (0,T].$$

If k is constant, then

$$\frac{c\varrho}{k}\vartheta_t = \Delta\vartheta \equiv \sum_{i=1}^{\mathrm{n}} \frac{\partial^2\vartheta}{\partial x_i^2}\,.$$

If furthermore c and ϱ are constants, then after rescaling of time we have

$$\vartheta_t = \Delta\vartheta\,. \tag{3.21}$$

This is the standard **heat equation**, the prototype equation of **parabolic** equations.

We need complementary conditions to specify a particular solution of (3.21):

1. *Initial condition*: $\vartheta(x,0) = \vartheta_0(x)$, where $\vartheta_0(x)$ is a given function.

2. *Boundary condition*: if $\partial\Omega \neq \varnothing$, some boundary condition of u at $x \in \partial\Omega$ for all time have to be introduced by physical reason to specify a unique solution.

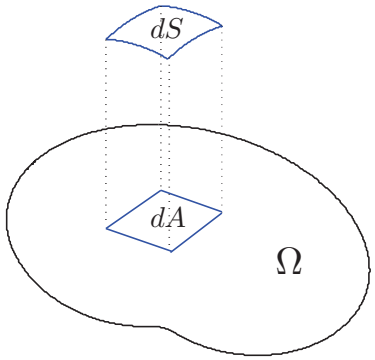
   (a) *Dirichlet condition*: $\vartheta(x,t) = g(x,t)$ for all $x \in \partial\Omega$ and $t \geqslant 0$, where g is a given function.

   (b) *Neumann condition*: $\dfrac{\partial\vartheta}{\partial\mathbf{N}} = 0$ for all $x \in \partial\Omega$ and $t \geqslant 0$, where $\dfrac{\partial\vartheta}{\partial\mathbf{N}} \equiv \mathbf{N}\cdot\nabla\vartheta$ and g is a given function.

   (c) *Robin condition*: $\dfrac{\partial\vartheta}{\partial\mathbf{N}} + hu = g$ for all $x \in \partial\Omega$ and $t \geqslant 0$, where h and g are given functions.

3.2.4 The wave equations

Consider the membrane (of a drum) as a graph of a function $z = u(x_1, x_2)$ for $(x_1, x_2) \in \Omega$.



Suppose that the energy stored in the membrane is given by

$$E(u) = \int_\Omega T\Big(\frac{dS}{d\mathbb{A}} - 1\Big)d\mathbb{A} = \int_\Omega T\big(\sqrt{1+|\nabla u|^2} - 1\big)\,d\mathbb{A}\,,$$

where T is called the tension of a membrane. In other words, to deform a membrane from its unforced equilibrium state to a surface S given by $z = u(x_1, x_2)$ requires the input of the energy shown above.

Question: If the deformation of the membrane is due to a small external force f, what is the relation between f and u?

Suppose that an small external force $\triangle f = \triangle f(x_1, x_2)$ is suddenly exerted onto the membrane (so that the total force added on the membrane is $f + \triangle f$), and the membrane deforms to the surface $z = (u + \triangle u)(x_1, x_2)$ slowly. We note that $\triangle f$ is a function of $\triangle u$ and $\triangle f \to 0$ as $\triangle u \to 0$. Then the extra energy needed to deform the membrane is $E(u + \triangle u) - E(u)$, while this extra work is done by the force $f + \triangle f$ given by

$$\int_\Omega (f + \triangle f)\triangle u \, dx \, .$$

Therefore,

$$E(u + \triangle u) - E(u) = \int_\Omega (f + \triangle f)\triangle u \, dx \, .$$

Let φ be a given $\mathscr{C}^1$ function and $\triangle u = t\varphi$. Then if $t \neq 0$,

$$\frac{E(u + t\varphi) - E(u)}{t} = \int_\Omega (f + \triangle f)\varphi \, dx \, .$$

Since $\triangle f \to 0$ as $t \to 0$, we find that

$$\lim_{t\to 0} \frac{E(u + t\varphi) - E(u)}{t} = \int_\Omega f\varphi \, dx \, . \tag{3.22}$$

On the other hand, assuming that u is a smooth function,

$$\begin{aligned}
\delta E(u;\varphi) \equiv \lim_{t\to 0} \frac{E(u + t\varphi) - E(u)}{t} &= \lim_{t\to 0} \int_\Omega T \frac{\sqrt{1 + |\nabla u + t\nabla\varphi|^2} - \sqrt{1 + |\nabla u|^2}}{t} \, d\mathbb{A} \\
&= \int_\Omega T \Big(\frac{\partial}{\partial t}\Big|_{t=0} \sqrt{1 + |\nabla u + t\nabla\varphi|^2} \Big) \, d\mathbb{A} = \int_\Omega T \frac{\nabla u \cdot \nabla\varphi}{\sqrt{1 + |\nabla u|^2}} \, d\mathbb{A} \\
&= \int_\Omega \operatorname{div}\Big(\frac{T\varphi\nabla u}{\sqrt{1 + |\nabla u|^2}} \Big) \, d\mathbb{A} - \int_\Omega \varphi \operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1 + |\nabla u|^2}} \Big) \, d\mathbb{A} \, .
\end{aligned}$$

where we have used $\operatorname{div}(\varphi \boldsymbol{F}) = \varphi \operatorname{div}\boldsymbol{F} + \boldsymbol{F} \cdot \nabla\varphi$ to conclude the last equality. By the divergence theorem, with $\mathbf{N}$ denoting the outward-pointing unit normal on $\partial\Omega$,

$$\delta E(u;\varphi) = \int_{\partial\Omega} \frac{T\varphi\nabla u}{\sqrt{1 + |\nabla u|^2}} \cdot \mathbf{N} \, d\mathbb{A} - \int_\Omega \varphi \operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1 + |\nabla u|^2}} \Big) \, d\mathbb{A} \, ;$$

thus (3.22) implies that

$$\int_\Omega \Big[\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1 + |\nabla u|^2}} \Big) + f \Big] \varphi \, d\mathbb{A} - \int_{\partial\Omega} \frac{T}{\sqrt{1 + |\nabla u|^2}} \frac{\partial u}{\partial \mathbf{N}} \varphi \, d\mathbb{A} = 0 \quad \text{for all } \mathscr{C}^1\text{-function } \varphi \, . \tag{3.23}$$

In particular,

$$\int_\Omega \Big[\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1 + |\nabla u|^2}} \Big) + f \Big] \varphi \, d\mathbb{A} = 0 \quad \text{for all } \mathscr{C}^1\text{-function } \varphi \text{ that vanishes on } \partial\Omega \, . \tag{3.24}$$

The above identity implies that

$$\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)+f=0 \qquad \text{in} \quad \Omega\,. \tag{3.25}$$

Therefore,

1. If the membrane is constrained on the boundary; that is, the boundary of the membrane is fixed (for example, $u=0$ on $\partial\Omega$), then u satisfies that

$$-\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)=f \qquad \text{in} \quad \Omega\,, \tag{3.26a}$$

$$u=0 \qquad \text{on} \quad \partial\Omega\,. \tag{3.26b}$$

2. If the membrane is not constrained on the boundary (such as the banners), then (3.23) and (3.25) imply that

$$\int_{\partial\Omega}\frac{T}{\sqrt{1+|\nabla u|^2}}\frac{\partial u}{\partial \mathbf{N}}\varphi\, d\mathbb{A}=0 \quad \text{for all } \mathscr{C}^1\text{-function } \varphi\,.$$

Therefore, $\dfrac{\partial u}{\partial \mathbf{N}}=0$ on $\partial\Omega$ (where we assume that $T>0$ everywhere) which shows that u satisfies

$$-\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)=f \qquad \text{in} \quad \Omega\,, \tag{3.27a}$$

$$\frac{\partial u}{\partial \mathbf{N}}=0 \qquad \text{on} \quad \partial\Omega\,. \tag{3.27b}$$

Remark 3.32. If $u=0$ on the boundary, we will not have an extra boundary condition (3.27b) (even though at the first glance it seems the case) since if $u=0$ on $\partial\Omega$, then all possible displacement ${}_{\vartriangle}u$ should also satisfy that ${}_{\vartriangle}u=0$ on $\partial\Omega$; thus φ also has to vanish on $\partial\Omega$ in the derivation of (3.23). In other words, if the membrane is constrained, instead of (3.23) we should obtain (3.24) directly.

Remark 3.33. By expanding the derivatives, we find that

$$\begin{aligned}\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)&=\frac{\operatorname{div}(T\nabla u)}{\sqrt{1+|\nabla u|^2}}+T\nabla u\cdot\nabla\frac{1}{\sqrt{1+|\nabla u|^2}}\\&=\frac{\operatorname{div}(T\nabla u)}{\sqrt{1+|\nabla u|^2}}-T\sum_{i,j=1}^{2}\frac{u_{x_i}u_{x_j}u_{x_ix_j}}{\sqrt{1+|\nabla u|^2}^3}\,.\end{aligned}$$

Therefore, if $|\nabla u|\ll 1$ (which is a valid assumption for the case of drums), we find that

$$\operatorname{div}\Big(\frac{T\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)\approx\operatorname{div}(T\nabla u)\,;$$

thus (3.26) can be approximated by

$$\begin{cases}-\operatorname{div}(T\nabla u)=f & \text{in} \quad \Omega\,,\\ u=0 & \text{on} \quad \partial\Omega\,.\end{cases} \tag{D}$$

while (3.27) can be approximated by

$$\begin{cases}-\operatorname{div}(T\nabla u)=f & \text{in} \quad \Omega\,,\\ \dfrac{\partial u}{\partial \mathbf{N}}=0 & \text{on} \quad \partial\Omega\,.\end{cases} \tag{N}$$

• **Equation for vibrating membrane**

Let T be the tension, ϱ be the density, and f be the density of the external force which may depend on x and t.

d'Alembert's principle:

$$\int_\Omega \left[-T\nabla u\cdot\nabla\varphi + (f-\varrho u_{tt})\varphi\right] dx = 0$$

for all φ compatible with the existence constraints. Therefore,

1. Membrane fastened on the boundary:

$$\begin{cases} \varrho u_{tt} - \operatorname{div}(T\nabla u) = f & \text{in} \quad \Omega\times(0,T]\,, \\ u = g & \text{on} \quad \partial\Omega\times(0,T]\,, \\ u(x,0)=g(x)\,,\ u_t(x,0)=h(x) & \text{for all } x\in\Omega\,. \end{cases}$$

2. Membrane with free boundary:

$$\begin{cases} \varrho u_{tt} - \operatorname{div}(T\nabla u) = f & \text{in} \quad \Omega\times(0,T]\,, \\ \dfrac{\partial u}{\partial \mathbf{N}} = 0 & \text{on} \quad \partial\Omega\times(0,T]\,, \\ u(x,0)=g(x)\,,\ u_t(x,0)=h(x) & \text{for all } x\in\Omega\,. \end{cases}$$

3.2.5 The Navier-Stokes equations

Aside from the equation of continuity (3.18), at least an equation for the fluid velocity $\boldsymbol{u}$ is required to complete the system. Consider that conservation of momentum $\boldsymbol{m}=\varrho\boldsymbol{u}$. By the fact that the rate of change of momentum of a body is equal to the resultant force acting on the body, the conservation of momentum states that

$$\frac{d}{dt}\int_{\mathcal{U}} \boldsymbol{m}\,dx = -\int_{\partial\mathcal{U}} \boldsymbol{m}(\boldsymbol{u}\cdot\mathbf{N})\,dS + \int_{\partial\mathcal{U}} \boldsymbol{\sigma}\,dS + \int_{\mathcal{U}} \boldsymbol{f}\,dx\,, \tag{3.28}$$

where $\mathbf{N}$ is the outward-pointing unit normal of $\partial\mathcal{U}$, $\boldsymbol{f}$ is the external force (such as the gravity) on the fluid system, and $\boldsymbol{\sigma}$ is the stress (應力) exerted by the fluid given by

$$\boldsymbol{\sigma} = 2\mu\mathrm{Def}\boldsymbol{u}\mathbf{N} - p\mathbf{N}\,,$$

where μ is called the dynamical viscosity (which may depend on $\boldsymbol{u}$) and $\mathrm{Def}\boldsymbol{u}$, called the rate of strain tensor, is the symmetric part of the gradient of $\boldsymbol{u}$ given by

$$(\mathrm{Def}\boldsymbol{u})_{ij} = \frac{1}{2}\Big(\frac{\partial u^i}{\partial x_j} + \frac{\partial u^j}{\partial x_i}\Big)\,.$$

In other words, if $\boldsymbol{\sigma}=(\sigma_1,\sigma_2,\sigma_3)$, then

$$\sigma_i = \mu\sum_{j=1}^{3}\Big(\frac{\partial u^i}{\partial x_j} + \frac{\partial u^j}{\partial x_i}\Big)\mathrm{N}_j - p\mathrm{N}_i\,.$$

Assuming the smoothness of the variables, (3.28) and the divergence theorem imply that for each $1 \leqslant i \leqslant 3$,

$$\int_{\mathcal{U}} \Big[m_t^i + \sum_{j=1}^{\mathrm{n}} \frac{\partial (m^i u^j)}{\partial x_j} + \frac{\partial p}{\partial x_i} - \sum_{j=1}^{3} \frac{\partial}{\partial x_j} \Big[\mu \Big(\frac{\partial u^i}{\partial x_j} + \frac{\partial u^j}{\partial x_i} \Big) \Big] + f_i \Big] \, dx = 0$$

for all regular domain $\mathcal{U} \subseteq \Omega$. As a consequence, we obtain the momentum equation

$$(\varrho \boldsymbol{u})_t + \mathrm{div}(\varrho \boldsymbol{u} \otimes \boldsymbol{u}) + \nabla p = \mathrm{div}(\mu \mathrm{Def} \boldsymbol{u}) + \boldsymbol{f} \qquad \text{in} \quad \Omega \times (0, \infty) \,, \tag{3.29}$$

where for a matrix $a = [a_{ij}]$, $(\mathrm{div} a)_i \equiv \sum_{j=1}^{3} \frac{\partial a_{ij}}{\partial x_j}$.

- **Newtonian and non-Newtonain fluids**

  1. Newtonian fluids: the viscosity μ is a constant.
  2. Non-Newtonian fluids: the viscosity μ is a function of $\boldsymbol{u}$.

Consider the Newtonian case. If the fluids under consideration is incompressible, we let $\varrho = 1$ and (3.19) and (3.29) together imply the Navier-Stokes equations

$$\boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla p = \mu \Delta \boldsymbol{u} + \boldsymbol{f} \qquad \text{in} \quad \Omega \times (0, T) \,, \tag{3.30a}$$

$$\mathrm{div} \boldsymbol{u} = 0 \qquad \text{in} \quad \Omega \times (0, T) \,, \tag{3.30b}$$

where we have used the incompressibility condition (3.19) to deduce that

$$\sum_{j=1}^{3} \frac{\partial}{\partial x_j} \Big[\mu \Big(\frac{\partial u^i}{\partial x_j} + \frac{\partial u^j}{\partial x_i} \Big) \Big] = \mu \sum_{j=1}^{3} \frac{\partial}{\partial x_j} \Big(\frac{\partial u^i}{\partial x_j} + \frac{\partial u^j}{\partial x_i} \Big) = \mu \sum_{j=1}^{3} \frac{\partial^2 u^i}{\partial x_j^2} = \mu \Delta u^i \,.$$

Initial conditions: $\boldsymbol{u}(x, 0) = \boldsymbol{u}_0(x)$ for all $x \in \Omega$.
Boundary condition:

1. **No-slip boundary condition**: $\boldsymbol{u} = \boldsymbol{0}$ on $\partial\Omega$.
2. **Navier boundary condition**: $\boldsymbol{u} \cdot \mathbf{N} = 0$ and $\mathbf{N} \times (\mu \mathrm{Def} \boldsymbol{u} \mathbf{N}) = \alpha(\mathbf{N} \times \boldsymbol{u})$ on $\partial\Omega$ for some constant $\alpha > 0$. This condition is based on the assumption that the traction force due to the viscous effect is proportional to the fluid velocity on the boundary.

- **Some brief introduction about stress/traction**

  - What is the stress/traction?

  Let Σ be a small piece of surface centered at x with area δA and "outward-pointing" unit normal $\boldsymbol{n}$. The stress exerted by the fluid on the side toward which $\boldsymbol{n}$ points on the surface Σ ($\boldsymbol{n}$ 所指向的這一側的流體對曲面 Σ 所施的應力) is defined as

  $$\boldsymbol{\sigma}(x, t, \boldsymbol{n}) = \lim_{\delta A \to 0} \frac{\boldsymbol{\delta F}}{\delta A} \,,$$

  where $\boldsymbol{\delta F}$ is the force exerted on the surface by the fluid on that side (only one side is involved).

- General properties of the stress:
  1. For a unit vector $\boldsymbol{n} = (n_1, n_2, n_3)$, $\boldsymbol{\sigma}(x, t, -\boldsymbol{n}) = -\boldsymbol{\sigma}(x, t, \boldsymbol{n})$.
  2. At a given point x, suppose that $\sigma(x, t, \mathbf{e}_j) = \tau_{1j}\mathbf{e}_1 + \tau_{2j}\mathbf{e}_2 + \tau_{3j}\mathbf{e}_3$ for $1 \leqslant j \leqslant 3$, where $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the standard basis of $\mathbb{R}^3$ and $\tau_{ij} = \tau_{ij}(x, t)$. Then
  $$\boldsymbol{\sigma}(x, t, \boldsymbol{n}) = \boldsymbol{\sigma}(x, t, \mathbf{e}_1)n_1 + \boldsymbol{\sigma}(x, t, \mathbf{e}_2)n_2 + \boldsymbol{\sigma}(x, t, \mathbf{e}_3)n_3 = \Big(\sum_{i,j=1}^{3} \tau_{ij}n_j\Big)\mathbf{e}_i \qquad (\star)$$
  or equivalently,
  $$\boldsymbol{\sigma}(x, t, \boldsymbol{n}) = \begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}.$$
  3. By the conservation of angular momentum, $\tau_{ij} = \tau_{ji}$ for all $1 \leqslant i, j \leqslant 3$. In other words, the matrix (called the stress tensor) $\tau = [\tau_{ij}]$ is symmetric.

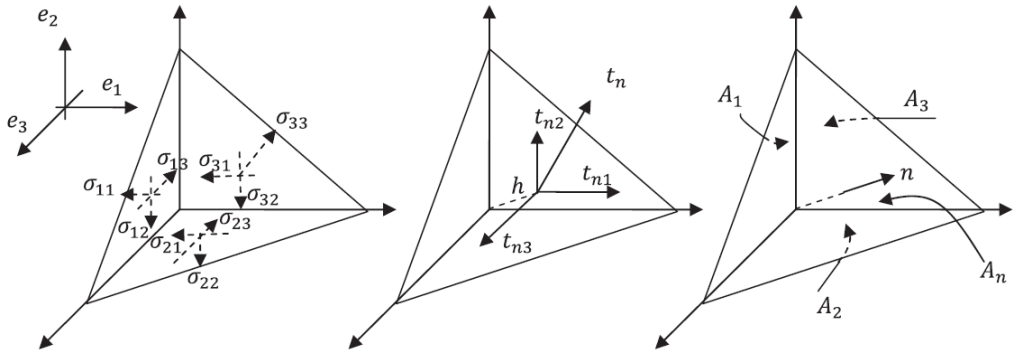


Figure 3.2: (a) On each side orthogonal to the coordinate axis, the stress is given by $\sigma(-\mathbf{e}_i) = \sum_{j=1}^{3} \sigma_{ij}\mathbf{e}_j$. (b) On the "slant" side, the stress is given by $\sigma(\boldsymbol{n}) = t_n = t_{n1}\mathbf{e}_1 + t_{n2}\mathbf{e}_2 + t_{n3}\mathbf{e}_3$. (c) By force balances, $\sigma(\boldsymbol{n})A_n = \sigma(\mathbf{e}_1)A_1 + \sigma(\mathbf{e}_2)A_2 + \sigma(\mathbf{e}_3)A_3$ which leads to $(\star)$.

- The reason why $2\mu\mathrm{Def}\boldsymbol{u}\mathbf{N}$ appears in the expression of $\boldsymbol{\sigma}$:
  1. Suppose that Σ is the xy-plane, $\mathbf{n} = (0, 0, 1)$, and $\boldsymbol{u} = (u, 0, 0)$. The larger the value $\dfrac{\partial u}{\partial x_3}$, the larger the traction due to the fluid; thus the traction should be proportion to $\dfrac{\partial u}{\partial x_3}$. Suppose that the traction, without considering the effect of pressure, is $\mu\dfrac{\partial u}{\partial x_3}$. Then $\boldsymbol{\sigma} = \mu\dfrac{\partial u}{\partial x_3}\mathbf{e}_1$.
  2. If $\boldsymbol{n} = (0, 0, 1)$ but instead $\boldsymbol{u} = (u, v, 0)$, choose a constant unit vector such that $\boldsymbol{u} = (\boldsymbol{u} \cdot \widehat{\mathbf{e}}_1)\widehat{\mathbf{e}}_1$, then $\boldsymbol{\sigma} = \mu\dfrac{\partial(\boldsymbol{u} \cdot \widehat{\mathbf{e}}_1)}{\partial x_3}\widehat{\mathbf{e}}_1 = \mu\dfrac{\partial \boldsymbol{u}}{\partial x_3}$.
  3. When $\boldsymbol{n}$ is arbitrary, by the fact that $\dfrac{\partial}{\partial x_3}$ is the directional derivative in the direction $\boldsymbol{n}$ when $\boldsymbol{n} = (0, 0, 1)$, it is naive to imagine that $\boldsymbol{\sigma} = \mu(\nabla \boldsymbol{u})\boldsymbol{n}$.
  4. Since the stress tensor has to be symmetric, we have $\boldsymbol{\sigma} = 2\mu\mathrm{Def}\boldsymbol{u}\boldsymbol{n}$.

3.3 Solving PDE using matlab®

The PDEs in the models that we derived above are of the form

$$u_t = A(u) + f \qquad \text{or} \qquad u_{tt} = A(u) + f \tag{3.31}$$

for some differential operator A; that is, for a given smooth function u, $A(u)$ is some functions of partial derivatives of u with respect to x. We are not going to talk about numerical method of solving PDEs (which is a big topic), but instead try to make use of the ODE solver (such as ode45 in matlab) which requires that we write $A(u)$ in terms of the value of u (so that the right-hand side of (3.31) can be expressed as $\varphi(x, t, u)$). We note that computers view functions as a map whose values are known on just discrete points (of interests), so to find a numerical solution u to the PDEs above is to find the "approximated" values of u on a given set of discrete points. Therefore, in order to make use of the ODE solver to solve the PDEs above, we only need to know how to compute the partial derivatives of u w.r.t. x in terms of the values of u on discrete points.

Caution: Making $A(u)$ in terms of values of u at discrete points does not always work to solve PDEs numerically!!!

- **Central differences**

Recall the Taylor Theorem that if w is a $(n+1)$-times differentiable function in x,

$$w(x+h) = \sum_{k=0}^{n} \frac{w^{(k)}(x)}{k!} h^k + \frac{w^{(n+1)}(\xi)}{(n+1)!} h^{n+1},$$

where ξ is a point between x and $x+h$. Now suppose that we are interested in the value of the solution u on the set of discrete points which consists of a regular partition $\mathcal{P} = \{0 = x_0 < x_1 < x_2 < \cdots < x_n = L\}$ of $[0, L]$. Write $\|\mathcal{P}\| = h = \frac{L}{n}$ and assume that the solution w is four times continuously differentiable in x. Then for x being one of $x_i's$,

$$\begin{aligned} w(x+h) &= w(x) + hw'(x) + \frac{h^2}{2} w''(x) + \frac{h^3}{6} w'''(x) + \mathcal{O}(h^4), \\ w(x-h) &= w(x) - hw'(x) + \frac{h^2}{2} w''(x) - \frac{h^3}{6} w'''(x) + \mathcal{O}(h^4), \end{aligned}$$

where the notation $\mathcal{O}(h^4)$ means that it is a function of h and the quotient of this function and h^4 is still bounded (when h is close to 0). More generally,

$$g(h) = \mathcal{O}(h^k) \text{ (as } h \to 0) \quad \text{if and only if} \quad \left|\frac{g(h)}{h^k}\right| \leqslant M \text{ (when } h \text{ is close to zero).}$$

Therefore,

$$\begin{aligned} w'(x) &= \frac{w(x+h) - w(x-h)}{2h} + \mathcal{O}(h^2), \\ w''(x) &= \frac{w(x+h) - 2w(x) + w(x-h)}{h^2} + \mathcal{O}(h^2). \end{aligned}$$

In other words, if w is four times continuously differentiable in x, the first and second derivatives of w at x can be made as accurate as possible using the values of w at $x \pm h$ and x by making h small enough. The finite difference scheme

$$w'(x) \approx \frac{w(x+h) - w(x-h)}{2h} \quad \text{and} \quad w''(x) \approx \frac{w(x+h) - 2w(x) + w(x-h)}{h^2} \tag{3.32}$$

of finding the approximated value of the first and second derivatives of w is called the central difference scheme.

Remark 3.34. If w is only three times continuously differentiable in x, then

$$\begin{aligned} w'(x) &= \frac{w(x+h) - w(x-h)}{2h} + \mathcal{O}(h)\,, \\ w''(x) &= \frac{w(x+h) - 2w(x) + w(x-h)}{h^2} + \mathcal{O}(h)\,. \end{aligned}$$

Remark 3.35. Let Δ_h be an operation defined by the following: if w is a function of x, then $\Delta_h w$ is a function given by

$$(\Delta_h w)(x) = \frac{w(x+h) - w(x-h)}{h}\,.$$

Then

$$\begin{aligned} (\Delta^2_{\frac{h}{2}} w)(x) \equiv (\Delta_{\frac{h}{2}} \Delta_{\frac{h}{2}} w)(x) &= \frac{(\Delta_{\frac{h}{2}} w)\big(x + \frac{h}{2}\big) - (\Delta_{\frac{h}{2}} w)\big(x - \frac{h}{2}\big)}{h} \\ &= \frac{\dfrac{w(x+h) - w(x)}{h} - \dfrac{w(x) - w(x-h)}{h}}{h} = \frac{w(x+h) - 2w(x) + w(x-h)}{h^2} \end{aligned}$$

which shows that the central difference scheme of computing the second derivative is the same as applying the central difference scheme of computing the first derivative twice (but with difference mesh size).

3.3.1 The 1-dimensional heat equations

We first consider the 1-d heat equations with Dirichlet boundary condition

$$\vartheta_t - \kappa \vartheta_{xx} = f(x,t) \qquad \text{in} \quad (0,L) \times \mathbb{R}^+\,, \tag{3.33a}$$

$$\vartheta = \vartheta_0 \qquad \text{on} \quad (0,L) \times \{0\}\,, \tag{3.33b}$$

$$\vartheta(0,t) = a(t)\,,\ \vartheta(L,t) = b(t) \qquad \text{on} \quad \{0,L\} \times \mathbb{R}^+\,. \tag{3.33c}$$

Let $\big\{0 = x_0 < x_1 < \cdots < x_{n+1} = L\big\}$ be a regular partition of $[0,L]$, and $h = L/(n+1)$. Define $\varphi_i(t) = \vartheta(x_i,t)$ and $f_i(t) = f(x_i,t)$. Then (3.33) implies that

$$\begin{aligned} \frac{d\varphi_i}{dt} - \frac{\kappa}{h^2}(\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}) = f_i(t) + \mathcal{O}(h^2) \qquad & \text{for all } 1 \leqslant i \leqslant n \text{ and } t > 0\,, \\ \varphi_i(0) = \vartheta_0(x_i) \qquad & \text{for all } 1 \leqslant i \leqslant n\,, \\ \vartheta_0(t) = a(t)\,,\ \vartheta_{n+1}(t) = b(t) \qquad & \text{for all } t > 0\,, \end{aligned}$$

where ϑ_0 is a given function independent of t, and a, b are given constants. Therefore, naively we look for the solution to the ODE

$$\frac{d}{dt}\begin{bmatrix}\phi_1(t)\\ \phi_2(t)\\ \phi_3(t)\\ \vdots\\ \vdots\\ \phi_{n-2}(t)\\ \phi_{n-1}(t)\\ \phi_n(t)\end{bmatrix} = \frac{\kappa}{h^2}\begin{bmatrix}-2 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0\\ 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots & 0\\ 0 & 1 & -2 & 1 & 0 & \cdots & \cdots & 0\\ \vdots & 0 & \ddots & \ddots & \ddots & 0 & \cdots & 0\\ \vdots & & \ddots & \ddots & \ddots & \ddots & \cdots & 0\\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 & 0\\ 0 & \cdots & \cdots & \cdots & 0 & 1 & -2 & 1\\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & -2\end{bmatrix}\begin{bmatrix}\phi_1(t)\\ \phi_2(t)\\ \phi_3(t)\\ \vdots\\ \vdots\\ \phi_{n-2}(t)\\ \phi_{n-1}(t)\\ \phi_n(t)\end{bmatrix} + \frac{\kappa}{h^2}\begin{bmatrix}a(t)\\ 0\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ 0\\ b(t)\end{bmatrix} + \begin{bmatrix}f_1(t)\\ f_2(t)\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ f_{n-1}(t)\\ f_n(t)\end{bmatrix}$$

with initial condition

$$\begin{bmatrix}\phi_1(0) & \phi_2(0) & \cdots & \phi_n(0)\end{bmatrix}^{\mathrm{T}} = \begin{bmatrix}\vartheta_0(x_1) & \vartheta_0(x_2) & \cdots & \vartheta_0(x_n)\end{bmatrix}^{\mathrm{T}}$$

and treat $\phi_i(t)$ as an approximated value of $\varphi_i(t)$.

Example 3.36. Now suppose that we look for the numerical solution of

$$\begin{aligned}\vartheta_t(x,t) - \vartheta_{xx}(x,t) &= x^2 \sin t && \text{for all } 0 < x < 1 \text{ and } t > 0\,,\\ \vartheta(x,0) &= 1 + x + \sin(\pi x) && \text{for all } 0 < x < 1\,,\\ \vartheta(0,t) = 1\,,\ \vartheta(1,t) &= 2 && \text{for all } t > 0\,.\end{aligned}$$

We first input the function $f(x,t)$, $\vartheta_0(x,t)$, $a(t)$ and $b(t)$ as follows:

```
function output = forcing(x,t)
output = x.^2*sin(t);
```

```
function output = theta_0(x)
output = 1 + x + sin(pi*x);
```

```
function output = a(t)
output = 1*ones(size(t));
```

```
function output = b(t)
output = 2*ones(size(t));
```

Next we provide the function "heat_RHS" as "ODE_RHS" before. Here the values κ, h, and the matrix $K = \begin{bmatrix}-2 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0\\ 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots & 0\\ 0 & 1 & -2 & 1 & 0 & \cdots & \cdots & 0\\ \vdots & 0 & \ddots & \ddots & \ddots & 0 & \cdots & 0\\ \vdots & & \ddots & \ddots & \ddots & \ddots & \cdots & 0\\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 & 0\\ 0 & \cdots & \cdots & \cdots & 0 & 1 & -2 & 1\\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & -2\end{bmatrix}$ will be part of the inputs (so that we do not have to adjust them every time we modify the equations and the data).

```
function yp = heat_RHS(t,y,kappa,h,K)
n = length(y);
x = [h:h:n*h]';
yp = kappa/h.^2*(K*y + [a(t);zeros(n-2,1);b(t)]) + forcing(x,t);
```

Finally, we have the main code as follows:

```
L = 1
n = 10;
kappa = 1;
h = L/(n+1);
T_end = 1;
x = [h:h:n*h]';
K = -2*eye(n) + diag(ones(n-1,1),1) + diag(ones(n-1,1),-1);

[t,y] = ode45(@(t,y) heat_RHS(t,y,kappa,h,K),[0 T_end],theta_0(x));
y = [a(t),y,b(t)];  % adding the values of the solution at the end-points
x = 0:h:(n+1)*h;
plot(x,y(end,:),'b');
```

Here we use the command "eye" and "diag" to produce the matrix K. We remark that "eye(n)" will produce an $n \times n$ identity matrix, and for a given vector V "diag(V,k)" will produce an $m \times m$ matrix whose k-th diagonal is the vector V, where $m = \text{length}(V) + k$. We also note that each row of y, obtained using the ODE solver in the penultimate (倒數第二) line of the codes, provides the approximated value of φ at $x_1, \cdots, x_n$ at each sampled time, so the last line of the codes is to add $\vartheta(0,t)$ and $\vartheta(L,t)$ into the solution (for the purpose of plotting the solution).

If one wants to see the evolution of the solution, we can do the following:

```
x = 0:h:(n+1)*h;
figure(1)
for j=1:length(t)
    plot(x,y(j,:),'b');
    drawnow;  % force matlab to run the for loop
end;
```

3.3.2 The 1-dimensional wave equations

Now we consider the 1-d wave equations with Neumann boundary condition

$$u_{tt} - c^2 u_{xx} = f(x,t) \qquad \text{in} \quad (0,L) \times \mathbb{R}^+ , \tag{3.34a}$$

$$u = u_0 \, , u_t = u_1 \qquad \text{on} \quad (0,L) \times \{0\} , \tag{3.34b}$$

$$u_x(0,t) = a(t) \, , \; u_x(L,t) = b(t) \qquad \text{on} \quad \{0,L\} \times \mathbb{R}^+ . \tag{3.34c}$$

For an integer $n \geqslant 2$, define $h = \dfrac{L}{n-1}$ and $x_i = (i-1)h$ for $1 \leqslant i \leqslant n$. Let $v_i(t) = u(x_i, t)$ for $1 \leqslant i \leqslant n$. Then (3.34a) and the central difference scheme (3.32) imply that

$$\frac{d^2 v_i}{dt^2} - c^2 \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} = f_i(t) + \mathcal{O}(h^2) \qquad \text{for all } 2 \leqslant i \leqslant n-1 \text{ and } t > 0\,. \tag{3.35}$$

where as in the previous section $f_i(t) = f(x_i, t)$. Unlike the case of PDEs with Dirichlet boundary condition, now $v_1(t) = u(0,t)$ and $v_n(t) = u(L,t)$ are also unknown, so to complete the system we need to know how to compute $\dfrac{dv_1}{dt}$ and $\dfrac{dv_n}{dt}$.

Let $x_0 = -h$ and $x_{n+1} = L + h$. Using the central difference scheme (3.32), (3.34c) implies that

$$\begin{aligned} a(t) = u_x(x_1, t) &= \frac{u(x_2, t) - u(x_0, t)}{2h} + \mathcal{O}(h^2)\,, \\ b(t) = u_x(x_n, t) &= \frac{u(x_{n+1}, t) - u(x_{n-1}, t)}{2h} + \mathcal{O}(h^2)\,. \end{aligned}$$

Therefore, even though $u(-h,t)$ and $u(L+h,t)$ are meaningless objects (since u is a function defined on $[0, L]$), it is reasonable to assume that $u(x_0, t) = u(x_2, t) + \mathcal{O}(h^3)$ and $u(x_{n+1}, t) = u(x_{n-1}, t) + \mathcal{O}(h^3)$. Using the central difference scheme (3.32), we obtain that

$$\begin{aligned} u_{xx}(x_1, t) &= \frac{u(x_1 + h, t) - 2u(x_1, t) + u(x_1 - h, t)}{h^2} = \frac{2}{h^2}\big[v_2(t) - v_1(t)\big] - \frac{2}{h} a(t) + \mathcal{O}(h)\,, \\ u_{xx}(x_n, t) &= \frac{u(x_n + h, t) - 2u(x_n, t) + u(x_n - h, t)}{h^2} = \frac{2}{h^2}\big[v_{n-1}(t) - v_n(t)\big] + \frac{2}{h} b(t) + \mathcal{O}(h)\,; \end{aligned}$$

thus

$$\begin{aligned} \frac{d^2 v_1}{dt^2} - \frac{2c^2}{h^2}(v_2 - v_1) &= f_1(t) + \mathcal{O}(h)\,, \\ \frac{d^2 v_n}{dt^2} - \frac{2c^2}{h^2}(v_{v-1} - v_n) &= f_n(t) + \mathcal{O}(h)\,. \end{aligned}$$

Similar to the derivation in Section 3.3.1, naively we consider

$$\frac{d^2}{dt^2}\begin{bmatrix} \mathrm{v}_1(t) \\ \mathrm{v}_2(t) \\ \mathrm{v}_3(t) \\ \vdots \\ \vdots \\ \mathrm{v}_{n-2}(t) \\ \mathrm{v}_{n-1}(t) \\ \mathrm{v}_n(t) \end{bmatrix} = \frac{c^2}{h^2}\begin{bmatrix} -2 & 2 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} \mathrm{v}_1(t) \\ \mathrm{v}_2(t) \\ \mathrm{v}_3(t) \\ \vdots \\ \vdots \\ \mathrm{v}_{n-2}(t) \\ \mathrm{v}_{n-1}(t) \\ \mathrm{v}_n(t) \end{bmatrix} + \frac{c^2}{h^2}\begin{bmatrix} -a(t) \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ b(t) \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f_{n-1}(t) \\ f_n(t) \end{bmatrix}$$

with initial conditions

$$\begin{aligned} \begin{bmatrix} \mathrm{v}_1(0) & \mathrm{v}_2(0) & \cdots & \mathrm{v}_n(0) \end{bmatrix}^{\mathrm{T}} &= \begin{bmatrix} u_0(x_1) & u_0(x_2) & \cdots & u_0(x_n) \end{bmatrix}^{\mathrm{T}}\,, \\ \begin{bmatrix} \mathrm{v}_1'(0) & \mathrm{v}_2'(0) & \cdots & \mathrm{v}_n'(0) \end{bmatrix}^{\mathrm{T}} &= \begin{bmatrix} u_1(x_1) & u_1(x_2) & \cdots & u_1(x_n) \end{bmatrix}^{\mathrm{T}}\,, \end{aligned}$$

and treat $\mathrm{v}_i(t)$ as an approximated value of $v_i(t)$. We note that in order to use the ODE solver to solve the ODE above, we need to assign $\boldsymbol{w} = \boldsymbol{v}'(t)$, where $\boldsymbol{v} = (\mathrm{v}_1, \cdots, \mathrm{v}_n)^{\mathrm{T}}$, and write the system above as

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{w}\end{bmatrix} = \begin{bmatrix}\boldsymbol{w}\\ \frac{c^2}{h^2}K\boldsymbol{v}\end{bmatrix} + \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{f}(t)\end{bmatrix} = \begin{bmatrix}\mathrm{I}_n & 0\\ 0 & \frac{c^2}{h^2}K\end{bmatrix}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{w}\end{bmatrix} + \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{f}(t)\end{bmatrix}, \tag{3.36}$$

where I_n is the $n \times n$ identity matrix and $\boldsymbol{f} = (f_1, \cdots, f_n)^{\mathrm{T}}$.

Once (3.36) is obtained, it should be straight forward, as in the case of solving heat equations, to solve the ODE system numerically using the ODE solver. Here we only provide the code of the right-hand side function:

```
function yp = wave_RHS(t,y,c,h,K)
n = length(y);
x = [0:h:(n-1)*h]';
yp = c^2/h^2*[eye(n), zeros(n,n); zeros(n,n),K]*y + [zeros(n,1);forcing(x,t)];
```

while K should be provided in the main code as

```
K = -2*eye(n) + diag([2;ones(n-2,1)],1) + diag([ones(n-2,1);2],-1);
```

We note that the first n rows of the solution y obtained using the ODE solver corresponds to the approximated value of u at $\{x_1, \cdots, x_n\}$, while the rest n rows of y corresponds to the approximated value of u_t at $\{x_1, \cdots, x_n\}$.

3.3.3 The 1-dimensional conservation laws

We have to **warn** the readers that the usual central difference scheme (to approximate the partial derivatives w.r.t. x) together with the ODE solver is not a useful tool of solving the PDEs from conservation laws. In order to demonstrate this fact, we look at the numerical solution of the equation

$$u_t + u_x = q(x,t) \qquad \text{in} \quad (0,L)\times(0,T)\,, \tag{3.37a}$$

$$u(x,0) = u_0(x) \qquad \text{on} \quad (0,L)\times\{t=0\}\,, \tag{3.37b}$$

$$u(0,t) = u(L,t) = 0 \qquad \text{for all } t > 0\,. \tag{3.37c}$$

Let $\mathcal{P} = \{0 = x_0 < x_1 < \cdots < x_{n+1} = L\}$ be a regular partition of $[0, L]$, $h = L/(n+1)$, and define $u_i(t) = u(x_i, t)$ for $0 \leqslant i \leqslant n+1$. Then (3.37) implies that

$$\frac{du_i}{dt} + u_x(x_i,t) = q(x_i,t) \qquad \text{for all } 1 \leqslant i \leqslant n \text{ and } t > 0\,.$$

Using the central difference scheme (3.32) to approximate $u_x(x_i, t)$, we find that

$$\frac{du_i}{dt} + \frac{u_{i+1}(t) - u_{i-1}(t)}{2h} = q(x_i,t) + \mathcal{O}(h^2) \qquad \text{for all } 1 \leqslant ilen \text{ and } t > 0$$

where $u_0(t) = u_{n+1}(t) = 0$. The ODE above motivates the following ODE

$$\frac{d}{dt}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \vdots \\ \vdots \\ v_{n-2} \\ v_{n-1} \\ v_n \end{bmatrix} = \frac{1}{2h}\begin{bmatrix} 0 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & -1 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\ \vdots & & 0 & 1 & 0 & -1 & 0 \\ \vdots & & & 0 & 1 & 0 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \vdots \\ \vdots \\ v_{n-2} \\ v_{n-1} \\ v_n \end{bmatrix} + \begin{bmatrix} q_1(t) \\ q_2(t) \\ q_3(t) \\ \vdots \\ \vdots \\ \vdots \\ q_{n-2}(t) \\ q_{n-1}(t) \\ q_n(t) \end{bmatrix}$$

with initial condition

$$\begin{bmatrix} v_1(0) & v_2(0) & \cdots & v_n(0) \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} u_0(x_1) & u_0(x_2) & \cdots & u_0(x_n) \end{bmatrix}^{\mathrm{T}}$$

and treat $v_i(t)$ as approximated value of $u_i(t)$. So the main code is

```
L = 10
n = 100;
h = L/(n+1);
T_end = 10;
x = [h:h:n*h]';
K = diag(ones(n-1,1),-1) - diag(ones(n-1,1),1);


[t,y] = ode45(@(t,y) cl_RHS(t,y,h,K),[0 T_end],u_0(x));
y = [zeros(size(t)),y,zeros(size(t))];  % adding the values at the end-points
```

where cl_RHS is given by

```
function yp = cl_RHS(t,y,h,K)
n = length(y);
x = [h:h:n*h]';
yp = 1/(2*h)*K*y + source_q(x,t)];
```

Example 3.37. We first consider the case $L = 10$, $q(x,t) = (x-L)\cos x \sin t + \sin(x)\sin(t) + (x-L)\sin(x)\cos t$ and $u_0 = 0$. We note that the solution is indeed $u(x,t) = (x-L) * \sin(x) * \sin t$ (which is a smooth function so that the central difference scheme (3.32) provides good approximation of the derivatives). Knowing the exact solution of the PDE enables us to compare the numerical solution and the exact solution.

We still need

```
function output = source_q(x,t)
output = (x-10).*cos(x)*sin(t) + sin(x)*sin(t) + (x-10).*sin(x)*cos(t);
```

and

```
function output = u_0(x)
output = zeros(size(x));
```

to run simulations. To see the outcome, we use

```
x = 0:h:(n+1)*h;
figure(1)
for j=1:length(t)
    plot(L/2,30,'.');  % this is to fix the windows
    hold on;
    plot(L/2,-30,'.');  % this is to fix the windows
    plot(x,(x-L).*sin(x)*sin(t(j)),'r');
    plot(x,y(j,:),'b');
    hold off;
    drawnow;  % force matlab to run the for loop
end;
```

You should be able to see that the numerical solution is on top of the exact solution (which should imply that there is no bug in our code).

We next consider the case $L = 10$, $q(x,t) = x(x-L)\cos t + (2x-L)\sin t$ and $u_0 = 0$. The exact solution is $u(x,t) = x(x-L)\sin t$. Now we modify the function source_q and the exact solution in the comparison of the numerical solution and the exact solution as follows:

```
function output = source_q(x,t)
output = x.*(x-10).*cos(t) + (2*x-10)*sin(t);
```

and change the line in magenta by

```
plot(x,x.*(x-L).*sin(t(j)),'r');
```

You will see a sawtooth like graph of the numerical solution, while the exact solution is still smooth.

Finally, you can change the source to

```
function output = source_q(x,t)
output = abs(x-5)-5;
```

and you will find that the numerical solution becomes a garbage immediately.

3.3.4 Built-in PDE solver in matlab®

There is a built-in solver for PDE in matlab®. The PDE has to be of the form

$$m\frac{\partial^2 u}{\partial t^2} + d\frac{\partial u}{\partial t} - \mathrm{div}(c\nabla u) + au = f \qquad \text{in} \quad \Omega\,, \tag{3.38}$$

where either the Dirichlet, Neumann or mixed type boundary condition can be imposed. The unknown u can be a scalar or vector-valued function.

The main tool of solving PDE of form (3.38) in matlab® is the command "solvepde".

Keywords to check in matlab®:

1. **solvepde**
2. **Parametrized Function for 2-D Geometry Creation**
3. **generateMesh**
4. **applyBoundaryCondition**
5. **setInitialConditions**
6. **Coefficient for specifyCoefficients**
7. **CoefficientAssignment Properties**

Chapter 4

Optimization Problems and Calculus of Variations

4.1 Examples of Optimization Problems

4.1.1 Heron's Principle

Given a straight line L and two points a, b on a plane P, find a point x on L such that $|\overline{ax}| + |\overline{bx}|$ is minimal.

Theorem 4.1. *If x is a point of L such that the sum $|\overline{ax}| + |\overline{bx}|$ is the least possible, then the lines $\overline{ax}$ and $\overline{bx}$ form equal angles with the line L.*

4.1.2 Steiner's Tree Problem

The Steiner tree problem is superficially similar to the minimum spanning tree problem: given a set V of points (vertices), interconnect them by a network (graph) of shortest length, where the length is the sum of the lengths of all edges. The difference between the Steiner tree problem and the minimum spanning tree problem is that, in the Steiner tree problem, extra intermediate vertices and edges may be added to the graph in order to reduce the length of the spanning tree.

4.1.3 Dido's Problem (Isoperimetric Problem)

For a simple closed curve C in the plane, let $\ell(C)$ denote the length of the curve. The isoperimetric problem is to find a curve C satisfying $\ell(C) = L$ which encloses the largest area.

If $A(C)$ denotes the area enclosed by the curve C, then the ***isoperimetric inequality*** provides that

$$\ell(C)^2 \geqslant 4\pi A(C) \qquad \text{for every simple closed curve } C\,, \tag{4.1}$$

and "=" holds if and only if C is a circle.

Sketch of the proof. Let $\mathscr{P}_n$ denote the collection of simple closed polygon with $2n$ sides and with length L. We look for one P in $\mathscr{P}_n$ which encloses the largest area. Let

$$P_n = [A_1, A_2, \cdots, A_n, A_{n+1}, \cdots, A_{2n}, A_1]$$

be a polygon in $\mathscr{P}_n$ which encloses the largest area. We use the notion $A_j = A_k$ if $j = k \pmod{2n}$.

Claim I: P_n is convex.

Claim II: For all $j \in \mathbb{N}$, $|\overline{A_jA_{j+1}}| = |\overline{A_{j+1}A_{j+2}}|$.

Claim III: For all $j \in \mathbb{N}$, $[A_j, A_{j+1}, \cdots, A_{j+n}, A_j]$ and $[A_{j+n}, A_{j+n+1}, \cdots, A_{j+2n}, A_{j+n}]$ encloses the same area.

Claim IV: For $1 < j < n+1$, $\overline{A_1A_j} \perp \overline{A_jA_{n+1}}$ at A_j.

Proof of Claim IV: If $\overline{A_1A_j}$ is not perpendicular to $\overline{A_jA_{n+1}}$ at A_j, we can adjust the position of A_1 to A_1', and adjust accordingly the positions of $A_2, \cdots, A_{j-1}$ to $A_2', \cdots, A_{j-1}'$ so that the polygon $[A_1, A_2, \cdots, A_j, A_1]$ is the identical (in shape) to $[A_1', A_2', \cdots, A_{j-1}', A_j, A_1']$. We note that the area enclosed by the polygon $[A_1', \cdots, A_{j-1}', A_j, A_{j+1}, \cdots, A_{n+1}, A_1']$ is larger than the area enclosed by the polygon $[A_1, \cdots, A_{n+1}, A_1]$. **(End of proof of Claim IV)**

By Claim IV, $A_j's$ locates on a circle (with diameter $|A_1A_{n+1}|$). Let r_n be the radius of the circle in which P_n is inscribed. Then $4nr_n \sin\dfrac{\pi}{2n} = L$ and the area A_n enclosed by P_n is

$$A_n = nr_n^2 \sin\frac{\pi}{n} = \frac{L^2}{8n}\cot\frac{\pi}{2n}\,;$$

thus $A_{n+1} \geqslant A_n$ for all $n \in \mathbb{N}$ (**Exercise!**). The circle C with radius r has length L and encloses the largest area among all simple closed curves with length L and $L^2 = 4\pi A$. ▫

On the other hand, the minimization problem can be reformulated by looking for "minimizer" in the space of piecewise differentiable closed curve; that is, we look for curves C that can be parameterized, using the arc-length, by vector-valued function $(x(s), y(s))$ in the set

$$\mathcal{A} = \left\{(x(s), y(s)) \,\middle|\, x, y \in \mathscr{D}^1([0, L]; \mathbb{R}), x(0) = x(1), y(0) = y(1), \dot{x}^2(s) + \dot{x}^2(s) = 1\right\},$$

where $\mathscr{D}^1([0,1]; \mathbb{R})$ consists of continuous, piecewise continuously differentiable functions on $[0, L]$. Then the problem above is equivalent to the minimization problem

$$\min_{(x,y)\in\mathcal{A}} \int_0^L \left[x(s)\dot{y}(s) - \dot{x}(s)y(s)\right] ds\,.$$

4.1.4 Minimal Surface of Revolution

This is a problem of finding a curve C connecting (x_0, y_0) and (x_1, y_1), where $x_0 < x_1$, such that its surface of revolution has the least surface area. Given a function $y = y(x)$ satisfying

$y(x_0) = y_0$ and $y(x_1) = y_1$, the surface of revolution of the curve $C = \big\{(x, y(x)) \,\big|\, y : [x_0, x_1] \to \mathbb{R} \text{ is differentiable}, y(x_0) = y_0, y(x_1) = y_1\big\}$ is given by

$$2\pi \int_{x_0}^{x_1} y\sqrt{1 + y'^2(x)}\, dx\,.$$

Therefore, the problem of minimal surface of revolution is to find a function $y \in \mathcal{A} \equiv \big\{y \in \mathscr{D}^1([x_0, x_1]) \,\big|\, y(x_0) = y_0, y(x_1) = y_1\big\}$ which minimizes the functional

$$I(y) = 2\pi \int_{x_0}^{x_1} y\sqrt{1 + y'^2(x)}\, dx\,.$$

4.1.5 Newton's Problem

The Newton problem is to find a curve C connecting (x_0, y_0) and (x_1, y_1), where $x_0 < x_1$, such that its surface of revolution has the least resistance from the air when it moves along x-axis with speed v (or velocity $(v, 0, 0)$).

Let u be the normal component of the velocity (given some surface of revolution) $\big($thus $u = \dfrac{dy}{ds}v = \dfrac{y'v}{\sqrt{1+y'^2}}\big)$. Suppose that for each surface element dS (at point (x, y, z)), the resistance force is $\big[\varphi(u)dS\big]\mathbf{N}$ for some function φ, where $\mathbf{N}$ is the unit normal of the surface with negative first component (which means the resistance force points to the left). If the surface of revolution is given by the curve $y = y(x)$, then with ds denoting the infinitesimal arc-length, for each slice of the surface the total force acting on this slice is $2\pi y\varphi(u)ds(\mathbf{N}\cdot\mathbf{e}_1)$ (the vertical component cancels out); thus by the fact that $\dfrac{dy}{ds} = (\mathbf{N}\cdot\mathbf{e}_1)$, the total resistance force (in magnitude) is

$$I(y) = 2\pi \int_{x_0}^{x_1} y\varphi(u)ds\frac{dy}{ds} = 2\pi \int_{x_0}^{x_1} yy'\varphi\big(\frac{y'v}{\sqrt{1+y'^2}}\big)dx\,.$$

Therefore, the Newton problem can be formulated as "finding a function $y \in \mathcal{A} \equiv \big\{y \in \mathscr{D}^1([x_0, x_1]) \,\big|\, y(x_0) = y_0, y(x_1) = y_1\big\}$ which minimizes $I(y)$".

Newton's model: $\varphi(u) = u^2$.

4.1.6 Brachistochrone Problem

A brachistochrone curve, meaning "shortest time" or curve of fastest descent, is the curve that would carry an idealized point-like body, starting at rest and moving along the curve, without friction, under constant gravity, to a given end point in the shortest time. For given two point $(0, 0)$ and (a, b), where $b < 0$, what is the brachistochrone curve connecting $(0, 0)$ and (a, b)?

Given a curve parameterized by $\big\{(f(y), y) \,\big|\, y \in [b, 0]\big\}$ for some continuously differentiable function f, the total time required to travel from $(0, 0)$ to (a, b) is given by

$$T(f) = \int_b^0 \frac{\sqrt{1 + f'^2(y)}}{\sqrt{-2gy}}\, dy\,.$$

Therefore, the brachistochrone problem can be formulated as finding $f \in \mathcal{A} = \{h \in \mathscr{C}^1([0,b]) \,|\, h(0) = 0, h(b) = a\}$ such that $T(f)$ is minimized. In other words, the minimizer h satisfies that

$$T(h) = \inf_{f \in \mathcal{A}} \int_b^0 \frac{\sqrt{1 + f'(y)^2}}{\sqrt{-2gy}}\, dy\,.$$

4.2 Simplest Problem in Calculus of Variations

Let $[a,b] \subseteq \mathbb{R}$, $L : [a,b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be continuous. We consider the problem of minimizing the functional

$$I(y) = \int_a^b L(x, y(x), y'(x))\, dx$$

for $y \in \mathscr{C}^1([a,b])$ or $\mathscr{D}^1([a,b])$, and y satisfies the boundary condition $y(a) = A_0, y(b) = B_0$, where $\mathscr{C}^1([a,b])$ denotes the space of continuously differentiable functions on $[a,b]$, and $\mathscr{D}^1([a,b])$ denotes the space of continuous, piecewise continuously differentiable functions on $[a,b]$. In other words, with $\mathcal{A}$ denoting either the set $\{y \in \mathscr{C}^1([a,b]) \,|\, y(a) = A_0, y(b) = B_0\}$ or $\{y \in \mathscr{D}^1([a,b]) \,|\, y(a) = A_0, y(b) = B_0\}$, we consider the minimization problem

$$\inf_{y \in \mathcal{A}} \int_a^b L(x, y(x), y'(x))\, dx\,. \tag{4.2}$$

The function L is called the ***Lagrangian***.

In the following discussion, we write $L = L(x,y,p)$ and let $\arg\min_{z \in \mathcal{A}} I(z)$ denote the minimizer, if exists, of the minimization problem $\min_{z \in \mathcal{A}} I(z)$. In other word, if $y = \arg\min_{z \in \mathcal{A}} I(z)$, then $y \in \mathcal{A}$ and

$$I(y) \leqslant I(z) \qquad \forall\, z \in \mathcal{A}\,.$$

Remark 4.2. Let

$$\mathcal{X} = \{y \in \mathscr{C}^1([a,b]) \,|\, y(a) = A_0, y(b) = B_0\}$$
$$\mathcal{Y} = \{y \in \mathscr{D}^1([a,b]) \,|\, y(a) = A_0, y(b) = B_0\}\,.$$

Then $\arg\min_{z \in \mathcal{X}} I(z)$, if exists, equals $\arg\min_{z \in \mathcal{Y}} I(z)$. To see this, we first note that $\min_{z \in \mathcal{X}} I(z) \geqslant \min_{z \in \mathcal{Y}} I(z)$; thus for $\arg\min_{z \in \mathcal{X}} I(z) \neq \arg\min_{z \in \mathcal{Y}} I(z)$ to hold, we must have $\widehat{y} \in \mathcal{Y} \backslash \mathcal{X}$ such that $I(\widehat{y}) < \min_{z \in \mathcal{X}} I(z)$. By smooth $\widehat{y}$ at corners, we obtain $\bar{y} \in \mathcal{X}$ such that $I(\bar{y}) < \min_{z \in \mathcal{X}} I(z)$, a contradiction.

However, it is possible that there are only minimizers in $\mathscr{D}^1([a,b])$. See Example 4.15 for the detail.

4.2.1 First Variation of I

Let $\mathcal{A} = \{y \in \mathscr{D}^1([a,b]) \,|\, y(a) = A_0, y(b) = B_0\}$ and $\mathcal{N} = \{\eta \in \mathscr{C}^1([a,b]) \,|\, \eta(a) = \eta(b) = 0\}$, called the ***admissible set*** and the ***test function space***, respectively. For $y \in \mathcal{A}$, $\eta \in \mathcal{N}$

and $\epsilon \in \mathbb{R}$, let $J(\epsilon) = I(y + \epsilon\eta)$ and consider the following quotient

$$\frac{J(\epsilon) - J(0)}{\epsilon} = \frac{1}{\epsilon} \int_a^b \big[L(x, y(x) + \epsilon\eta(x), y'(x) + \epsilon\eta'(x)) - L(x, y(x), y'(x))\big]\, dx$$

Assume that L_y and L_p are continuous, then

$$\lim_{\epsilon \to 0} \frac{J(\epsilon) - J(0)}{\epsilon} = \underbrace{\int_a^b \big[L_y(x, y(x), y'(x))\eta(x) + L_p(x, y(x), y'(x))\eta'(x)\big]\, dx}_{\equiv\, \delta I(y;\eta) \text{ or } \frac{\delta I}{\delta \eta}(y)} .$$

This limit, denoted by $\delta I(y;\eta)$ or $\dfrac{\delta I}{\delta \eta}(y)$, is called the ***first variation*** of I at y along η.

Theorem 4.3. *If $y = \arg\min_{z \in \mathcal{A}} I(z)$ is a minimizer of I, then $\delta I(y;\eta) = 0$ for all $\eta \in \mathcal{N}$.*

Definition 4.4. The integral equation $\delta I(y;\eta) = 0$ for all $\eta \in \mathcal{N}$ is called the ***weak form*** of the ***Euler-Lagrange equation*** (associated with the minimization problem (4.2)).

• **Basic Lemmas**

Lemma 4.5. *If $y \in \mathscr{C}([a,b])$ and $\displaystyle\int_a^b y(x)\eta(x)\, dx = 0$ for all $\eta \in \mathscr{C}([a,b])$, then $y \equiv 0$.*

Lemma 4.6. *If $y \in \mathscr{C}([a,b])$ and $\displaystyle\int_a^b y(x)\eta'(x)\, dx = 0$ for all $\eta \in \mathcal{N}$, then $y \equiv c$ for some constant c.*

Proof. Let $\eta(x) = \displaystyle\int_a^x \big(y(t) - c\big)\, dt$, where the constant c is chosen so that $\displaystyle\int_a^b \big(y(t) - c\big)\, dt = 0$. Then $\eta \in \mathcal{N}$ and

$$\int_a^b \big|y(x) - c\big|^2\, dx = \int_a^b \big(y(x) - c\big)\eta'(x)\, dx = -c \int_a^b \eta'(x)\, dx = c\big(\eta(a) - \eta(b)\big) = 0\,.$$

Therefore, $y(x) = c$ for all $x \in [a,b]$. ▫

Lemma 4.7. *If $y, z \in \mathscr{C}([a,b])$ satisfy*

$$\int_a^b \big[y(x)\eta(x) + z(x)\eta'(x)\big]\, dx = 0 \qquad \forall\, \eta \in \mathcal{N}\,, \tag{4.3}$$

then $z \in \mathscr{C}^1([a,b])$ and $z'(x) = y(x)$ for all $x \in [a,b]$.

Proof. Let $z_1(x) = \displaystyle\int_a^x y(t)\, dt$. Integration-by-parts provides that

$$\int_a^b y(x)\eta(x)\, dx = z_1(x)\eta(x)\big|_{x=a}^{x=b} - \int_a^b z_1(x)\eta'(x)\, dx = -\int_a^b z_1(x)\eta'(x)\, dx\,;$$

thus (4.3) implies that

$$\int_a^b \big[z(x) - z_1(x)\big]\eta'(x)\, dx = 0 \qquad \forall\, \eta \in \mathcal{N}\,.$$

By Lemma 4.6, $z(x) - z_1(x) = C$ for some constant C. Therefore, $z(x) = C + \displaystyle\int_a^x y(t)\, dt$ which implies that $z \in \mathscr{C}^1([a,b])$ and $z'(x) = y(x)$. ▫

Lemma 4.8. *Suppose that $y, z \in \mathscr{C}([a,b])$ and z is not a constant function. If*

$$\int_a^b y(x)\eta'(x)\,dx = 0 \qquad \forall\, \eta \in \mathcal{N} \text{ and } \eta \text{ satisfies } \int_a^b z(x)\eta'(x)\,dx = 0\,,$$

then there are constants $\lambda, \mu \in \mathbb{R}$ such that $y(x) = \lambda z(x) + \mu$.

Proof. Let $\eta(x) = \displaystyle\int_a^x \big(y(t) - \lambda z(t) - \mu\big)\,dt$, where λ, μ are chosen so that $\eta(b) = 0$ and $\displaystyle\int_a^b z(x)\eta'(x)\,dx = 0$; that is,

$$\begin{aligned} \lambda \int_a^b z(x)\,dx + \mu \int_a^b dx &= \int_a^b y(x)\,dx\,, \\ \lambda \int_a^b z^2(x)\,dx + \mu \int_a^b z(x)\,dx &= \int_a^b y(x)z(x)\,dx\,. \end{aligned}$$

Since z is not a constant, the Cauchy-Schwarz inequality implies that the system above has a unique solution (λ, μ). Since $\eta \in \mathcal{N}$ and satisfies $\displaystyle\int_a^b z(x)\eta'(x)\,dx = 0$, we have

$$\int_a^b \big|y(x) - \lambda z(x) - \mu\big|^2\,dx = \int_a^b \big(y(x) - \lambda z(x) - \mu\big)\eta'(x)\,dx = -\mu \int_a^b \eta'(x)\,dx = 0\,;$$

thus $y(x) = \lambda z(x) + \mu$ for all $x \in [a,b]$. ▫

4.2.2 The Euler-Lagrange Equation

Recall that the weak form of the Euler-Lagrange equation associated with the minimization problem (4.2) is $\delta I(y;\eta) = 0$ for all $\eta \in \mathcal{N}$.

Theorem 4.9. *Suppose that L, L_y, L_p are continuous. If $\widehat{y} \in \mathcal{A}$ is a minimizer of the minimization problem* (4.2)*, then*

$$\frac{d}{dx} L_p(x, \widehat{y}(x), \widehat{y}'(x)) = L_y(x, \widehat{y}(x), \widehat{y}'(x)) \tag{4.4}$$

for point x at which $\widehat{y}'$ is continuous.

Proof. Apply Theorem 4.3 and Lemma 4.7 to each interval on which $\widehat{y}$ is of class $\mathscr{C}^1$. ▫

Definition 4.10. Equation (4.4) is called (the ***strong form*** of) the Euler-Lagrange equation (associated with the minimization problem (4.2)).

Remark 4.11.

1. Theorem 4.9 is essentially due to Du Bois-Reymond, so (4.4) is also called the Du Bois-Reymond equation.

2. If $\widehat{y} \in \mathscr{C}^2([a,b])$ and L_{px}, L_{yp}, L_{pp} are continuous, then $\widehat{y}$ satisfies the following second order ODE

$$\begin{aligned} &L_{pp}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}''(x) \\ &\quad = L_y(x, \widehat{y}(x), \widehat{y}'(x)) - L_{px}(x, \widehat{y}(x), \widehat{y}'(x)) - L_{py}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}'(x)\,. \end{aligned}$$

This is the equation that Euler originally derived/obtained.

Example 4.12. Now we consider the brachistochrone problem. We rewritten the minimization problem as

$$\inf_{h \in \mathcal{A}} -\int_0^b \frac{\sqrt{1+y'(x)^2}}{\sqrt{-2gx}}\,dx$$

where $\mathcal{A} = \big\{y \in \mathscr{D}^1([0,b]) \,\big|\, y(0) = 0, y(b) = a\big\}$. Therefore, $L(x,y,p) = -\dfrac{\sqrt{1+p^2}}{\sqrt{-2gx}}$ which implies that the Euler-Lagrange equation for the brachistochrone problem is

$$\frac{d}{dx}\frac{y'}{\sqrt{-2gx}\sqrt{1+y'^2}} = 0\,.$$

Therefore, if $\widehat{y} \in \mathcal{A}$ is a minimizer, then in each interval where $\widehat{y}'$ is continuous,

$$\frac{\widehat{y}'}{\sqrt{-2gx}\sqrt{1+\widehat{y}'^2}} = C$$

for some constant C. The equation above shows that $\widehat{y}'^2 = -2C^2gx(1+\widehat{y}'^2)$ which, together with the fact that $\widehat{y}'$ must be non-positive, implies that

$$\widehat{y}'(x) = -\sqrt{\frac{-2C^2gx}{1+2C^2gx}}\,.$$

As a consequence, if $\widehat{y} \in \mathscr{C}^1([0,b])$,

$$\widehat{y}(x) = -\int_0^x \sqrt{\frac{-2C^2gt}{1+2C^2gt}}\,dt\,.$$

and the constant C is determined by the condition $\widehat{y}(b) = a$.

Example 4.13. The Euler-Lagrange equation for the minimal surface of revolution problem is

$$\frac{d}{dx}\frac{yy'}{\sqrt{1+y'^2}} = \sqrt{1+y'^2}\,,$$

and the Euler-Lagrange equation for Newton's problem (with $\varphi(u) = u^2$) is

$$\frac{d}{dx}\frac{yy'^2(y'^2+3)}{(1+y'^2)^2} = \frac{y'^3}{1+y'^2}\,.$$

Theorem 4.14. *Suppose that $\widehat{y} \in \mathscr{D}^1([a,b])$ satisfies the Euler-Lagrange equation (4.4), and $x \in (a,b)$. If L_{px}, L_{py} are continuous at $(x, \widehat{y}(x), \widehat{y}'(x))$, $L_{pp}(x, \widehat{y}(x), \widehat{y}'(x)) \neq 0$, and $\widehat{y}'$ is continuous at x, then $\widehat{y}''(x)$ exists.*

Proof. Since $\widehat{y} \in \mathcal{A}$ is a minimizer of the minimization problem (4.2) and $\widehat{y}'$ is continuous at x, by Theorem 4.9 we find that

$$\frac{d}{dx} L_p(x, \widehat{y}(x), \widehat{y}'(x)) = L_y(x, \widehat{y}(x), \widehat{y}'(x)) \,.$$

Note that

$$\begin{aligned}
\frac{d}{dx} L_p(x, \widehat{y}(x), \widehat{y}'(x)) &= \lim_{\epsilon \to 0} \frac{L_p(x+\epsilon, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))}{\epsilon} \\
&= \lim_{\epsilon \to 0} \Big[\frac{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))}{\epsilon} \\
&\qquad + \frac{L_p(x+\epsilon, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon))}{\epsilon} \Big] \,.
\end{aligned}$$

By the mean value theorem,

$$\begin{aligned}
&L_p(x+\epsilon, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon)) \\
&\quad = L_p(x+\epsilon, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) \\
&\qquad + L_p(x, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon)) \\
&\quad = L_{px}(x+\epsilon\theta_1, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon))\epsilon \\
&\qquad + L_{py}(x, \widehat{y}(x) + \theta_2(\widehat{y}(x+\epsilon) - \widehat{y}(x)), \widehat{y}'(x+\epsilon))(\widehat{y}(x+\epsilon) - \widehat{y}(x))
\end{aligned}$$

for some $\theta_1 = \theta_1(\epsilon, x)$ and $\theta_2 = \theta_2(\epsilon, x)$ satisfying $|\theta_1|, |\theta_2| \leqslant 1$. Therefore, by the continuity of L_{px} and L_{py} at $(x, \widehat{y}(x), \widehat{y}'(x))$ and $\widehat{y}'$ at x,

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{L_p(x+\epsilon, \widehat{y}(x+\epsilon), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon))}{\epsilon} \\
&\qquad = L_{px}(x, \widehat{y}(x), \widehat{y}'(x)) + L_{py}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}'(x) \,;
\end{aligned}$$

thus

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))}{\epsilon} \\
&\qquad = L_y(x, \widehat{y}(x), \widehat{y}'(x)) - L_{px}(x, \widehat{y}(x), \widehat{y}'(x)) - L_{py}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}'(x)
\end{aligned} \tag{4.5}$$

exists.

Suppose the contrary that $\widehat{y}''(x)$ does not exist. Then

$$\#\big\{0 < |\epsilon| < \delta \,\big|\, \widehat{y}'(x+\epsilon) \neq \widehat{y}'(x)\big\} = \infty \qquad \forall\, \delta > 0 \tag{4.6}$$

for otherwise there exists $\delta > 0$ such that $\#\big\{0 < |\epsilon| < \delta \,\big|\, \widehat{y}'(x+\epsilon) \neq \widehat{y}'(x)\big\} < \infty$; thus there exists $\epsilon^* > 0$ such that $\widehat{y}'(x+\epsilon) = \widehat{y}'(x)$ for all $|\epsilon| < \epsilon^*$ which then leads to a contradiction that

$$\lim_{\epsilon \to 0} \frac{\widehat{y}'(x+\epsilon) - \widehat{y}'(x)}{\epsilon} = 0 \,.$$

Let $\{\epsilon_j\}_{j=1}^{\infty}$ be sequence converging to 0 such that

$$\liminf_{j \to \infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j} < \limsup_{j \to \infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j} \,. \tag{4.7}$$

Using (4.6), $\big\{j \in \mathbb{N} \,\big|\, \widehat{y}'(x+\epsilon_j) \neq \widehat{y}'(x)\big\} = \{j_\ell\}_{\ell=1}^{\infty}$; thus by the definition of L_{pp} and the continuity of $\widehat{y}'$ at x,

$$\lim_{\ell\to\infty} \frac{L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon_{j_\ell})) - L_p(x,\widehat{y}(x),\widehat{y}'(x))}{\widehat{y}'(x+\epsilon_{j_\ell}) - \widehat{y}'(x)} = L_{pp}(x,\widehat{y}(x),\widehat{y}'(x))\,.$$

The condition $L_{pp}(x,\widehat{y}(x),\widehat{y}'(x)) \neq 0$ further implies that

$$\lim_{\ell\to\infty} \frac{\widehat{y}'(x+\epsilon_{j_\ell}) - \widehat{y}'(x)}{L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon_{j_\ell})) - L_p(x,\widehat{y}(x),\widehat{y}'(x))} = \frac{1}{L_{pp}(x,\widehat{y}(x),\widehat{y}'(x))}\,.$$

We then conclude from (4.5) that

$$\begin{aligned}
&\lim_{\ell\to\infty} \frac{\widehat{y}'(x+\epsilon_{j_\ell}) - \widehat{y}'(x)}{\epsilon_{j_\ell}} \\
&\quad= \lim_{\ell\to\infty} \Big[\frac{\widehat{y}'(x+\epsilon_{j_\ell}) - \widehat{y}'(x)}{L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon_{j_\ell})) - L_p(x,\widehat{y}(x),\widehat{y}'(x))} \frac{L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon_{j_\ell})) - L_p(x,\widehat{y}(x),\widehat{y}'(x))}{\epsilon_{j_\ell}}\Big] \\
&\quad= \frac{L_y(x,\widehat{y}(x),\widehat{y}'(x)) - L_{px}(x,\widehat{y}(x),\widehat{y}'(x)) - L_{py}(x,\widehat{y}(x),\widehat{y}'(x))\widehat{y}'(x)}{L_{pp}(x,\widehat{y}(x),\widehat{y}'(x))}\,. \qquad (4.8)
\end{aligned}$$

If $\#\big\{j \in \mathbb{N} \,\big|\, \widehat{y}'(x+\epsilon_j) = \widehat{y}'(x)\big\} = \infty$, then with $\{j_\ell\}_{\ell=1}^{\infty} = \big\{j \in \mathbb{N} \,\big|\, \widehat{y}'(x+\epsilon_j) = \widehat{y}'(x)\big\}$, (4.5) shows that

$$\begin{aligned}
&L_y(x,\widehat{y}(x),\widehat{y}'(x)) - L_{px}(x,\widehat{y}(x),\widehat{y}'(x)) - L_{py}(x,\widehat{y}(x),\widehat{y}'(x))\widehat{y}'(x) \\
&\qquad= \lim_{\epsilon\to 0} \frac{L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon)) - L_p(x,\widehat{y}(x),\widehat{y}'(x))}{\epsilon} \\
&\qquad= \lim_{j\to\infty} \frac{L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon_j)) - L_p(x,\widehat{y}(x),\widehat{y}'(x))}{\epsilon_j} = 0\,;
\end{aligned}$$

thus (4.8) yields that

$$\lim_{j\to\infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j} = 0\,,$$

a contradiction to (4.7). Therefore, $\#\big\{j \in \mathbb{N} \,\big|\, \widehat{y}(x+\epsilon_j) = \widehat{y}(x)\big\} < \infty$; however, this would imply that

$$\lim_{j\to\infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j} = \frac{L_y(x,\widehat{y}(x),\widehat{y}'(x)) - L_{px}(x,\widehat{y}(x),\widehat{y}'(x)) - L_{py}(x,\widehat{y}(x),\widehat{y}'(x))\widehat{y}'(x)}{L_{pp}(x,\widehat{y}(x),\widehat{y}'(x))}\,,$$

still a contradiction to (4.7). □

Example 4.15. Let $\mathcal{A} = \big\{y \in \mathscr{D}^1([0,1]) \,\big|\, y(0) = y(1) = 0\big\}$. Consider the minimization problem

$$\inf_{y\in\mathcal{A}} \int_0^1 \big(y'(x)^2 - 1\big)^2\,dx\,;$$

that is, we assume $L(x,y,p) = (p^2-1)^2$. The Euler-Lagrange equation associated with this minimization problem is

$$\frac{d}{dx}\frac{d}{dp}\Big|_{p=y'(x)} (p^2-1)^2 = 0$$

which, together with the fact that $L_{pp}(x,y,p) = 12p^2 - 4$, implies that if $p^2 \neq \frac{1}{3}$ the minimizer $\widehat{y}$ satisfies

$$2\widehat{y}'^2\widehat{y}'' + (\widehat{y}'^2 - 1)\widehat{y}'' = 0\,.$$

Therefore, $\widehat{y}''(3\widehat{y}'^2 - 1) = 0$ on points at which $\widehat{y}'$ is continuous and $\widehat{y}'^2 \neq \frac{1}{3}$. Therefore, $\widehat{y}'' = 0$ if $\widehat{y}'^2 \neq \frac{1}{3}$ which implies that $\widehat{y}'$ is piecewise constant. The minimizer is then saw-tooth like function with slope ± 1, and there are only $\mathscr{D}^1$-minimizers.

Remark 4.16. Suppose $L_{pp} \neq 0$ and $\widehat{y} = \arg\min_{z \in \mathcal{A}} I(z)$. If $\widehat{y}'$ is continuous in a neighborhood of x, then $\widehat{y}''$ exists in a neighborhood of x and is continuous there.

Remark 4.17 (Remark on the extensions of the simplest problem of Calculus of variations)**.**

1. **Higher derivatives**: The Lagrangian might involves higher order derivatives of y. For example, we can consider the minimization problem

$$\inf_{y \in \mathcal{A}} \int_a^b L(x, y(x), y'(x), y''(x))\, dx\,,$$

where $\mathcal{A} = \big\{y \in \mathscr{D}^2([a,b]) \,\big|\, y(a) = A_0, y(b) = B_0, y'(a) = A_1, y'(b) = B_1\big\}$. We note that the corresponding test function space is $\mathcal{N} = \big\{y \in \mathscr{D}^2([a,b]) \,\big|\, y(a) = y(b) = y'(a) = y'(b) = 0\big\}$.

If $\widehat{y}$ is a minimizer, then $J(\epsilon) = I(\widehat{y} + \epsilon\eta)$ attains its minimum at $\epsilon = 0$ for all $\eta \in \mathcal{N}$. This implies $J'(0) = 0$ for all $\eta \in \mathcal{N}$, and this condition gives the weak form of the Euler-Lagrange equation associated with this minimization problem: write $L = L(x,y,p,q)$,

$$\begin{aligned}\int_a^b \Big[& L_y(x, \widehat{y}(x), \widehat{y}'(x), \widehat{y}''(x))\eta(x) + L_p(x, \widehat{y}(x), \widehat{y}'(x), \widehat{y}''(x))\eta'(x) \\ & + L_q(x, \widehat{y}(x), \widehat{y}'(x), \widehat{y}''(x))\eta''(x)\Big]\, dx = 0 \qquad\qquad \forall\, \eta \in \mathcal{N}\,.\end{aligned}$$

2. **Free ends**: This is to consider the minimization problem

$$\inf_{y \in \mathscr{D}^1([a,b])} \int_a^b L(x, y(x), y'(x))\, dx\,.$$

In this case, the test function space is then $\mathcal{N} = \mathscr{C}^1([a,b])$. The same argument implies that

$$L_p(b, \widehat{y}(b), \widehat{y}'(b))\eta(b) - L_p(a, \widehat{y}(a), \widehat{y}'(a))\eta(a) = 0 \qquad \forall\, \eta \in \mathscr{C}^1([a,b])\,.$$

Therefore,

   (a) The Euler-Lagrange/Du Bois-Reymond equation holds.

   (b) $L_p(b, \widehat{y}(b), \widehat{y}'(b)) = L_p(a, \widehat{y}(a), \widehat{y}'(a)) = 0$ - this is called the ***natural boundary condition***.

3. **Several dependent variables**: Let

$$\mathcal{A} = \left\{ \boldsymbol{y} = (y_1, \cdots, y_{\mathrm{n}}) : [a,b] \to \mathbb{R}^{\mathrm{n}} \,\middle|\, y_j \in \mathscr{D}^1([a,b]) \text{ for } 1 \leqslant j \leqslant \mathrm{n}\,,\, \boldsymbol{y}(a) = \boldsymbol{A}_0, \boldsymbol{y}(b) = \boldsymbol{B}_0 \right\}$$

or (when considering minimization problems with free ends)

$$\mathcal{A} = \left\{ \boldsymbol{y} = (y_1, \cdots, y_{\mathrm{n}}) : [a,b] \to \mathbb{R}^{\mathrm{n}} \,\middle|\, y_j \in \mathscr{D}^1([a,b]) \text{ for } 1 \leqslant j \leqslant \mathrm{n} \right\} \equiv \mathscr{D}^1([a,b]; \mathbb{R}^{\mathrm{n}})\,,$$

and $L : [a,b] \times \mathbb{R}^{\mathrm{n}} \times \mathbb{R}^{\mathrm{n}} \to \mathbb{R}$. Consider the minimization problem

$$\inf_{\boldsymbol{y} \in \mathcal{A}} \int_a^b L(x, \boldsymbol{y}(x), \boldsymbol{y}'(x))\, dx\,.$$

Write $L = L(x, y_1, \cdots, y_n, p_1, \cdots, p_n)$. Then the Du Bois-Reymond equation is

$$\frac{d}{dx} L_{p_i}(x, \boldsymbol{y}(x), \boldsymbol{y}'(x)) = L_{y_i}(x, \boldsymbol{y}(x), \boldsymbol{y}'(x)) \qquad \text{for} \quad i = 1, 2, \cdots, \mathrm{n}\,. \tag{4.9}$$

When considering free ends problem, natural boundary conditions

$$L_{p_i}(b, \widehat{\boldsymbol{y}}(b), \widehat{\boldsymbol{y}}'(b)) = L_{p_i}(a, \widehat{\boldsymbol{y}}(a), \widehat{\boldsymbol{y}}'(a)) = 0 \qquad \text{for} \quad i = 1, 2, \cdots, \mathrm{n} \tag{4.10}$$

have to be imposed for the minimizer $\boldsymbol{y}$.

4. **Several independent variables**: Let $\Omega \subseteq \mathbb{R}^{\mathrm{n}}$ be bounded open set, and $L : \Omega \times \mathbb{R} \times \mathbb{R}^{\mathrm{n}} \to \mathbb{R}$. Consider the minimization problem

$$\inf_{y \in \mathcal{A}} \int_\Omega L(x, y(x), Dy(x))\, dx\,,$$

where $\mathcal{A}$ could be

   (a) $\mathcal{A} = \left\{ y \in \mathscr{D}^1(\bar{\Omega}) \,\middle|\, y = f \text{ on } \partial\Omega \right\}$ (with corresponding $\mathcal{N} = \left\{ \eta \in \mathscr{C}^1(\bar{\Omega}) \,\middle|\, \eta = 0 \text{ on } \partial\Omega \right\}$) when considering the fixed-end problem, or

   (b) $\mathcal{A} = \mathscr{D}^1(\bar{\Omega})$ (with corresponding $\mathcal{N} = \mathscr{C}^1(\bar{\Omega})$) when considering the free-end problem.

   Define $J(\epsilon) = I(\widehat{y} + \epsilon\eta)$, where $\widehat{y} \in \mathcal{A}$ is a possible minimizer, $\eta \in \mathcal{N}$ and $\alpha \in \mathbb{R}$. The weak form of the Euler-Lagrange equation is $J'(0) = 0$.

5. **Non-affine admissible set**: We note that in Dido's problem the admissible set $\mathcal{A}$ is not an affine space (a translation of a vector space). In a minimization problem, the admissible set $\mathcal{A}$ in general is not an affine space so there is no obvious test function spaces $\mathcal{N}$ to work on. See Example 4.19 for deriving the weak form of the Euler-Lagrange equation for minimizers.

Example 4.18 (The minimal surface)**.** Suppose that $\Omega \subseteq \mathbb{R}^2$ is a bounded set with boundary parameterized by $(x(t), y(t))$ for $t \in I$, and $C \subseteq \mathbb{R}^3$ is a closed curve parameterized by $(x(t), y(t), f(x(t), y(t)))$ for some given function f. We want to find a surface having C as

its boundary with minimal surface area. Then the goal is to find a function u with the property that $u = f$ on $\partial\Omega$ that minimizes the functional

$$A(w) = \int_\Omega \sqrt{1 + |\nabla w|^2}\, dA\,.$$

Let $\varphi \in \mathscr{C}^1(\overline{\Omega})$, and define

$$\delta A(u;\varphi) = \lim_{t\to 0} \frac{A(u+\epsilon\varphi) - A(u)}{\epsilon} = \int_\Omega \frac{\nabla u \cdot \nabla\varphi}{\sqrt{1+|\nabla u|^2}}\, dx\,.$$

If u minimize A, then $\delta A(u;\varphi) = 0$ for all $\varphi \in \mathscr{C}^1(\Omega)$ satisfying $\varphi = 0$ on $\partial\Omega$. Assuming that $u \in \mathscr{C}^2(\overline{\Omega})$, by the divergence theorem (Theorem 3.31) we find that u satisfies

$$\operatorname{div}\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0\,,$$

or expanding the bracket using the Leibnitz rule, we obtain the ***minimal surface equation***

$$(1+u_y^2)u_{xx} - 2u_x u_y u_{xy} + (1+u_x^2)u_{yy} = 0 \qquad \forall\,(x,y) \in \Omega\,. \tag{4.11}$$

Example 4.19 (Isoperimetric Inequality - revisit)**.** We rephrase Dido's problem as finding a simply closed curve C enclosing a fixed number A of area with shortest perimeter. Let

$$\mathcal{A} = \Big\{ \boldsymbol{r}(t) = (x(t), y(t)) \in \mathscr{D}^1([0,1]) \,\Big|\, \boldsymbol{r}(0) = \boldsymbol{r}(1)\,,\ \int_0^1 (x\dot{y} - y\dot{x})dt = 2\mathrm{A} \Big\}$$

and $I(\boldsymbol{r}) = \inf\limits_{r\in\mathcal{A}} \int_0^1 |\boldsymbol{r}'(t)|\, dt$. We would like to study the minimization problem $\inf\limits_{r\in\mathcal{A}} I(\boldsymbol{r})$.

The difficulty of this particular formulation is that $\mathcal{A}$ is not an affine space so there is "no" corresponding test functions space to compute the first variation as before. To see how we derive the Euler-Lagrange equation for this minimization problem for a minimizer $\widehat{\boldsymbol{r}} = (\widehat{x}, \widehat{y})$, we introduce a family of curves $\boldsymbol{r}(t;\epsilon) = (x(t;\epsilon), y(t;\epsilon)) \in \mathcal{A}$, where $\epsilon \in \mathbb{R}$ is a parameter that will be passed to the limit, such that

1. $\boldsymbol{r}(t;0) = \widehat{\boldsymbol{r}}(t)$; 2. $\boldsymbol{r}(0;\epsilon) = \boldsymbol{r}(1;\epsilon)$; 3. $\boldsymbol{r}$ is also differentiable in ϵ.

Denote $\delta\boldsymbol{r}(t) = (\delta x(t), \delta y(t)) = \dfrac{d}{d\epsilon}\Big|_{\epsilon=0} \boldsymbol{r}(t;\epsilon)$. Since $\boldsymbol{r} \in \mathcal{A}$,

$$\frac{d}{d\epsilon}\Big|_{\epsilon=0} \int_0^1 \big[x(t;\epsilon)\dot{y}(t;\epsilon) - y(t;\epsilon)\dot{x}(t;\epsilon)\big]\, dt = 0$$

which implies that $\delta\boldsymbol{r}$ satisfies

$$\int_0^1 \big[(\delta x)\dot{\widehat{y}} + \widehat{x}(\dot{\delta y}) - (\delta y)\dot{\widehat{x}} - \widehat{y}(\dot{\delta x})\big]\, dt = 0\,. \tag{4.12}$$

For each possible minimizer $\widehat{\boldsymbol{r}}$, the relation above induces a linear vector space

$$\mathscr{N}_{\widehat{\boldsymbol{r}}} = \Big\{ \delta\boldsymbol{r} = (\delta x, \delta y) \in \mathscr{C}^1([0,1]) \,\Big|\, \int_0^1 \big[\widehat{x}(\dot{\delta y}) - \widehat{y}(\dot{\delta x})\big]\, dt = 0 \Big\}\,.$$

Now we look for a minimizer $\widehat{\boldsymbol{r}} \in \mathscr{C}^2([0,1])$. We note that Remark 4.2 implies that if we are able to find a minimizer in $\mathscr{C}^2([0,1])$ (thus a $\mathscr{C}^1$-minimizer), it must also be a minimizer in $\mathscr{D}^1([0,1])$. Since $\widehat{\boldsymbol{r}} \in \mathscr{C}^2([0,1])$ is a minimizer, the function $J(\epsilon) \equiv I(\boldsymbol{r}(t;\epsilon))$ attains its minimum at $\epsilon = 0$. This yields that $J'(0) = 0$ or more precisely,

$$\int_0^1 \frac{\widehat{\boldsymbol{r}}'(t) \cdot (\delta \boldsymbol{r})'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\, dt = 0\,,$$

where we note that $\delta \boldsymbol{r} \in \mathscr{N}_{\widehat{\boldsymbol{r}}}$. In other words, $\widehat{\boldsymbol{r}}$ satisfies

$$\int_0^1 \frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} \cdot (\delta \boldsymbol{r})'(t)\, dt = 0 \qquad \forall\, \delta \boldsymbol{r} \in \mathscr{N}_{\widehat{\boldsymbol{r}}}\,, \tag{4.13}$$

and Lemma 4.8 implies that there exists $\lambda_1, \lambda_2, \mu_1, \mu_2 \in \mathbb{R}$ such that

$$\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} = \big(-\lambda_1 \widehat{y}(t) + \mu_1, \lambda_2 \widehat{x}(t) + \mu_2\big)\,.$$

Since $\widehat{\boldsymbol{r}} = (\widehat{x}, \widehat{y}) \in \mathscr{C}^2([0,1])$, we differentiate the equation above and obtain that

$$\Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' = \big(-\lambda_1 \widehat{y}'(t), \lambda_2 \widehat{x}'(t)\big)\,.$$

Therefore, taking the inner product of the equation above with the unit tangent vector $\dfrac{\widehat{\boldsymbol{r}}'}{|\widehat{\boldsymbol{r}}'|}$, we find that

$$0 = \Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big) \cdot \Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' = \big(-\lambda_1 \widehat{y}'(t), \lambda_2 \widehat{x}'(t)\big) \cdot \Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big) = (\lambda_2 - \lambda_1) \frac{\widehat{x}'(t)\widehat{y}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} \quad \forall\, t \in [0,1]$$

which implies that $\lambda_2 = \lambda_1$; thus

$$\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} = \lambda(-\widehat{y}(t), \widehat{x}(t)\big) + (\mu_1, \mu_2)\,.$$

Note that $\lambda \neq 0$ for otherwise the unit tangent vector is constant which implies that $\widehat{\boldsymbol{r}}$ is a parametrization of a straight line. Therefore, with $\widetilde{\boldsymbol{r}} = \big(\widetilde{x}(t), \widetilde{y}(t)\big)$ denoting the vector

$$\big(\widetilde{x}(t), \widetilde{y}(t)\big) = \big(\widehat{x}(t) + \frac{\mu_2}{\lambda}, \widehat{y}(t) - \frac{\mu_1}{\lambda}\big)\,,$$

we have

$$\frac{\widetilde{\boldsymbol{r}}'(t)}{|\widetilde{\boldsymbol{r}}'(t)|} = \lambda(-\widetilde{y}(t), \widetilde{x}(t)\big)\,.$$

Finally, taking the inner product of the equation above with the (position) vector $\widetilde{\boldsymbol{r}}$, we conclude that

$$\frac{d}{dt}\big|\widetilde{\boldsymbol{r}}(t)\big|^2 = 0\,.$$

Therefore, the closed curve having fixed length and enclosing the largest area must be a circle.