# Mathematical Modeling 數學建模

**Ching-hsiao Arthur Cheng 鄭經斅**

# Contents

# 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 stages of modeling is to analyze the relationship between related variables and their dimensions. This method is called ***dimensional analysis***. Generally, the dimension of a variable is a combination of basic physical dimensions such as mass, length, time, charge, and temperature. For example, the dimension of speed is length per unit of time, and the unit of measurement is meters per second, miles per hour, or other units. Dimensional analysis is based on the important principle that the relationship between variables must have dimensional homogeneity; that is, the relationship between variables must be independent of the unit of measurement of its variables. Any meaningful equation must have the same dimensions on the left-hand side and the right-hand side. Checking whether this rule is followed is the most basic step for dimensional analysis.

**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.

Note that ”dimensions” are more abstract than ”units”: mass is a dimension, and kilograms is a scale unit whose dimension is mass. For each dimension, different standard systems will specify different units. For example, the dimension of force is mass × length/(time squared), and its units under the centimeter-gram-second (cgs) and the metric (MKS) systems are $g\cdot\ cm/s^2$ and $kg\cdot\ m/s^2$, respectively. In principle, the dimensions of other physical quantities can also be defined as the basic dimensions, which can replace the above-mentioned dimensions. For example, momentum, energy, or current can all be selected as the basic dimensions.

Some physicists do not think that temperature is a fundamental dimension, because temperature is expressed as the energy per degree of freedom of the particle, which can be expressed in terms of energy (or mass, length, time). Some physicists do not think that the amount of charge is the basic dimension, and think that the amount of charge can be expressed in terms of mass, length, and time. In addition, some physicists suspect that nature has physical quantities with incompatible fundamental dimensions.

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 $q$ which does not change after changing unit of every fundamental dimension is called dimensionless and is denoted by $[q] = 1$.

**Example 1.2.** Let $F$, $v$, $a$ and $p$ demote the force, the velocity, the acceleration and the pressure, respectively. Then

$$[F] = MLT^{-2}, \quad [v] = LT^{-1}, \quad [a] = LT^{-2}, \quad [p] = ML^{-1}T^{-2}\,.$$

The numbers $\pi$ and $e$ are dimensionless. The quantity $\dfrac{F}{ma}$, where $m$ denotes the mass, is also dimensionless.

## 1.1 Dimensional Methods

The cornerstone of dimensional analysis is known as the ***Buckingham Pi theorem*** which states that any physical law that expresses the relationship between multiple dimensional physical quantities corresponds to an equivalent law of the relationship between non-dimensional quantities.

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

**Example 1.3.** 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 $\rho$. Therefore,

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

or equivalently,

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

**Example 1.4.** 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 $\rho$, 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 $q_j$ can be written in terms of the fundamental dimensions as

$$[q_j] = L_1^{a_{1j}} L_2^{a_{2j}} \cdots L_n^{a_{nj}}$$

for some exponents $a_{1j}, a_{2j}, \cdots, a_{nj}$. The $n \times m$ matrix

$$\begin{array}{c} \\ L_1 \\ \vdots \\ L_n \end{array} \begin{array}{c} \begin{array}{ccc} q_1 & \cdots & q_m \end{array} \\ \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \end{array} \tag{1.1}$$

containing the exponents is called the ***dimension matrix*** (of $q_1, \cdots, q_m$ w.r.t. dimensions $L_1, \cdots, L_n$). The entries in the $j$-th column give the exponents for $q_j$ in terms of the powers of $L_1, \cdots, L_n$.

We note that the choices of different independent fundamental dimensions results in different dimension matrices; however, the rank of dimension matrices is well-defined.

**Remark 1.5.** We briefly explained why then rank of dimension matrices could be well-defined using one particular example. Suppose that two fundamental dimensions are chosen to describe a physical law: the first fundamental dimensions are simply the length, the mass, and the time denoted respectively by $L$, $M$, $T$, and the second fundamental dimensions are the force, the velocity, and the energy denoted respectively by $F$, $V$, $E$. Note that $F = LMT^{-2}$, $V = LT^{-1}$, $E = L^2MT^{-2}$ and $L = F^{-1}E$, $M = V^{-2}E$, $T = F^{-1}V^{-1}E$. Define two $3 \times 3$ matrices

$$\mathrm{A} = \begin{array}{c|ccc} & F & V & E \\ \hline L & 1 & 1 & 2 \\ M & 1 & 0 & 1 \\ T & -2 & -1 & -2 \end{array} \qquad \text{and} \qquad \mathrm{B} = \begin{array}{c|ccc} & L & M & T \\ \hline F & -1 & 0 & -1 \\ V & 0 & -2 & -1 \\ E & 1 & 1 & 1 \end{array}$$

that are collections of exponents in these relations. Then the dimension matrix $\mathrm{D}_1$ and $\mathrm{D}_2$ of a set of quantities $q_1, \cdots, q_m$ w.r.t. the first fundamental dimension $L$, $M$, $T$ and the second fundamental dimensions $F$, $V$, $E$, respectively, has the relation

$$\mathrm{D}_2 = \mathrm{B}\mathrm{D}_1 \qquad \text{and} \qquad \mathrm{D}_1 = \mathrm{A}\mathrm{D}_2 \,.$$

Since $\mathrm{AB} = \mathrm{I}_3$, we find that the rank of $\mathrm{D}_1$ and $\mathrm{D}_2$ are the same.

**Definition 1.6.** Let $q_1, \cdots, q_m$ be dimensioned quantities.

1. A quantity $\pi$ is called a ***dimensionless combinations*** of $q_1, \cdots, q_m$ if $\pi = q_1^{\alpha_1} \cdots q_m^{\alpha_m}$ for some numbers $\alpha_1, \cdots, \alpha_m$ and $[\pi] = 1$.

2. A collection $\{\pi_1, \cdots, \pi_k\}$ of dimensionless combinations of $q_1, \cdots, q_m$ is said to be ***maximal*** if any dimensionless quantities $\pi$ formed from $q_1, \cdots, q_m$ can be expressed as $\pi = \pi_1^{c_1} \cdots \pi_k^{c_k}$ for some unique $c_1, \cdots, c_k$.

**Remark 1.7.** The number $k$ in the definition above is in fact the nullity of the dimension matrix. To see this, we first note that the fact that $[q_j] = L_1^{a_{1j}} L_2^{a_{2j}} \cdots L_n^{a_{nj}}$ for $1 \leqslant j \leqslant m$ implies that

$$\begin{aligned}\left[q_1^{\alpha_1} q_2^{\alpha_2} \cdots q_m^{\alpha_m}\right] = [q_1]^{\alpha_1}[q_2]^{\alpha_2} \cdots [q_m]^{\alpha_m} &= \prod_{j=1}^{m} (L_1^{a_{1j}} L_2^{a_{2j}} \cdots L_n^{a_{nj}})^{\alpha_j} \\ &= L_1^{a_{11}\alpha_1 + a_{12}\alpha_2 + \cdots + a_{1m}\alpha_m} L_2^{a_{21}\alpha_1 + a_{22}\alpha_2 + \cdots + a_{2m}\alpha_m} \cdots L_n^{a_{n1}\alpha_1 + a_{n2}\alpha_2 + \cdots + a_{nm}\alpha_m} .\end{aligned}$$

Therefore,

$$\begin{array}{l} q_1^{\alpha_1} q_2^{\alpha_2} \cdots q_m^{\alpha_m} \text{ is a dimensionless} \\ \text{combination of } q_1,\, q_2,\, \cdots,\, q_m \end{array} \quad \Leftrightarrow \quad \begin{cases} a_{11}\alpha_1 + a_{12}\alpha_2 + \cdots + a_{1m}\alpha_m = 0\,, \\ a_{21}\alpha_1 + a_{22}\alpha_2 + \cdots + a_{2m}\alpha_m = 0\,, \\ \qquad\qquad\vdots \\ a_{n1}\alpha_1 + a_{n2}\alpha_2 + \cdots + a_{nm}\alpha_m = 0\,. \end{cases}$$

In other words, with $D = [a_{ij}]_{n \times m}$ denoting the dimension matrix (of $q_1, \cdots, q_m$ w.r.t. fundamental dimensions $L_1, \cdots, L_n$), the statement above shows that

$$q_1^{\alpha_1} \cdots q_m^{\alpha_m} \text{ is a dimensionless combination of } q_1, \cdots, q_m \text{ if and only if } D \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \mathbf{0}. \tag{1.2}$$

This shows that every dimensionless combination $\pi$ of $q_1, q_2, \cdots, q_m$ corresponds to a unique vector $\boldsymbol{\alpha}$ in the null space of the dimension matrix $D$ and vice versa.

Now suppose that $\{\pi_1, \pi_2, \cdots, \pi_k\}$, where $\pi_j = q_1^{\beta_{1j}} q_2^{\beta_{2j}} \cdots q_m^{\beta_{mj}}$, is a collection of dimensionless combinations of $q_1, q_2, \cdots, q_m$. Define $B = [\beta_{ij}]_{m \times k}$ and $\boldsymbol{\beta}_j$ is the $j$-th column of $B$; that is, $\boldsymbol{\beta}_j = [\beta_{1j}, \beta_{2j}, \cdots, \beta_{mj}]^{\mathrm{T}}$ for $1 \leqslant j \leqslant k$. Then with $N(D)$ denoting the null space of $D$, (1.2) shows that $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_k\} \subseteq N(D)$. Similar computation also shows that

$$\pi_1^{c_1} \pi_2^{c_2} \cdots \pi_k^{c_k} = q_1^{\beta_{11}c_1 + \beta_{12}c_2 + \cdots + \beta_{1k}c_k} q_2^{\beta_{21}c_1 + \beta_{22}c_2 + \cdots \beta_{2k}c_k} \cdots q_m^{\beta_{m1}c_1 + \beta_{m2}c_2 + \cdots + \beta_{mk}c_k} .$$

Therefore, the definition of maximal collection of dimensionless combinations implies that

$\{\pi_1, \pi_2, \cdots, \pi_k\}$ is a maximal collection of dimensionless combinations of $q_1, \cdots, q_m$

$\Leftrightarrow$ for all dimensionless combination $\pi$ of $q_1, \cdots, q_k$ there exists a unique vector $(c_1, \cdots, c_k) \in \mathbb{R}^k$ such that $\pi = \pi_1^{c_1} \pi_2^{c_2} \cdots \pi_k^{c_k}$

$\Leftrightarrow$ for all $[\alpha_1, \cdots, \alpha_m]^{\mathrm{T}} \in N(D)$ there exists a unique vector $(c_1, \cdots, c_k) \in \mathbb{R}^k$ such that $\beta_{j1}c_1 + \beta_{j2}c_2 + \cdots + \beta_{jk}c_k = \alpha_j$ for $1 \leqslant j \leqslant m$

$\Leftrightarrow$ for all $\boldsymbol{\alpha} \in N(D)$ there exists a unique $\boldsymbol{c} \in \mathbb{R}^k$ such that $B\boldsymbol{c} = \boldsymbol{\alpha}$

$$\Leftrightarrow \begin{cases} 1.\ \ \forall\, \boldsymbol{\alpha} \in N(D), \boldsymbol{\alpha} \in \operatorname{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_k), \\ 2.\ \ \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_k\} \text{ is a linearly independent set.} \end{cases}$$

$\Leftrightarrow$ $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_k\}$ is a basis of $N(D)$.

The statement above shows that $k =$ the nullity of $D$, the dimension of $N(D)$. Moreover, if $r$ is the rank of $D$, then the dimension theorem in Linear Algebra implies that $k = m - r$.

Any fundamental dimension $L_k$ has the property that its units can be changed upon multiplication by the appropriate conversion factor to obtain a new value in a new system of units. Let $\{[L_1]_1, \cdots, [L_n]_1\}$ and $\{[L_1]_2, \cdots, [L_n]_2\}$ be two particular choices of units for fundamental dimension. For example, if $L_1$ is the time dimension, then $[L_1]_1$ and $[L_1]_2$ could be chosen as second and year, respectively. Then for each $1 \leqslant k \leqslant n$, $[L_k]_2 = \lambda_k [L_k]_1$ for some dimensionless constant $\lambda_k > 0$. The units of derived quantities $q$ then can be changed in the fashion that if

$$[q] = L_1^{b_1} L_2^{b_2} \cdots L_n^{b_n}, \tag{1.3}$$

and $v_1(q)$ denotes the value of $q$ in the system of units $\{[L_1]_1, \cdots, [L_n]_1\}$, then

$$v_2(q) = \lambda_1^{b_1} \lambda_2^{b_2} \cdots \lambda_n^{b_n} v_1(q) \tag{1.4}$$

gives the value of $q$ in the new system of units $\{[L_1]_2, \cdots, [L_n]_2\}$.

**Definition 1.8.** Let $q_1, q_2, \cdots, q_m$ be dimensioned quantities. The physical law

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

is said to be ***unit free*** if for all positive real numbers $\lambda_1, \cdots, \lambda_n$,

$$f\big(v_1(q_1), \cdots, v_1(q_m)\big) = 0 \quad \text{if and only if} \quad f\big(v_2(q_1), \cdots, v_2(q_m)\big) = 0\,,$$

where $v_1(q_j)$ and $v_2(q_j)$ are related by (1.4) if $q_j$ obeys (1.3).

**Theorem 1.9** (Buckingham's Pi Theorem)**.** *Suppose that*

$$f(q_1, q_2, \cdots, q_m) = 0 \tag{1.6}$$

*is a unit free physical law that relates the dimensioned quantities $q_1, q_2, \cdots, q_m$. Then there exists a maximal collection $\{\pi_1, \pi_2, \cdots, \pi_k\}$ of dimensionless combinations of $q_1, \cdots, q_m$ and the physical law* (1.6) *is equivalent to an equation*

$$F(\pi_1, \cdots, \pi_k) = 0$$

*expressed only in terms of the dimensionless quantities.*

*Proof.* Let $D = [a_{ij}]_{n\times m}$ be the dimension matrix of $q_1$, $\cdots$, $q_m$ w.r.t. a given fundamental dimensions $L_1$, $\cdots$, $L_n$, and $r = \text{rank}(D)$. Suppose that $\pi = q_1^{\alpha_1} q_2^{\alpha_2} \cdots q_m^{\alpha_m}$ is a dimensionless quantities. Then with $\boldsymbol{\alpha}$ denoting the column vector $[\alpha_1, \cdots, \alpha_m]^{\mathrm{T}}$, (1.2) implies that

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

where $\mathbf{0}$ denotes the zero vector in $\mathbb{R}^n$. Since $\text{rank}(D) = r$, 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 uniquely 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 $\boldsymbol{\alpha}$ so that $\alpha_1, \cdots, \alpha_r$ is uniquely determined by $\alpha_{r+1}, \cdots, \alpha_m$. Assume that the vector $\boldsymbol{\alpha}(1:r)$ is given by

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

and let $\pi_1, \cdots, \pi_{m-r}$ be given by

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

Then $\{\pi_1, \cdots, \pi_{m-r}\}$ is a a maximal collection of dimensionless combinations of $q_1, \cdots, q_r$ (by the dimension theorem in Linear Algebra). Define

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

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

$$v_2(q_j) = \lambda_1^{a_{1j}} \lambda_2^{a_{2j}} \cdots \lambda_n^{a_{nj}} v_1(q_j), \quad 1 \leqslant j \leqslant r,$$

we must have $G\big(v_2(q_1), \cdots, v_2(q_r), \pi_1, \cdots, \pi_{m-r}\big) = 0$. Since the columns of $D(:, 1:r)$ are linearly independent and $n \geqslant r$, there exist $\lambda_1, \cdots, \lambda_n$ (might not be unique if $n > r$) such that

$$\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{nr} \end{bmatrix} \begin{bmatrix} \log \lambda_1 \\ \log \lambda_2 \\ \vdots \\ \log \lambda_n \end{bmatrix} = \begin{bmatrix} -\log v_1(q_1) \\ -\log v_1(q_2) \\ \vdots \\ -\log v_1(q_r) \end{bmatrix} \tag{1.7}$$

Choose $\lambda_1, \cdots, \lambda_n$ satisfying (1.7). Then in the new system of units $v_2(q_j) = 1$ for all $1 \leqslant j \leqslant r$; thus we establish that as long as $q_1, \cdots, q_r$ satisfy $G(q_1, \cdots, q_r, \pi_1, \cdots, \pi_{m-r}) = 0$, there exists a system of units such that $v_2(q_1) = \cdots = v_2(q_r) = 1$. This implies that $G$ is independent of $q_1, \cdots, q_r$ and we have

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

**Example 1.10** (Example 1.3 - 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 $F, \rho, A, v$. 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, $F = k\rho Av^2$) for some (dimensionless) constant $k$.

**Example 1.11** (Example 1.4 - 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 $t, r, \rho, E$. 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 1.12.** 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 $\theta$ as a function of $r$ and $t$.

Clearly the temperature $\theta$ 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

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

This physical law has 6 dimensioned quantities

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

The dimension matrix (with the order of dimension $T, L, \Theta, E$) 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

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

In other words,

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

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

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

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

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

**Example 1.13.** In this example we determine the relation between the power $P$ that must be applied to keep a ship of length $\ell$ moving at a constant speed $V$. Assume that $P$ depends on the density of water $\varrho$, the acceleration due to gravity $g$, and the viscosity of water $\nu$ (in length-squared per time), as well as $\ell$ and $V$. The physical law is given by

$$f(P, \varrho, g, \nu, \ell, V) = 0\,.$$

Suppose that the fundamental dimension is the time $T$, the length $L$, and the mass $M$. Then

$$[P] = ML^2T^{-3}\,,\ [\varrho] = ML^{-3}\,,\ [g] = LT^{-2}\,,\ [\nu] = L^2T^{-1}\,,\ [\ell] = L \quad \text{and} \quad [V] = LT^{-1}\,.$$

Therefore, the dimension matrix (in the order $T$, $L$, $M$) is

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

which has rank 3. By the Pi Theorem, there are three dimensionless quantities $\pi_1$, $\pi_2$ and $\pi_3$ and the physical law $f(P, \varrho, g, \nu, \ell, V) = 0$ is equivalent to $F(\pi_1, \pi_2, \pi_3) = 0$ (or sometimes $\pi_1 = F(\pi_2, \pi_3)$).

If $\pi = P^{\alpha_1}\varrho^{\alpha_2}g^{\alpha_3}\nu^{\alpha_4}\ell^{\alpha_5}V^{\alpha_6}$ is dimensionless, then

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

Three choices of $(\alpha_1, \cdots, \alpha_6)$ are

$$(1, -1, 0, 0, -2, -3), \quad \big(0, 0, -\frac{1}{2}, 0, -\frac{1}{2}, 1\big) \quad \text{and} \quad (0, 0, 0, -1, 1, 1)$$

which implies that the physical law is equivalent to

$$\frac{P}{\varrho\ell^2V^3} = F\Big(\frac{V}{\sqrt{\ell g}}, \frac{V\ell}{\nu}\Big).$$

The two dimensionless quantities $\dfrac{V}{\sqrt{\ell g}}$ and $\dfrac{V\ell}{\nu}$ are called the Froude number Fr and the Reynolds number Re, respectively, so that the equality above can be rewritten as

$$\frac{P}{\varrho\ell^2V^3} = F(\mathrm{Fr}, \mathrm{Re}).$$

**Example 1.14.** 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*,

$$\frac{d^2h}{dt^2} = -\frac{R^2g}{(R+h)^2}, \qquad h(0) = 0, \quad h'(0) = V. \tag{1.8}$$

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 $\big(-1, 2, -1\big)$ to form

$$\pi_1 = \frac{h}{R}\,, \qquad \pi_2 = \frac{tV}{R}\,, \qquad \pi_3 = \frac{V^2}{gR}\,.$$

Therefore, the Pi theorem "implies" that there exists a function $\widetilde{\Phi}$ such that $\widetilde{\Phi}(\pi_1, \pi_2, \pi_3) = 0$ which "implies" that $\pi_1 = \Phi(\pi_2, \pi_3)$; thus

$$\frac{h}{R} = \Phi\big(\frac{tV}{R}, \frac{V^2}{gR}\big)\,.$$

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

$$0 = h'(t_{\max}) = R\frac{d}{dt}\Big|_{t=t_{\max}} \Phi\big(\frac{tV}{R}, \frac{V^2}{gR}\big) = V\frac{\partial \Phi}{\partial \pi_2}\big(\frac{t_{\max}V}{R}, \frac{V^2}{gR}\big)\,.$$

The above relation "implies" that $\dfrac{t_{\max}V}{R}$ is a function of $\dfrac{V^2}{gR}$; thus

$$\frac{t_{\max}V}{R} = F\big(\frac{V^2}{gR}\big)\,.$$

## 1.2 Characteristic Scales and Scaling

The "characteristic scales" are some specific chosen values of dimensioned quantities in the problem under consideration. The use of characteristic scales can help us reduce mathematical model into dimensionless form, and a good choice of characteristic scales sometimes can even simplify complicated models into simple ones.

**Example 1.15** (Example 1.14 - revisit)**.** In this example we choose characteristic time scale $t_c$ and length scale $\ell_c$ to recast the ODE (1.8)

$$\frac{d^2h}{dt^2} = -\frac{R^2 g}{(R+h)^2}\,, \qquad h(0) = 0\,, \quad h'(0) = V\,. \tag{1.8}$$

We note that in pratice we know the values of $R$, $g$ and $V$, so we should choose characteristic scales according to these values.

Define the dimensionless time $\bar{t} = t/t_c$ and dimensionless height $\bar{h} = h/\ell_c$ $\big($so that

$\bar{h}(\bar{t}) = \dfrac{h(t_c\bar{t})}{\ell_c}$). With the dimensionless time $\bar{t}$ and dimensionless height $\bar{h}$, ODE (1.8) is equivalent to the dimensionless ODE

$$\frac{d^2\bar{h}}{d\bar{t}^2} = -\frac{t_c^2 g}{\ell_c}\frac{1}{(1+\frac{\ell_c}{R}\bar{h})^2}\,, \qquad \bar{h}(0) = 0\,, \quad \bar{h}'(0) = \frac{t_c V}{\ell_c}\,. \tag{1.9}$$

Since the dimension of the known dimensioned quantities in (1.8) are

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

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 (1.9) becomes

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

2. Let $t_c = R/V$ and $\ell_c = V^2/g$. Then (1.9) becomes

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

3. Let $t_c = \sqrt{R/g}$ and $\ell_c = R$. Then (1.9) becomes

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

4. Let $t_c = \sqrt{R/g}$ and $\ell_c = V^2/g$. Then (1.9) becomes

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

5. Let $t_c = V/g$ and $\ell_c = R$. Then (1.9) becomes

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

6. Let $t_c = V/g$ and $\ell_c = V^2/g$. Then (1.9) becomes

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

We note that these six ODEs are equivalent; however, we look for further simplification if the parameter $\epsilon$ is very small (or very large).

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 $\epsilon$ (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 1.16.** 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 rate of change of the population $\dfrac{dp}{dt}$ 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).

To complete the system, as in ODE (1.8) we need to impose an initial condition so that the complete equation is

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

In the logistic model above, the dimension of $t$ is time, and the dimension of population is named "population". Let $t_c$ and $p_c$ denote the characteristic time scale and the characteristic population scale, respectively. Introducing the dimensionless time $\bar{t} = t/t_c$ and the dimensionless population $\bar{p} = p/p_c$ $\big($so that $\bar{p}(\bar{t}) = \dfrac{p(t_c\bar{t})}{p_c}\big)$, we obtain the following dimensionless ODE

$$\frac{d\bar{p}}{d\bar{t}} = rt_c\bar{p}\big(1 - \frac{p_c}{K}\bar{p}\big)\,, \qquad \bar{p}(0) = \frac{p_0}{p_c}\,. \tag{1.10}$$

Apparently, we should choose the characteristic time scale $t_c = 1/r$, while two characteristic population scales can be chosen: $p_c = K$ or $p_c = p_0$. Moreover, there is a dimensionless quantity $\epsilon = \dfrac{p_0}{K}$ in the system.

1. $p_c = K$: the (1.10) becomes $\dfrac{d\bar{p}}{d\bar{t}} = \bar{p}\big(1 - \bar{p}\big)$ with initial data $\bar{p}(0) = \epsilon$.

2. $p_c = p_0$: the (1.10) becomes $\dfrac{d\bar{p}}{d\bar{t}} = \bar{p}\big(1-\epsilon\bar{p}\big)$ with initial data $\bar{p}(0)=1$.

If $\epsilon \ll 1$, we are tempted to delete the terms involving $\epsilon$ (or simply setting $\epsilon = 0$) in the scaled problem. Only case 2 provides a reasonable interpretation of the real phenomena.

**Example 1.17.** 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:

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

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

where $\rho$ 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 $p$, $\mu$ is the dynamical viscosity with dimension of mass per length per time , and

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

Let $\ell_c$ 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

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

where $\boldsymbol{v}\cdot\nabla_y\boldsymbol{v}$, $\Delta_y\boldsymbol{v}$ and $\mathrm{div}_y\boldsymbol{v}$ 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

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

## 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 1.18.** 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 $p$ and $r$.

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 $\infty$ 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 1.19** (Hölder's inequality)**.** Suppose that one knows that for some $p, q, r, s \in \mathbb{R}$, we have the following inequality

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

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

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

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 1.18 shows that $pr = 1$, $qs = 1$ and $-n = -nr - ns$; thus $r = \frac{1}{p}$, $s = \frac{1}{q}$ and we have

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

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

**Remark 1.20.** 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 1.21.** 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 $\Omega$ 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 1.22** (Sobolev's inequalities)**.** The simplest Sobolev's inequalities is of the form

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

where $C$ is a generic constant independent of $f$, and $\mathscr{C}_c^1(\mathbb{R}^n)$ 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 $f \in \mathscr{C}_c^1(\mathbb{R}^n)$. For given constants $M, L > 0$, define $u(x) = Mf(Lx)$. Then $u \in \mathscr{C}_c^1(\mathbb{R}^n)$; 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

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

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.

**Definition 2.3.** The ordinary differential equation

$$F(t,y,y',\cdots,y^{(n-1)},y^{(n)}) = y^{(n)}(t) - f(t,y,y',\cdots,y^{(n-2)},y^{(n-1)}) = 0 \tag{2.1}$$

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

$$\begin{aligned} &F(t,cy,cy',\cdots,cy^{(n-1)},cy^{(n)}) - F(t,0,0,\cdots,0) \\ &\qquad = c\big[F(t,y,y',\cdots,y^{(n-1)},y^{(n)}) - F(t,0,0,\cdots,0)\big] \end{aligned} \qquad \forall\, c \in \mathbb{R}\,. \tag{2.2}$$

The ODE (2.1) is said to be ***nonlinear*** if it is not linear.

**Remark 2.4.** 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 $\boldsymbol{z} = \big(y, y', y'', \cdots, y^{(n-1)}\big)^{\mathrm{T}}$ and write the ODE above as

$$\boldsymbol{z}'(t) = \frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \\ f(t, z_1, z_2, \cdots, z_n) \end{bmatrix} = \boldsymbol{f}(t, \boldsymbol{z}) \tag{2.3}$$

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

## 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 1.16 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 rate of change of the amount of the plutonium $\frac{dp}{dt}$ 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 (for population growrth or decay of radioactive substance).

**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

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

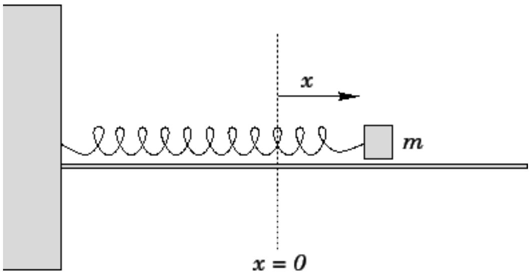


Figure 2.1: The spring-mass system

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 since the function

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

satisfies (2.2).

**Example 2.8.** In this example we study a closed circuit shown in the figure below.

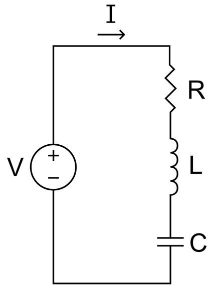


Figure 2.2: A closed circuit

In the figure above, V is the voltage (電壓) source powering the circuit, I is the current (電流) admitted through the circuit, R is the effective resistance (電阻) of the combined load, source, and components, L is the inductance of the inductor (電感) component, and C the capacitance of the capacitor (電容) component.

An electric current (電流) is the rate of flow of electric charge (電荷) past a point or region:

$$\mathrm{I}(t) = \frac{d\mathrm{Q}}{dt}\,.$$

A capacitor (電容) consists of two conductors separated by a non-conductive region which can either be a vacuum or an electrical insulator material known as a dielectric (介電質). From Coulomb's law (庫倫定律) a charge on one conductor will exert a force on the charge carriers within the other conductor, attracting opposite polarity charge and repelling like polarity charges, thus an opposite polarity charge will be induced on the surface of the other conductor. The conductors thus hold equal and opposite charges on their facing surfaces, and the dielectric develops an electric field. An ideal capacitor is characterized by a constant capacitance C which is defined as the ratio of the positive or negative charge Q on each conductor to the voltage V between them:

$$\mathrm{C} = \frac{\mathrm{Q}}{\mathrm{V}} \qquad \text{or} \qquad \mathrm{Q} = \mathrm{CV}\,.$$

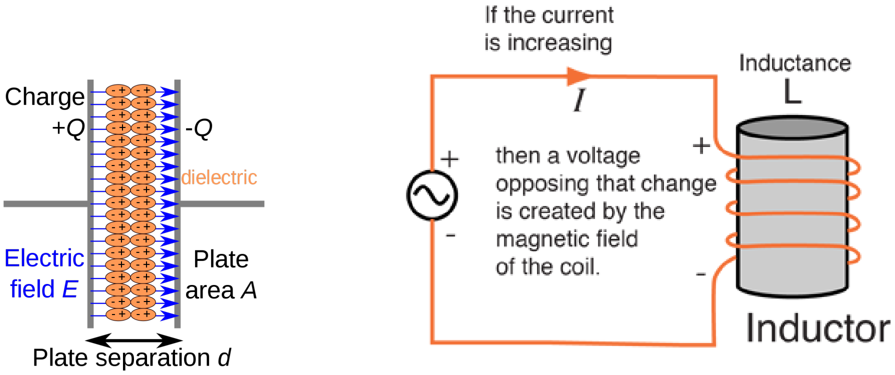
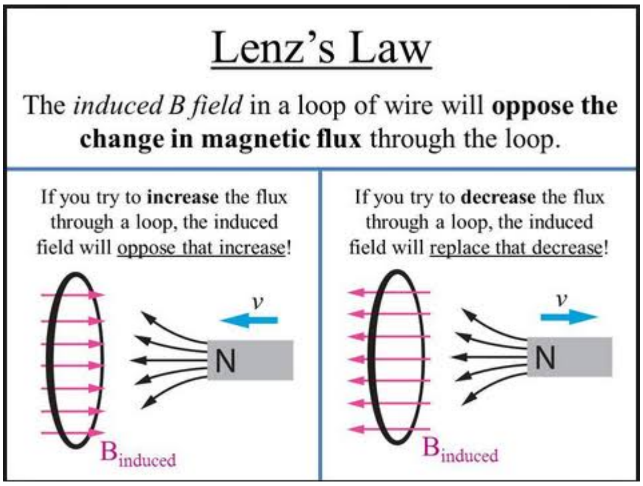

Charge +Q
-Q
dielectric
Electric field *E*
Plate area *A*
Plate separation *d*
If the current is increasing
*I*
Inductance L
+
-
then a voltage opposing that change is created by the magnetic field of the coil.
+
-
Inductor

Figure 2.3: Left: capacitor, Right: inductance

Inductance (電感) is the tendency of an electrical conductor to oppose a change in the electric current flowing through it, and is defined as the ratio of the induced voltage to the rate of change of current causing it:

$$\mathrm{V}(t) = \mathrm{L}\frac{d\mathrm{I}}{dt}\,.$$

The design of inductance is based on Lenz's law (冷次定律) which states that "the current induced in a circuit due to a change in a magnetic field is directed to oppose the change in flux and to exert a mechanical force which opposes the motion" (磁通量的改變而產生的感應電流，其方向為抗拒磁通量改變的方向).


Lenz's Law

The *induced B field* in a loop of wire will **oppose the change in magnetic flux** through the loop.

If you try to **increase** the flux through a loop, the induced field will oppose that increase!

If you try to **decrease** the flux through a loop, the induced field will replace that decrease!

In a closed circuit (a circuit without interruption, providing a continuous path through which a current can flow) shown in Figure 2.2, one has

$$\mathrm{V}(t) = \mathrm{I}(t)\mathrm{R} + \mathrm{L}\frac{d\mathrm{I}}{dt} + \frac{1}{\mathrm{C}}\mathrm{Q}(t)\,.$$

By the definition of the electric current I, we find that Q satisfies

$$\mathrm{L}\frac{d^2\mathrm{Q}}{dt^2}+\mathrm{R}\frac{d\mathrm{Q}}{dt}+\frac{1}{\mathrm{C}}\mathrm{Q}=\mathrm{V}(t)\,.$$

To complete the model, initial conditions have to be imposed so that we have

$$\mathrm{L}\frac{d^2\mathrm{Q}}{dt^2}+\mathrm{R}\frac{d\mathrm{Q}}{dt}+\frac{1}{\mathrm{C}}\mathrm{Q}=\mathrm{V}(t)\,,\qquad \mathrm{Q}(t_0)=\mathrm{Q}_0\,,\quad \mathrm{Q}'(t_0)=\mathrm{I}_0\,.$$

We note that the IVP above is essentially the same as the IVP (2.5).

**Example 2.9** (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.

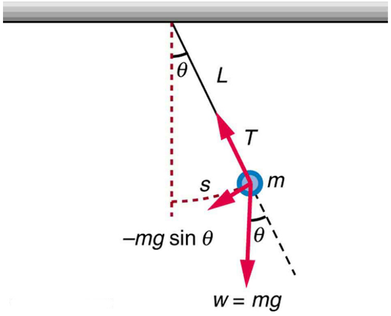


Figure 2.4: A simple pendulum system

Let $\theta(t)$ denote the angle, measured from the vertical dashed line (see Figure 2.4), 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 in the IVP above is a **nonlinear** ODE. We also note that when the angle of oscillation is very small; that is, $\theta\approx 0$, then by the fact that $\lim\limits_{\theta\to 0}\frac{\sin\theta}{\theta}=1$ we find that in this case

$$mL\ddot{\theta}\approx -mg\theta\,;$$

thus we obtain a simplified model for simple pendulum

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

The simplified model for oscillating pendulum is essentially "the same as" the model for the spring-mass system without the friction and the external force.

**Example 2.10** (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 $\delta pq$, $\delta > 0$. 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 obtain 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 $F(t, c\boldsymbol{z}, c\dot{\boldsymbol{z}}) - F(t, 0, 0) \neq c\big[F(t, \boldsymbol{z}, \dot{\boldsymbol{z}}) - F(t, 0, 0)\big]$ if $c \neq 1$, where

$$F(t, \boldsymbol{z}, \dot{\boldsymbol{z}}) = \dot{\boldsymbol{z}} - \boldsymbol{f}(t, \boldsymbol{z})\,.$$

**Example 2.11** (SIR model for spread of diseases)**.** This example presents a classical model, called the SIR model, of disease transmission within a population. The total population is divided into three groups: individuals susceptible to disease, infected individuals, and "removed" individuals. The removed class counts those individuals who are not infected and not susceptible; in other words, immune, quarantined, or dead. Individuals may move from one class to another; for example, an individual may move from the infected class to the removed class upon recovery. Thus the model accounts for the interdependency of the different classes within the population.

The fundamental relation of the SIR model is the relation

$$N = S(t) + I(t) + R(t)\,, \tag{2.7}$$

where $N$ is the total population size, taken to be constant; $S(t)$ is the size of the susceptible population, $I(t)$ is the size of the infected population, and $R(t)$ is the size of the removed population. We note that (2.7) shows that the rate of change of $S$, $I$ and $R$ must obey the following identity

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0\,. \tag{2.8}$$

The derivation of the SIR model is similar to the prey-predator model: the roles of the infected group and the susceptible group are respectively similar to the predator and the prey in the prey-predator model, except that the assumption of a fixed amount of total population prohibits the growth of the susceptible group. The population of the infected group, without the presence of the susceptible group, decays due to the recovery from the disease and increases due to contact with the susceptible group. On the other hand, the only way an individual leaves the susceptible group is by becoming infected (due to contact with the infected group). Therefore, we obtain the following differential equation

$$\frac{dS}{dt} = -bS(t)I(t)\,, \tag{2.9a}$$
$$\frac{dI}{dt} = -\gamma I(t) + bS(t)I(t)\,, \tag{2.9b}$$

where $b$ is termed effective disease transmission, and $\gamma$ is the recovery rate. Because of (2.8), we find that

$$\frac{dR}{dt} = \gamma I(t)\,. \qquad (2.9c).$$

The equation above explains the term recovery rate.

Sometimes (2.9) is written in the following form:

$$\frac{dS}{dt} = -\frac{\beta I(t)}{N} S(t)\,, \tag{2.10a}$$
$$\frac{dI}{dt} = -\gamma I(t) + \frac{\beta I(t)}{N} S(t)\,, \tag{2.10b}$$
$$\frac{dR}{dt} = \gamma I(t)\,, \tag{2.10c}$$

where $\beta = Nb$ is called the transmission rate.

In epidemiology (流行病學), the **basic reproduction number**, denoted by $R_0$, of an infection is the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. In the SIR model, $R_0 = \beta/\gamma$.

**Example 2.12.** 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 2.5).

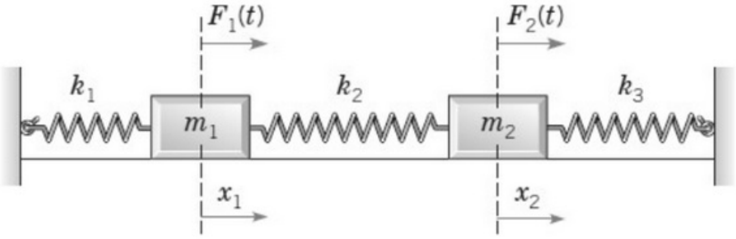


Figure 2.5: A two-mass, three-spring system

Let $L_1$, $L_2$, $L_3$ be the length of the unconstrained springs, and $\ell_1$, $\ell_2$, $\ell_3$ be the increment of the springs in equilibrium. Then

$$k_1\ell_1 = k_2\ell_2 = k_3\ell_3 \,. \tag{2.11}$$

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

$$\begin{aligned} m_1\frac{d^2x}{dt^2} &= -k_1(x-L_1)+k_2(y-x-L_2)+F_1\,, && \text{(2.12a)}\\ m_2\frac{d^2y}{dt^2} &= -k_2(y-x-L_2)+k_3(L_1+L_2+L_3+\ell_1+\ell_2+\ell_3-y-L_3)+F_2\\ &= -k_2(y-x-L_2)+k_3(L_1+L_2+\ell_1+\ell_2+\ell_3-y)+F_2\,. && \text{(2.12b)} \end{aligned}$$

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 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

$$\begin{aligned} m_1\frac{d^2x_1}{dt^2} &= -k_1x_1+k_2(x_2-x_1)+F_1\,, && \text{(2.13a)}\\ m_2\frac{d^2x_2}{dt^2} &= -k_2(x_2-x_1)-k_3x_2+F_2\,. && \text{(2.13b)} \end{aligned}$$

We note that (2.13) is "the same as" letting $L_1 = L_2 = L_3 = \ell_1 = \ell_2 = \ell_3 = 0$ in (2.12).

Equation (2.13) a second order linear ODE, and it becomes an IVP if initial conditions $x_1(t_0) = x_{10}$, $x_2(t_0) = x_{20}$, $x_1'(t_0) = x_{11}$ and $x_2'(t_0) = x_{21}$ are imposed.

**Example 2.13** (Planetary motion)**.** In this example, we consider the orbit of a planet moving around the Sun in the solar system. Suppose that planet under consideration is Earth. Since Earth moves on the ecliptic plane（黃道面）, we can treat the orbit of Earth as a plane curve on the $xy$-plane. Let the origin of the $xy$-plane be the center of mass of the Sun, and the location of Earth at time $t$ be $\boldsymbol{r}(t) = \boldsymbol{x}(t)\mathbf{i}+y(t)\mathbf{j}$, where $\mathbf{i}$ and $\mathbf{j}$ are pre-chosen but fixed directions of Cartesian coordinates. Then Newton's second law of motion implies that

$$-\frac{GMm}{\|\boldsymbol{r}(t)\|^3}\boldsymbol{r}(t) = m\boldsymbol{r}''(t)\,, \tag{2.14}$$

where $M$ and $m$ denote the mass of the Sun and Earth, respectively, and $\|\boldsymbol{r}(t)\|$ is the distance from Earth to the Sun at time $t$.

We note that the two unknowns of the ODE (2.14) are indeed $x(t)$ and $y(t)$. To study the motion of Earth better, a polar coordinate representation of the ODE is need. We introduce a polar coordinate system in which the pole of the polar coordinate system is the Sun, and the polar axis is $\mathbf{i}$. Let $\big(r(t), \theta(t)\big)$ be the polar coordinate of the location of Earch at time $t$; that is, $\boldsymbol{r}(t) = r(t)\cos\theta(t)\mathbf{i} + r(t)\sin\theta(t)\mathbf{j}$, and define two vectors

$$\widehat{r}(t) = \cos\theta(t)\mathbf{i} + \sin\theta(t)\mathbf{j} \qquad \text{and} \qquad \widehat{\theta}(t) = -\sin\theta(t)\mathbf{i} + \cos\theta(t)\mathbf{j}$$

accordingly. Then $\boldsymbol{r}(t) = r(t)\widehat{r}(t)$. By the fact that

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

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, (2.14) 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 find that the polar coordinate $\big(r(t), \theta(t)\big)$ of Earch must satisfy the nonlinear ODE

$$-\frac{GM}{r^2} = r'' - r(\theta')^2\,, \tag{2.15a}$$

$$2r'\theta' + r\theta'' = 0\,. \tag{2.15b}$$

Since (2.15) is a second-order ODE, to make it an initial value problems we need to specify the values of $r(t_0)$, $\theta(t_0)$, $r'(t_0)$ and $\theta'(t_0)$.

**Kepler's second law**: Note that (2.15b) implies that $(r^2\theta')' = 0$; thus $r^2\theta'$ is a constant. Let $\ell$ be the constant angular momentum so that

$$\ell = mr^2\theta' = m\mathrm{r}_0\mathrm{v}_0\,, \tag{2.16}$$

where $\mathrm{r}_0$ and $\mathrm{v}_0$ denote the perihelion distance (近日點距) and the speed at the perihelion (近日點), respectively. Note that (2.16) shows that $\theta'$ is sign-definite (unless $\ell = 0$), so $\theta : I \to \mathbb{R}$ is one-to-one.

Let $t_1 < t_2$. The area swept out in the time interval $[t_1, t_2]$ is given by

$$\int_{t_1}^{t_2} \frac{1}{2} r^2(t)\theta'(t)\,dt = \int_{t_1}^{t_2} \frac{\ell}{2m}\,dt = \frac{\ell(t_2 - t_1)}{2m} = \frac{\mathrm{r}_0\mathrm{v}_0}{2}(t_2 - t_1)\,; \tag{2.17}$$

thus we conclude that

> A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.

This is Kepler's second law of planetary motion.

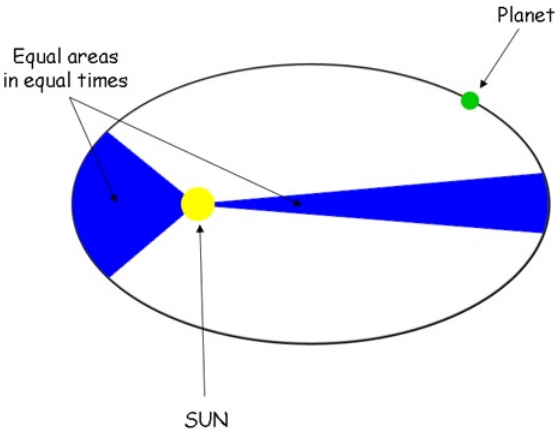


Figure 2.6: Kepler's second law of planetary motion

**Remark 2.14.** The angular momentum of a moving object relative to a point is the cross product of the particle's position vector $\boldsymbol{r}$ (relative to the point) and its momentum vector $\boldsymbol{p}$ (relative to the point as well). Therefore, the angular momentum of the planet relative to the Sun is

$$\boldsymbol{r} \times m\boldsymbol{r}' = mr\,\widehat{r} \times (r'\,\widehat{r} + r\theta'\widehat{\theta}) = mr^2\theta'\widehat{r} \times \widehat{\theta} = mr^2\theta'\mathbf{k}\,;$$

thus the quantity $mr^2\theta'$ is the angular momentum of the planet relative to the sun. (2.15b) (or (2.16)) then implies that the angular momentum is a constant, so-called the ***conservation of angular momentum***（角動量守恆）.

**Example 2.15.** Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a continuously 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" 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 continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we can 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 2.16** (Existence and Uniqueness of Solution/Fundamental theorem of ODE)**.** *Consider the initial value problem*

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}) \,, \qquad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \in \mathbb{R}^n \,, \tag{2.18}$$

*where $\boldsymbol{x}$ and $\boldsymbol{f}$ are functions with values in $\mathbb{R}^n$. If $\boldsymbol{f}$ and the first partial derivatives of $\boldsymbol{f}$ with respect to all its variables, possibly except $t$, are continuous functions in some rectangular domain $R = [a, b] \times [c_1, d_1] \times [c_2, d_2] \times \cdots \times [c_n, d_n]$ that contains the point $(t_0, \boldsymbol{x}_0)$ in the interior, then the initial value problem* (2.18) *has a unique solution $\boldsymbol{x} = \boldsymbol{\varphi}(t)$ in some interval $I = (t_0 - h, t_0 + h)$ for some positive number $h$. Moreover, the solution is continuously differentiable on $I$.*

**Remark 2.17.** Since every $n$-th order IVP can be rewritten in the form of (2.18), by the theorem above we also conclude that every $n$-th order IVP has a unique solution provided that the right-hand side function has required properties. To be more precise, following Remark 2.4 we rewrite the IVP

$$y^{(n)} = f(t, y, y', \cdots, y^{(n-1)}) \,, \qquad y(t_0) = y_0 \,, y'(t_0) = y_1 \,, \cdots, y^{(n-1)}(t_0) = y_{n-1} \tag{2.19}$$

as

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}) \,, \qquad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \,,$$

where

$$\boldsymbol{x} = \begin{bmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} \quad \text{and} \quad \boldsymbol{f}(t, \boldsymbol{x}) = \mathrm{N}\boldsymbol{x} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ f(t, \boldsymbol{x}) \end{bmatrix}$$

in which $\mathrm{N} = [n_{ij}]$ is the constant matrix given by $n_{k,k+1} = 1$ for $1 \leqslant k \leqslant n-1$ and $n_{ij} = 0$ elsewhere. Then $\dfrac{\partial \boldsymbol{f}}{\partial x_k} = \mathrm{N}\mathbf{e}_k + \begin{bmatrix} 0 & \cdots & 0 & \dfrac{\partial f}{\partial y^{(k)}} \end{bmatrix}^{\mathrm{T}}$ so $\dfrac{\partial \boldsymbol{f}}{\partial x_k}$ is continuous if and only if $\dfrac{\partial f}{\partial y^{(k)}}$ is continuous. This verifies the statement above.

In particular, if the ODE in IVP (2.18) is linear; that is,

$$f(t, y, y', \cdots, y^{(n-1)}) = a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y + g(t) \,, \tag{2.20}$$

then clearly the first partial derivative of $f$ w.r.t. all the "y-variables" are continuous if $a_0, a_1, \cdots, a_{n-1}$ are continuous (on an open interval containing $t_0$). Therefore, if the coefficients (that is, $a_0, a_1, \cdots, a_{n-1}$ in (2.20)) and the forcing (that is, $g$ in (2.20)) of a linear ODE are continuous in an open interval containing $t_0$, then the solution of IVP (2.19) is uniquely determine by the initial data $y_0, \cdots, y_{n-1}$.

## 2.2 Some Basic Techniques of Solving ODEs

### 2.2.1 Separation of variables

The simplest ODE takes the form

$$x'(t) = g(t)f(x(t))\,,$$

where $f$ and $g$ are given (Lipschitz) continuous functions. Formally we let $\Phi$ and $G$ be an anti-derivative of $\frac{1}{f}$ and $g$, respectively. Then

$$\frac{d}{dt}\Phi(x(t)) = \Phi'(x(t))x'(t) = \frac{x'(t)}{f(x(t))} = g(t)$$

which implies that $\Phi(x(t)) = G(t) + C$ for some constant $C$. A general solution $x(t)$ then is obtained by inverting the function $\Phi$.

If an initial condition $x(t_0) = x_0$ is provided, then we can choose $\Phi$ and $G$ satisfying $\Phi(x_0) = G(t_0)$ so that

$$\Phi(x(t)) - \Phi(x_0) = \int_{t_0}^{t} \frac{d}{dt}\Phi(x(s))\,ds = \int_{t_0}^{t} g(s)\,ds = G(t) - G(t_0)\,;$$

thus $\Phi(x(t)) = G(t)$ which further shows that $x(t) = \Phi^{-1}(G(t))$ (as long as $\Phi$ has an inverse function).

**Example 2.18.** Consider the logistic equation

$$p' = rp\Big(1 - \frac{p}{K}\Big)$$

introduced in Example 1.16. Letting $f(p) = rp\Big(1 - \frac{p}{K}\Big)$, we have

$$\int \frac{dp}{f(p)} = \frac{K}{r}\int \frac{dp}{p(K-p)} = \frac{1}{r}\int \Big(\frac{1}{p} + \frac{1}{K-p}\Big)\,dp = \frac{1}{r}\big(\ln|p| - \ln|K-p|\big) + C\,.$$

Therefore, an anti-derivative of $\frac{1}{f}$ is

$$\Phi(p) = \frac{1}{r}\ln\Big|\frac{p}{K-p}\Big| + C$$

whose inverse function, when considering the case $0 < p < K$ (which is the case if $0 < p(t_0) < K$), is given by

$$\Phi^{-1}(t) = \frac{Ke^{r(t-C)}}{1 + e^{r(t-C)}} = \frac{KDe^{rt}}{1 + De^{rt}}\,,$$

where $D = e^{-Cr}$. Solutions to the logistic equation above is then $p(t) = \frac{KDe^{rt}}{1 + De^{rt}}$.

We note that in practice the following procedure is often used:

$$\begin{aligned}\frac{dp}{dt} = rp\Big(1 - \frac{p}{K}\Big) &\Rightarrow \frac{dp}{p(1 - p/K)} = rdt \Rightarrow \int \frac{dp}{p(1-p/K)} = \int rdt\\ &\Rightarrow \int \frac{Kdp}{p(K-p)} = rt + C \Rightarrow \int \Big(\frac{1}{p} + \frac{1}{K-p}\Big)\,dp = rt + C\\ &\Rightarrow \ln|p| - \ln|K-p| = rt + C \Rightarrow \ln\Big|\frac{p}{K-p}\Big| = rt + C\end{aligned}$$

and then find $p$ as before. One should be able to see the similarity between finding $\Phi$ and the procedure above.

### 2.2.2 The method of integrating factors

A first-order ODE does not always take the form given in the previous sub-section. The simplest first-order ODE, when the right-hand side $f$ does not take that form, is linear. Consider the first-order linear ODE

$$x'(t) + q(t)x(t) = r(t)\,,$$

where $p, q$ are given continuous functions defined on a certain interval. Let $Q$ denote an antiderivative of $q$. Note that

$$\frac{d}{dt}\big[e^{Q(t)}x(t)\big] = e^{Q(t)}x'(t) + e^{Q(t)}q(t)x(t) = e^{Q(t)}\big[x'(t) + q(t)x(t)\big]\,;$$

thus

$$\frac{d}{dt}\big[e^{Q(t)}x(t)\big] = e^{Q(t)}r(t)\,. \tag{2.21}$$

The equation above implies that

$$e^{Q(t)}x(t) = \int e^{Q(t)}r(t)\,dt$$

so we have

$$x(t) = e^{-Q(t)}\int e^{Q(t)}r(t)\,dt\,.$$

Suppose now we are given an initial condition $x(t_0) = x_0$. Then we integrate both sides of (2.21) from $t_0$ to $t$ and obtain that

$$\int_{t_0}^{t} \frac{d}{ds}\big[e^{Q(s)}x(s)\big]\,ds = \int_{t_0}^{t} e^{Q(s)}r(s)\,ds\,.$$

The Fundamental Theorem of Calculus further implies that

$$e^{Q(t)}x(t) - e^{Q(t_0)}x(t_0) = \int_{t_0}^{t} e^{Q(s)}r(s)\,ds\,;$$

thus

$$x(t) = e^{Q(t_0)-Q(t)}x_0 + \int_{t_0}^{t} e^{Q(s)-Q(t)}r(s)\,ds\,. \tag{2.22}$$

Formula (2.22) gives the solution to the initial value problem

$$x'(t) + q(t)x(t) = r(t)\,, \qquad x(t_0) = x_0\,.$$

**Example 2.19.** The solution to the initial value problem

$$p'(t) = rp(t)\,, \qquad p(0) = p_0$$

is $p(t) = p_0e^{rt}$. The solution $p$ can also be obtained using separation of variables.

### 2.2.3 Second-order linear ODEs

Consider the second-order linear ODE

$$x''(t) + b(t)x'(t) + c(t)x(t) = f(t)\,, \tag{2.23}$$

where $b, c$ and $f$ are given continuous functions.

We first consider the case $f \equiv 0$. In this case, the ODE is said to be ***homogeneous***, and the theory of differential equations shows that the solution space (that is, the collection of solutions) is two dimensional. In other words, there exist two linearly independent solutions $\varphi_1$ and $\varphi_2$ such that every solution $x$ can be written as the linear combination of $\varphi_1$ and $\varphi_2$ or equivalently,

$$x(t) = C_1\varphi_1(t) + C_2\varphi_2(t) \quad \text{for some constants } C_1 \text{ and } C_2.$$

**Remark 2.20.** Suppose that $b$ and $c$ are continuous. Let $\varphi_1$ and $\varphi_2$ be solutions of the homogeneous ODE

$$x'' + b(t)x' + c(t)x = 0 \tag{2.24}$$

satisfying the initial conditions $\varphi_1(t_0) = \varphi_2'(t_0) = 1$ and $\varphi_1'(t_0) = \varphi_2(t_0) = 0$ (the existence of such $\varphi_1$ and $\varphi_2$ are guaranteed by the continuity of $b$ and $c$). Then the solution of the IVP

$$x'' + b(t)x' + c(t)x = 0\,, \qquad x(t_0) = x_0\,, x'(t_1) = x_1$$

can be expressed as $x(t) = x_0\varphi_1(t) + x_1\varphi_1(t)$; thus $\{\varphi_1, \varphi_2\}$ is a basis of the solution space of (2.24). This shows that the solution space of (2.24) is two dimensional.

- **The case that $b$ and $c$ are constants**

In general, it is not easy to find linearly independent solutions to homogeneous ODEs. Nevertheless, if $b(t) = b$ and $c(t) = c$ are constant functions, we can find linearly independent solution by looking at the characteristic equation

$$r^2 + br + c = 0\,. \tag{2.25}$$

Suppose that $r_1$ and $r_2$ are two real zeros of (2.25) (it is possible that $r_1 = r_2$). Then $b = -(r_1 + r_2)$ and $c = r_1r_2$. Define $y = x' - r_1x$. Then

$$\begin{aligned} y' - r_2y &= \frac{d}{dt}(x' - r_1x) - r_2(x' - r_1x) = x'' - (r_1 + r_2)x' + r_1r_2x \\ &= x'' + bx' + cx = 0\,. \end{aligned}$$

Therefore, separation of variables or the method of integrating factor implies that $y(t) = Ce^{r_2t}$. This in turn implies that $x$ satisfies $x' - r_1x = Ce^{r_2t}$, and the method of integrating factor then shows that

$$\frac{d}{dt}\big[e^{-r_1t}x(t)\big] = Ce^{(r_2-r_1)t}\,.$$

1. If $r_1 \neq r_2$, then
$$e^{-r_1 t}x(t) = \frac{C}{r_2 - r_1}e^{(r_2-r_1)t} + D\,;$$
thus $x(t) = \dfrac{C}{r_2 - r_1}e^{r_2 t} + De^{r_1}t$.
2. If $r_1 = r_2 = r$, ten
$$e^{-rt}x(t) = Ct + D\,;$$
thus $x(t) = Cte^{rt} + De^{rt}$.

How about if the characteristic equation (2.25) has complex roots? Since we assume that $b$ and $c$ are real, the complex roots of (2.25) must be conjugate to each other. Suppose that the zeros of (2.25) are $\alpha \pm i\beta$, where $\alpha, \beta \in \mathbb{R}$ and $\beta \neq 0$, so that $b = -2\alpha$ and $c = \alpha^2 + \beta^2$. Define $y(t) = e^{-\alpha t}x(t)$. Then $x(t) = e^{\alpha t}y(t)$ which shows that

$$\frac{d^2}{dt^2}\big[e^{\alpha t}y(t)\big] - 2\alpha\frac{d}{dt}\big[e^{\alpha t}y(t)\big] + (\alpha^2 + \beta^2)e^{\alpha t}y(t) = 0\,.$$

Expanding the derivatives, we find that

$$\alpha^2 e^{\alpha t}y(t) + 2\alpha e^{\alpha t}y'(t) + e^{\alpha t}y''(t) - 2\alpha^2 e^{\alpha t}y(t) - 2\alpha e^{\alpha t}y'(t) + (\alpha^2 + \beta^2)e^{\alpha t}y(t) = 0\,;$$

thus

$$y''(t) + \beta^2 y(t) = 0\,.$$

By the existence and uniqueness of the solution to (first-order) ODE, the solution $y$ to the ODE above exists and is uniquely determined by the initial condition $y(0) = y_0$ and $y'(0) = y_1$. The solution $y$ must be twice continuously differentiable. Since $y'' = -\beta^2 y$, we find that $y$ is four times continuously differentiable (on the interval the solution exists). Repeated this argument, we conclude that the solution $y$ to the initial value problem

$$y''(t) + \beta^2 y(t) = 0\,, \qquad y(0) = y_0\,, y'(0) = y_1 \tag{2.26}$$

is infinitely many times differentiable. Moreover, one has

$$y^{(2k)}(0) = (-\beta^2)^k y_0 \qquad y^{(2k+1)}(0) = (-\beta^2)^k y_1 \qquad \forall\, k \in \mathbb{N} \cup \{0\}\,.$$

This enables us to consider the Maclaurin series of $y$:

$$\begin{aligned}\sum_{k=0}^{\infty}\frac{y^{(k)}(0)}{k!}t^k &= \sum_{k=0}^{\infty}\Big[\frac{(-\beta^2)^k y_0}{(2k)!}t^{2k} + \frac{(-\beta^2)^k y_1}{(2k+1)!}t^{2k+1}\Big]\\ &= \sum_{k=0}^{\infty}\Big[y_0\frac{(-1)^k}{(2k)!}(\beta t)^{2k} + \frac{y_1}{\beta}\frac{(-1)^k}{(2k+1)!}(\beta t)^{2k+1}\Big]\,,\end{aligned}$$

and the use of the Maclaurin series of sine and cosine shows that

$$\sum_{k=0}^{\infty}\frac{y^{(k)}(0)}{k!}t^k = y_0\cos(\beta t) + \frac{y_1}{\beta}\sin(\beta t)\,.$$

Note that the right-hand side function are indeed a solution to the initial value problem (2.26), so the uniqueness of the solution shows that

$$y(t) = y_0 \cos(\beta t) + \frac{y_1}{\beta} \sin(\beta t) \,.$$

As a consequence, the solution to $x'' + bx' + cx = 0$, where $b$ and $c$ are real constants so that the characteristic equation (2.25) has complex roots $\alpha \pm \beta i$ for some $\alpha, \beta \in \mathbb{R}$, is given by

$$x(t) = C_1 e^{\alpha t} \cos(\beta t) + C_2 e^{\alpha t} \sin(\beta t) \,.$$

We summarize the discussion above as follows:

1. If (2.25) has two distinct real zeros $r_1$ and $r_2$, then $x'' + bx' + cx = 0$ has two linearly independent solutions
$$\varphi_1(t) = e^{r_1 t} \qquad \text{and} \qquad \varphi_2(t) = e^{r_2 t} \,.$$
2. If (2.25) has a repeated real zero $r$, then $x'' + bx' + cx = 0$ has two linearly independent solutions
$$\varphi_1(t) = e^{rt} \qquad \text{and} \qquad \varphi_2(t) = te^{rt} \,.$$
3. If (2.25) has complex zeros $\alpha \pm i\beta$, where $\alpha, \beta$ are real numbers, then $x'' + bx' + cx = 0$ has two linearly independent solutions
$$\varphi_1(t) = e^{\alpha t} \cos(\beta t) \qquad \text{and} \qquad \varphi_2(t) = e^{\alpha t} \sin(\beta t) \,.$$

When considering initial value problem, the constants $C_1$ and $C_2$ are determined by the initial conditions.

**Example 2.21.** Consider the spring-mass system

$$m\ddot{x} + kx = 0 \,, \qquad x(0) = x_0 \,, \quad x'(0) = v_0 \,. \tag{2.27}$$

Rewrite the equation above as $\ddot{x} + \omega^2 x = 0$, where $\omega = \sqrt{k/m}$. Since the corresponding characteristic equation has two complex zeros $\pm\omega i$, we find that

$$x(t) = C_1 \cos(\omega t) + C_2 \sin(\omega t) \,.$$

Using the initial data, we find that $C_1 = x_0$ and $C_2 = v_0/\omega$; thus the solution to (2.27) is given by

$$x(t) = x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t) = R \cos(\omega t - \phi) \,,$$

where $R = \sqrt{x_0^2 + \dfrac{v_0^2}{\omega^2}}$ and $\phi$ satisfies $\cos\phi = \dfrac{x_0}{R}$ and $\sin\phi = \dfrac{v_0}{R\omega}$.

- **The case that $b$ or $c$ is not a constant**

If $b$ or $c$ is not constant, there is a way to find a second solution which is linearly independent to a **known** non-zero solution. Suppose that $x = \varphi_1(t)$ satisfies

$$x''(t) + b(t)x'(t) + c(t)x(t) = 0\,. \tag{2.28}$$

We look for a solution $\varphi_2$ of the form $\varphi_2(t) = v(t)\varphi_1(t)$. Then

$$\begin{aligned} 0 &= \varphi_2''(t) + b(t)\varphi_2'(t) + c(t)\varphi_2(t) \\ &= v''(t)\varphi_1(t) + 2v'(t)\varphi_1'(t) + v(t)\varphi_1''(t) + b(t)\big[v'(t)\varphi_1(t) + v(t)\varphi_1'(t)\big] + c(t)v(t)\varphi(t) \\ &= v''(t)\varphi_1(t) + 2v'(t)\varphi_1'(t) + b(t)v'(t)\varphi_1(t) + v(t)\big[\varphi_1''(t) + b(t)\varphi_1'(t) + c(t)\varphi_1(t)\big] \\ &= v''(t)\varphi_1(t) + v'(t)\big[2\varphi_1'(t) + b(t)\varphi_1(t)\big]\,. \end{aligned}$$

The equation above is an first order ODE for $y(t) = v'(t)$ and can be solved using the method of integrating factor: since $y$ satisfies

$$y' + \frac{2\varphi_1'(t) + b(t)\varphi_1(t)}{\varphi_1(t)} y(t) = 0\,,$$

with $B$ denoting an anti-derivative of $b$ we have

$$y(t) = C\exp\Big(-\int \frac{2\varphi_1'(t) + b(t)\varphi_1(t)}{\varphi_1(t)}\,dt\Big) = C\exp\Big(-2\ln\big|\varphi(t)\big| - B(t)\Big) = \frac{C}{\varphi_1(t)^2}e^{-B(t)}\,.$$

Therefore, another solution $\varphi_2$ is given by

$$\varphi_2(t) = v(t)\varphi_1(t) = \varphi_1(t)\int \frac{1}{\varphi_1(t)^2}e^{-B(t)}\,dt\,. \tag{2.29}$$

**Example 2.22.** Given that $y = \varphi_1(t) = \dfrac{1}{t}$ is a solution of

$$2t^2x'' + 3tx' - x = 0 \qquad \text{for } t > 0\,, \tag{2.30}$$

find a linearly independent solution of the equation.

Rewrite (2.30) as

$$x'' + \frac{3}{2t}x' - \frac{1}{2t^2}x = 0\,.$$

Using (2.29), we find that a linearly independent second solution is given by

$$\varphi_2(t) = \varphi_1(t)\int \frac{1}{\varphi_1(t)^2}e^{-B(t)}\,dt = \frac{1}{t}\int t^2\exp\Big(-\frac{3}{2}\ln t\Big)\,dt = \frac{2}{3}\sqrt{t}\,.$$

Now we consider the general case that $f$ is not the zero function. In this case, the theory of differential equations shows that the solution to (2.23) can be expressed as

$$x(t) = C_1\varphi_1(t) + C_2\varphi_2(t) + x_p(t)$$

for some constants $C_1$ and $C_2$, where $\{\varphi_1, \varphi_2\}$ is a basis of the solution space of the corresponding homogeneous ODE (2.28), and $x_p$ is a function, called a particular solution of

(2.23). One such a particular solution can be found using the method of variation of parameters as follows. Suppose that $x_p(t) = C_1(t)\varphi_1(t) + C_2(t)\varphi_2(t)$ for some functions $C_1$ and $C_2$ to be determined.

First we assume that $C_1$, $C_2$ satisfy

$$C_1'(t)\varphi_1(t) + C_2'(t)\varphi_2(t) = 0\,.$$

Then $x_p'(t) = C_1(t)\varphi_1'(t) + C_2(t)\varphi_2'(t)$ which further implies that

$$\begin{aligned}
f(t) &= x_p''(t) + bx_p'(t) + cx_p(t)\\
&= C_1'(t)\varphi_1'(t) + C_1(t)\varphi_1''(t) + C_2'(t)\varphi_2'(t) + C_2(t)\varphi_2''(t) + bC_1'(t)\varphi_1(t) + bC_2'(t)\varphi_2(t)\\
&\quad + cC_1(t)\varphi_1(t) + cC_2(t)\varphi_2(t)\\
&= C_1'(t)\varphi_1'(t) + C_2'(t)\varphi_2'(t) + C_1(t)\big[\varphi_1''(t) + b\varphi_1'(t) + c\varphi_1(t)\big]\\
&\quad + C_2(t)\big[\varphi_2''(t) + b\varphi_2'(t) + c\varphi_2(t)\big]\\
&= C_1'(t)\varphi_1'(t) + C_2'(t)\varphi_2'(t)\,.
\end{aligned}$$

Therefore, $C_1$ and $C_2$ satisfy

$$\begin{aligned}
C_1'(t)\varphi_1(t) + C_2'(t)\varphi_2(t) &= 0\,,\\
C_1'(t)\varphi_1'(t) + C_2'(t)\varphi_2'(t) &= f(t)\,;
\end{aligned}$$

thus with $W[\varphi_1, \varphi_2]$ denoting the function $\varphi_1\varphi_2' - \varphi_2\varphi_1'$ (termed the **Wronskian** of $\varphi_1$ and $\varphi_2$), we have

$$C_1'(t) = -\frac{f(t)\varphi_2(t)}{W[\varphi_1,\varphi_2](t)} \qquad \text{and} \qquad C_2'(t) = \frac{f(t)\varphi_1(t)}{W[\varphi_1,\varphi_2](t)}\,.$$

As a consequence, a particular solution of (2.23) is given by

$$x_p(t) = -\varphi_1(t)\int \frac{f(t)\varphi_2(t)}{W[\varphi_1,\varphi_2](t)}\,dt + \varphi_2(t)\int \frac{f(t)\varphi_1(t)}{W[\varphi_1,\varphi_2](t)}\,dt\,. \tag{2.31}$$

**Example 2.23.** Consider the spring-mass system

$$m\ddot{x} + kx = F_0\,, \qquad x(0) = x_0\,, \quad x'(0) = v_0\,, \tag{2.32}$$

where $F_0$ is a given constant. Let $\varphi_1(t) = \cos(\omega t)$ and $\varphi_2(t) = \sin(\omega t)$, where $\omega = \sqrt{k/m}$. Example 2.21 shows that $\{\varphi_1, \varphi_2\}$ is a basis for the solution space of the corresponding homogeneous ODE $m\ddot{x} + kx = 0$; thus by the fact that

$$W[\varphi_1,\varphi_2](t) = \varphi_1(t)\varphi_2'(t) - \varphi_1'(t)\varphi_2(t) = \omega\big[\cos^2(\omega t) + \sin^2(\omega t)\big] = \omega\,,$$

formula (2.31) implies that a particular solution is given by

$$\begin{aligned}
x_p(t) &= -\cos(\omega t)\int \frac{F_0/m\cdot\sin(\omega t)}{\omega}\,dt + \sin(\omega t)\int \frac{F_0/m\cdot\cos(\omega t)}{\omega}\,dt\\
&= \frac{F_0}{m\omega^2}\big[\cos^2(\omega t) + \sin^2(\omega t)\big] = \frac{F_0}{k}\,.
\end{aligned}$$

Therefore, the general solution to the ODE in 2.32 is

$$x(t) = C_1\cos(\omega t) + C_2\sin(\omega t) + \frac{F_0}{k}$$

and the initial conditions imply that $C_1 = x_0 - \dfrac{F_0}{k}$ and $C_2 = \dfrac{v_0}{\omega}$.

**Example 2.24** (Kepler's laws of planetary motion)**.** In this example we prove Kepler's first and third laws of planetary motion. Recall that in Example 2.13 we have shown that the polar coordinate $(r, \theta)$ of the location of a planet moving around a single sun satisfy a nonlinear second order ODE

$$-\frac{GM}{r^2} = r'' - r(\theta')^2\,, \tag{2.15a}$$

$$2r'\theta' + r\theta'' = 0\,. \tag{2.15b}$$

Moreover, in the proof of the second law of planetary motion we see that $\theta$ is one-to-one, and Theorem 2.16 shows that $\theta$ is continuously differentiable.

**Kepler's first law**: Since $\theta$ is one-to-one and continuously differentiable, the inverse function of $\theta$ exists and is also continuously differentiable (the Inverse Function Theorem for functions of one variable). Write $t = t(\theta)$, and every function of $t$ can be viewed as a function of $\theta$ (via $f(t) \mapsto f(t(\theta))$).

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 $\theta$) 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,

$$r'' = -\frac{\ell^2}{m^2r^2}\ddot{u} = -\frac{\ell^2}{m^2}\ddot{u}u^2\,; \tag{2.33}$$

**Remark 2.25.** For those who are confused with the use of prime and dot in the derivation above, try the following:

1. Let $t = t(\theta)$ be the inverse function of $\theta = \theta(t)$. By the inverse function theorem (or differentiating the identity $\theta(t(\theta)) = \theta$),

$$t'(\theta) = \frac{1}{\theta'(t(\theta))}\,.$$

Because of (2.16) (which states that $mr^2\theta' = \ell$), we have

$$t'(\theta) = \frac{mr^2(t(\theta))}{\ell}\,. \tag{2.34}$$

2. Define $u(\theta) = \dfrac{1}{r(t(\theta))}$. Then (2.34) shows that

$$u'(\theta) = -\frac{1}{r(t(\theta))^2}\frac{d}{d\theta}r(t(\theta)) = -\frac{1}{r(t(\theta))^2}r'(t(\theta))t'(\theta) = -\frac{m}{\ell}r'(t(\theta))$$

which in turn further implies that

$$u''(\theta) = -\frac{m}{\ell} r''(t(\theta))t'(\theta) = -\frac{m^2}{\ell^2} r(t(\theta))^2 r''(t(\theta)) .$$

Therefore, rearranging terms shows that

$$r''(t(\theta)) = -\frac{\ell^2}{m^2}\frac{1}{r(t(\theta))^2}u''(\theta) = -\frac{\ell^2}{m^2}u(\theta)^2 u''(\theta)$$

which is exactly (2.33), except that here the derivatives w.r.t. $\theta$ is still denoted by primes not dots.

By the definition of $u$, the use of (2.16) and (2.33) in (2.15a) shows 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,

$$\ddot{u} + u = \frac{GMm^2}{\ell^2} = \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2} .$$

A particular solution $u_p$ to the ODE above is the constant function $u_p(\theta) = \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2}$; thus the general solution to the ODE above is

$$u(\theta) = C_1\cos\theta + C_2\sin\theta + \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2} = C\cos(\theta+\phi) + \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2} \tag{2.35}$$

for some non-negative constant $C$ $(=\sqrt{C_1^2+C_2^2})$ and angle $\phi$ that can be determined by the initial data. By the fact that $u = \frac{1}{r}$, we find that the polar equation for the orbit of the planet is

$$r = \frac{1}{C\cos(\theta+\phi) + \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2}} = \frac{A}{1+\mathrm{e}\cos(\theta+\phi)} , \tag{2.36}$$

where $A = \frac{\mathrm{r}_0^2\mathrm{v}_0^2}{GM}$ and $\mathrm{e} = AC$. The polar equation (2.36) represents a conic section (圓錐曲線) with eccentricity (離心率) e. Since we are considering the motion of a planet, (2.36) must represent an ellipse (so $0 < \mathrm{e} < 1$). This proves Kepler's first law of planetary motion:

> The orbit of every planet is an ellipse with the Sun at one of the two foci.

**Remark 2.26.** Since we have proved that the orbit of a planet must be an ellipse, unlike the case of parabola or hyperbola the angular parameter $\theta$ in (2.35) has not constraint and can be any real numbers. Therefore, the maximum of $u$ is given by the reciprocal of the perihelion and we have

$$\frac{1}{\mathrm{r}_0} = C + \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2} .$$

Therefore,

$$C = \frac{1}{\mathrm{r}_0} - \frac{GM}{\mathrm{r}_0^2\mathrm{v}_0^2} = \frac{1}{\mathrm{r}_0}\Big(1 - \frac{GM}{\mathrm{r}_0\mathrm{v}_0^2}\Big)$$

which further implies that the eccentricity e is given by $\mathrm{e} = \dfrac{\mathrm{r}_0 \mathrm{v}_0^2}{GM} - 1$ and the polar equation of the ellipse is given by

$$r = \frac{(1+\mathrm{e})\mathrm{r}_0}{1 + \mathrm{e}\cos(\theta+\phi)}\,. \tag{2.37}$$

**Remark 2.27.** The polar equation for a conic section is derived as follows. The pole of the polar coordinate is chosen to be (one of) the focus $F$, and the polar axis A is the ray perpendicular to and intersecting the directrix D. Let $P$ be a point on the conic section, and the polar coordinate of $P$ is $(r, \theta)$. If the distance from the focus to the directrix $L$, then the distance from $P$ to the focus $F$ is $r$, and the distance from $P$ to the directrix D is $L - r\cos\theta$. Therefore, by the definition of the eccentricity of conic sections,

$$\mathrm{e} = \frac{\text{the distance from } P \text{ to the focus } F}{\text{the distance from } P \text{ to the directrix D}} = \frac{r}{L - r\cos\theta}\,;$$

thus the polar equation for a conic section with eccentricity e is given by

$$r = \frac{\mathrm{e}L}{1 + \mathrm{e}\cos\theta}\,.$$

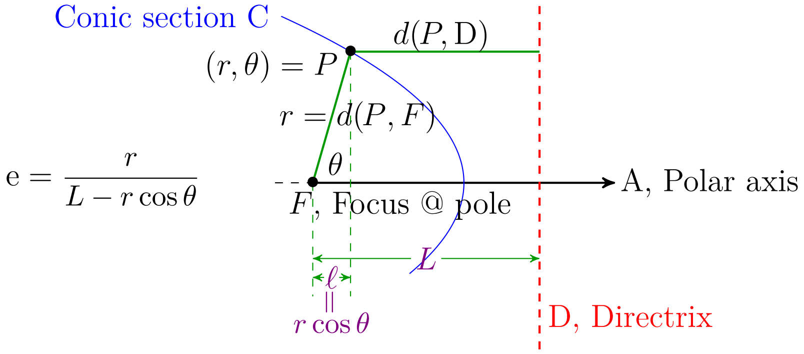


Figure 2.7: Polar representation of conic sections - Part 1

Suppose that the pole of a new polar coordinate system is the same focus but the polar axis A′ is given by $\theta = \phi$ in the polar coordinate system with polar axis A, then the polar equation for a conic section with eccentricity e is given by

$$r = \frac{\mathrm{e}L}{1 + \mathrm{e}\cos(\theta+\phi)}\,.$$

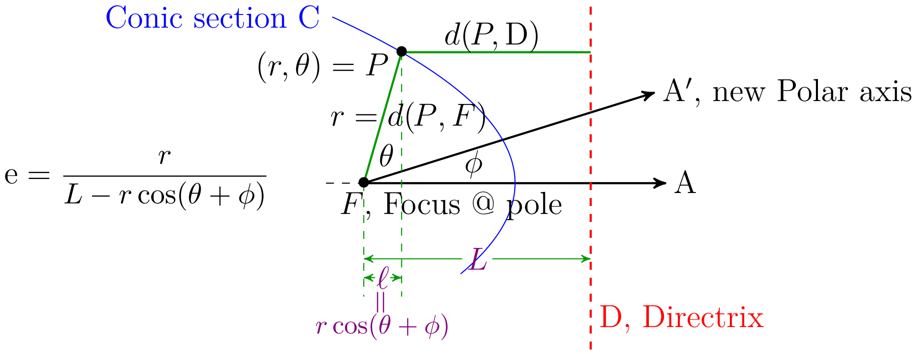


Figure 2.8: Polar representation of conic sections - Part 2

**Kepler's third law**: The third law of Kepler captures the relationship between the distance of planets from the Sun, and their orbital periods. Let $a, b$ be the semi-major axis and semi-minor axis of the orbit of a planet, and $\mathrm{T}$ be the orbital period. Using (2.17),

$$\pi ab = \int_0^{\mathrm{T}} \frac{1}{2} r^2 \theta' \, dt = \frac{\mathrm{r}_0 \mathrm{v}_0 \mathrm{T}}{2} \, .$$

Therefore, by the fact that $b = a\sqrt{1-\mathrm{e}^2}$,

$$\mathrm{T}^2 = \Big(\frac{2\pi ab}{\mathrm{r}_0 \mathrm{v}_0}\Big)^2 = \frac{4\pi^2 a^4}{\mathrm{r}_0^2 \mathrm{v}_0^2}(1-\mathrm{e}^2) = \frac{4\pi^2 a^4}{GM} \cdot \frac{2GM - \mathrm{r}_0 \mathrm{v}_0^2}{\mathrm{r}_0 GM} \, . \tag{2.38}$$

On the other hand, using (2.37) we find that $r_{\max} = r\big|_{\theta+\phi=\pi} = \mathrm{r}_0 \dfrac{1+\mathrm{e}}{1-\mathrm{e}}$; thus using the expression of e,

$$a = \frac{\mathrm{r}_0 + r_{\max}}{2} = \frac{\mathrm{r}_0}{1-\mathrm{e}} = \frac{\mathrm{r}_0 GM}{2GM - \mathrm{r}_0 \mathrm{v}_0^2} \, .$$

Using the identity above in (2.38), we conclude that

$$\mathrm{T}^2 = \frac{4\pi^2}{GM} a^3$$

which shows the third law of Kepler:

> The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

### 2.2.4 Linear systems with constant coefficients

A general linear system of ODEs takes the form

$$\begin{aligned} \frac{dx_1}{dt} &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + f_1(t) \, , \\ \frac{dx_2}{dt} &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n + f_2(t) \, , \\ &\vdots \qquad\qquad\qquad \vdots \\ \frac{dx_n}{dt} &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + f_n(t) \, , \end{aligned} \tag{2.39}$$

where the coefficients $a_{ij}$, where $1 \leqslant i, j \leqslant n$, and the forcing $f_1, \cdots, f_n$ are given functions. The linear system (2.39) is said to be **homogeneous** if $f_i(t) = 0$ for all $1 \leqslant i \leqslant n$; otherwise it is inhomogeneous.

Let $\boldsymbol{A}(t) = [a_{ij}(t)]_{n\times n}$, and $\boldsymbol{f}(t) = \big[f_1(t), \cdots, f_n(t)\big]^{\mathrm{T}}$. By Remark 2.17, we know that if $\boldsymbol{f}$ and $a_{ij}$ are continuous for all $i, j$, then for every given initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0 \in \mathbb{R}^n$ there exists a unique solution to the IVP

$$\boldsymbol{x}' = \boldsymbol{A}(t)\boldsymbol{x} + \boldsymbol{f}(t) \, , \qquad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \, .$$

In this sub-section, we always assume that $\boldsymbol{A}$ and $\boldsymbol{f}$ are continuous.

Imitating the method of integrating factor, we look for a matrix-valued function $\boldsymbol{M} = \boldsymbol{M}(t)$ such that

$$\frac{d}{dt}\big[\boldsymbol{M}(t)\boldsymbol{x}(t)\big] = \boldsymbol{M}\big[\boldsymbol{x}'(t) - \boldsymbol{A}(t)\boldsymbol{x}(t)\big] \tag{2.40}$$

so that once such an $\boldsymbol{M}$ is obtain, we have

$$\boldsymbol{M}(t)\boldsymbol{x}(t) = \int \boldsymbol{M}(t)\boldsymbol{f}(t)\,dt$$

and the inversion of $\boldsymbol{M}$ leads to the solution to the linear system. Note that $\boldsymbol{M}$ satisfies (2.40) if and only if $\boldsymbol{M}$ satisfies $\boldsymbol{M}'(t)\boldsymbol{x}(t) = -\boldsymbol{M}(t)\boldsymbol{A}(t)\boldsymbol{x}(t)$. Since $\boldsymbol{x}$ is unknown yet, we instead look for a matrix-valued function $\boldsymbol{M} = \boldsymbol{M}(t)$ satisfying

$$\boldsymbol{M}'(t) = -\boldsymbol{M}(t)\boldsymbol{A}(t)\,.$$

Note that the ODE above can be expressed as

$$\frac{d}{dt}m_{ij}(t) = -\sum_{k=1}^{n} m_{ik}(t)a_{kj}(t) \qquad \forall\, 1 \leqslant i, j \leqslant n$$

or in the form of linear system

$$\frac{d}{dt}\begin{bmatrix} m_{11} \\ m_{21} \\ \vdots \\ m_{n1} \\ m_{12} \\ \vdots \\ m_{nn} \end{bmatrix} = -\begin{bmatrix} a_{11}(t) & \cdots & a_{1n}(t) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & a_{21}(t) & \cdots & a_{2n}(t) & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \ddots & & & \vdots \\ \vdots & & & & \ddots & & \ddots & & \vdots \\ \vdots & & & & & \ddots & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & a_{n1}(t) & \cdots & a_{nn}(t) \end{bmatrix}\begin{bmatrix} m_{11} \\ m_{21} \\ \vdots \\ m_{n1} \\ m_{12} \\ \vdots \\ m_{nn} \end{bmatrix};$$

thus by the continuity of $A$, for any given initial condition $\boldsymbol{M}(0) = \boldsymbol{M}_0$, there exists a unique solution to the IVP

$$\frac{d}{dt}\boldsymbol{M}(t) = -\boldsymbol{M}(t)\boldsymbol{A}(t)\,, \qquad \boldsymbol{M}(0) = \boldsymbol{M}_0\,. \tag{2.41}$$

In the following, we look for solutions of a linear system when all the $a_{ij}$'s are constant functions. In other words, we look for vector-valued function $\boldsymbol{x}(t) = \big[x_1(t), \cdots, x_n(t)\big]^{\mathrm{T}}$ satisfying the ODE

$$\boldsymbol{x}'(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{f}(t)\,,$$

where $\boldsymbol{A} = [a_{ij}]_{n\times n}$ is a constant matrix, $\boldsymbol{f}(t) = \big[f_1(t), \cdots, f_n(t)\big]^{\mathrm{T}}$. This amounts to choose $\boldsymbol{M} = \boldsymbol{M}(t)$ satisfying

$$\frac{d}{dt}\boldsymbol{M}(t) = -\boldsymbol{M}(t)\boldsymbol{A}\,. \tag{2.42}$$

We note that the ODE above shows that $\boldsymbol{M}$ is infinitely many times differentiable such that for all $k \in \mathbb{N}$,

$$\boldsymbol{M}^{(k)}(t) = -\boldsymbol{M}^{(k-1)}(t)\boldsymbol{A} = (-1)^2\boldsymbol{M}^{k-2}\boldsymbol{A}^2 = \cdots = (-1)^k\boldsymbol{M}(0)\boldsymbol{A}^k\,;$$

thus $\boldsymbol{M}^{(k)}(0) = (-1)^k \boldsymbol{M}(0)\boldsymbol{A}^k$ for all $k \in \mathbb{N}$. Therefore, a good guess of the solution to the initial value problem (2.41), with $\boldsymbol{A}(t) \equiv \boldsymbol{A}$ being a constant matrix, is given by the Maclaurin series

$$\sum_{k=0}^{\infty} \frac{\boldsymbol{M}^{(k)}(0)}{k!} t^k = \sum_{k=0}^{\infty} \frac{(-1)^k \boldsymbol{M}(0) A^k}{k!} t^k = \boldsymbol{M}_0 \sum_{k=0}^{\infty} \frac{1}{k!} (-tA)^k \,.$$

Assuming that the derivative of power series is obtained by term by term differentiation, the "power series" given above is indeed the solution to (2.41) if $\boldsymbol{A}(t) \equiv \boldsymbol{A}$ is a constant matrix; thus we "conclude" (with the correct assumption that the solution to (2.42) is given by

$$\boldsymbol{M}(t) = \boldsymbol{M}(0) \sum_{k=0}^{\infty} \frac{1}{k!} (-t\boldsymbol{A})^k \,.$$

The Maclaurin series of the exponential function $y = e^x$ motivates the following

**Definition 2.28.** Let $\boldsymbol{B}$ be an $n \times n$ matrix. The exponential of $\boldsymbol{B}$, denoted by $e^{\boldsymbol{B}}$, is the series

$$e^{\boldsymbol{B}} = \mathrm{I} + \boldsymbol{B} + \frac{1}{2!}\boldsymbol{B}^2 + \frac{1}{3!}\boldsymbol{B}^3 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{B}^k \,.$$

Having defined the exponential of square matrices, we conclude that

$$\frac{d}{dt}\boldsymbol{M}(t) = -\boldsymbol{M}(t)\boldsymbol{A} \qquad \Leftrightarrow \qquad \boldsymbol{M}(t) = \boldsymbol{M}(0)e^{-t\boldsymbol{A}} \,. \tag{2.43}$$

**Remark 2.29.** We note that the exponential of square matrices is given by an infinite series, so in principle we should check the convergence of the series before we can define it. Nevertheless, we will treat the convergence of the series as a fact for this requires some additional knowledge in analysis.

Before proceeding, let us establish a fundamental identity

$$e^{t\boldsymbol{B}} e^{s\boldsymbol{B}} = e^{(t+s)\boldsymbol{B}} \qquad \text{for all square matrices } \boldsymbol{B} \text{ and } t, s \in \mathbb{R} \,. \tag{2.44}$$

To see this, we note that $e^{t\boldsymbol{B}}$ commutes with $\boldsymbol{B}$ (that is, $e^{t\boldsymbol{B}}\boldsymbol{B} = \boldsymbol{B}e^{t\boldsymbol{B}}$) for all $t \in \mathbb{R}$ and satisfies

$$\frac{d}{dt} e^{t\boldsymbol{B}} = e^{t\boldsymbol{B}}\boldsymbol{B} \,.$$

Therefore, for each given $s \in \mathbb{R}$,

$$\frac{d}{dt}\Big[e^{t\boldsymbol{B}} e^{s\boldsymbol{B}} - e^{(t+s)\boldsymbol{B}}\Big] = e^{t\boldsymbol{B}}\boldsymbol{B}e^{s\boldsymbol{B}} - e^{(t+s)\boldsymbol{B}}\boldsymbol{B} = \Big[e^{t\boldsymbol{B}} e^{s\boldsymbol{B}} - e^{(t+s)\boldsymbol{B}}\Big]\boldsymbol{B} \,.$$

Using (2.43),

$$e^{t\boldsymbol{B}} e^{s\boldsymbol{B}} - e^{(t+s)\boldsymbol{B}} = \Big[e^{0\boldsymbol{B}} e^{s\boldsymbol{B}} - e^{(0+s)\boldsymbol{B}}\Big] e^{t\boldsymbol{B}} = 0 \,;$$

thus (2.44) is established.

Now we come back to solve for the ODE $\boldsymbol{x}'(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{f}(t)$. We choose $\boldsymbol{M}(0) = \mathrm{I}$ so that an integrating factor $\boldsymbol{M}$ is given by $\boldsymbol{M}(t) = e^{-t\boldsymbol{A}}$. Therefore, (2.40) implies that

$$\frac{d}{dt}\big[e^{-t\boldsymbol{A}}\boldsymbol{x}(t)\big] = e^{-t\boldsymbol{A}}\big[\boldsymbol{x}'(t) - \boldsymbol{A}\boldsymbol{x}(t)\big] = e^{-t\boldsymbol{A}}\boldsymbol{f}(t)\,.$$

The equation above shows that

$$\boldsymbol{x}(t) = e^{t\boldsymbol{A}}\int e^{-t\boldsymbol{A}}\boldsymbol{f}(t)\,dt\,. \tag{2.45}$$

If an initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ is imposed, the unique solution to the IVP is given by

$$\boldsymbol{x}(t) = e^{(t-t_0)\boldsymbol{A}}\boldsymbol{x}_0 + e^{t\boldsymbol{A}}\int_{t_0}^{t} e^{-s\boldsymbol{A}}\boldsymbol{f}(s)\,ds\,. \tag{2.46}$$

- **The computation of $e^{t\boldsymbol{B}}$ for square matrix $\boldsymbol{B}$**

From linear algebra, every square matrix $\boldsymbol{B}$ has a Jordan decomposition; that is, every square matrix $\boldsymbol{B}$ can be expressed as

$$\boldsymbol{B} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}\,,$$

where $\boldsymbol{J}$ takes the Jordan canonical form

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{J}_2 & \ddots & \boldsymbol{O} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{J}_\ell \end{bmatrix}$$

in which each $\boldsymbol{O}$ is zero matrix, and each Jordan block $\boldsymbol{J}_r$ is a square matrix of the form $\lambda\mathbf{I}$ or

$$\boldsymbol{J}_r = \lambda\mathbf{I} + \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix} = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \lambda & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{bmatrix} \tag{2.47}$$

for some eigenvalue $\lambda$ of $\boldsymbol{B}$. By the fact that $\boldsymbol{B}^k = \boldsymbol{P}\boldsymbol{J}^k\boldsymbol{P}^{-1}$ we have

$$e^{t\boldsymbol{B}} = \sum_{k=0}^{\infty}\frac{1}{k!}t^k\boldsymbol{B}^k = \boldsymbol{P}\Big(\sum_{k=0}^{\infty}\frac{1}{k!}t^k\boldsymbol{J}^k\Big)\boldsymbol{P}^{-1}\,.$$

Since

$$\boldsymbol{J}^k = \begin{bmatrix} \boldsymbol{J}_1^k & & & \\ & \boldsymbol{J}_2^k & & \\ & & \ddots & \\ & & & \boldsymbol{J}_\ell^k \end{bmatrix},$$

we have

$$e^{t\boldsymbol{B}} = \boldsymbol{P}\begin{bmatrix} \sum_{k=0}^{\infty} \frac{1}{k!}t^k\boldsymbol{J}_1^k & & & \\ & \sum_{k=0}^{\infty} \frac{1}{k!}t^k\boldsymbol{J}_2^k & & \\ & & \ddots & \\ & & & \sum_{k=0}^{\infty} \frac{1}{k!}t^k\boldsymbol{J}_\ell^k \end{bmatrix}\boldsymbol{P}^{-1} = \boldsymbol{P}\begin{bmatrix} e^{t\boldsymbol{J}_1} & & & \\ & e^{t\boldsymbol{J}_2} & & \\ & & \ddots & \\ & & & e^{t\boldsymbol{J}_\ell} \end{bmatrix}\boldsymbol{P}^{-1}.$$

For each $r$, since $\mathbf{I}$ commutes with the matrix $\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}$, we have

$$\boldsymbol{J}_r^k = \left(\lambda\mathbf{I} + \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}\right)^k = \sum_{j=0}^{k} C_j^k \lambda^{k-j} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}^j .$$

By the fact that

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}^j = \begin{matrix} \overbrace{\hspace{5em}}^{j\text{ columns of 0's here}} \hspace{14em} \\ \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & & \ddots & 0 & 1 & 0 & \cdots & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & 0 & 1 & 0 \\ \vdots & & & & & \ddots & 0 & 1 \\ \vdots & & & & & & \ddots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \end{matrix},$$

if $\boldsymbol{J}_r$ is an $m \times m$ matrix, we have

$$\boldsymbol{J}_r^k = \begin{bmatrix} \lambda^k & k\lambda^{k-1} & C_2^k\lambda^{k-2} & \cdots & \cdots & C_{m-1}^k\lambda^{k-m+1} \\ 0 & \lambda^k & k\lambda^{k-1} & \ddots & \ddots & C_{m-2}^k\lambda^{k-m+2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & C_2^k\lambda^{k-2} \\ \vdots & \cdots & \cdots & 0 & \lambda^k & k\lambda^{k-1} \\ 0 & \cdots & \cdots & \cdots & 0 & \lambda^k \end{bmatrix}.$$

Here $C_m^k$ is the number $\dfrac{k(k-1)\cdots(k-m+1)}{m!}$ so that $C_m^k = 0$ if $m > k$. Therefore, if $\boldsymbol{J}_r$ is an $m \times m$ matrix taking the form (2.47), using

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{1}{k!} t^k C_\ell^k \lambda^{k-\ell} &= \sum_{k=\ell}^{\infty} \frac{1}{k!} t^k C_\ell^k \lambda^{k-\ell} = \sum_{k=\ell}^{\infty} t^k \frac{1}{\ell!(k-\ell)!} \lambda^{k-\ell} = \frac{t^\ell}{\ell!} \sum_{k=\ell}^{\infty} \frac{1}{(k-\ell)!} (\lambda t)^{k-\ell} \\ &= \frac{t^\ell}{\ell!} \sum_{j=0}^{\infty} \frac{1}{j!} (\lambda t)^j = \frac{t^\ell}{\ell!} e^{\lambda t} ,\end{aligned}$$

we find that

$$e^{t\boldsymbol{J}_r} = \sum_{k=0}^{\infty} \frac{(t\boldsymbol{J}_r)^k}{k!} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \cdots & \cdots & \cdots & \frac{t^{m-1}}{(m-1)!}e^{\lambda t} \\ 0 & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \ddots & \ddots & \frac{t^{m-2}}{(m-2)!}e^{\lambda t} \\ \vdots & 0 & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} \\ \vdots & \vdots & \ddots & \ddots & 0 & e^{\lambda t} & te^{\lambda t} \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & e^{\lambda t} \end{bmatrix} .$$

**Example 2.30.** Let $\boldsymbol{J} = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix}$. Then $\boldsymbol{J}$ takes the form $\begin{bmatrix} \boldsymbol{J}_1 & & \\ & \boldsymbol{J}_2 & \\ & & \boldsymbol{J}_3 \end{bmatrix}$ so that

$$e^{t\boldsymbol{J}} = \begin{bmatrix} e^{2t} & te^{2t} & \frac{t^2}{2}e^{2t} & 0 & 0 & 0 \\ 0 & e^{2t} & te^{2t} & 0 & 0 & 0 \\ 0 & 0 & e^{2t} & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{-3t} & te^{-3t} & 0 \\ 0 & 0 & 0 & 0 & e^{-3t} & 0 \\ 0 & 0 & 0 & 0 & 0 & e^{5t} \end{bmatrix} .$$

**Example 2.31.** Consider the ODE derived from the two masses three springs system:

$$\begin{aligned} m_1 \frac{d^2 x_1}{dt^2} &= -k_1 x_1 + k_2(x_2 - x_1) + F_1 , \\ m_2 \frac{d^2 x_2}{dt^2} &= -k_2(x_2 - x_1) - k_3 x_2 + F_2 . \end{aligned}$$

Let $\boldsymbol{y} = [x_1, x_1', x_2, x_2']^{\mathrm{T}}$. Then

$$\boldsymbol{y}'(t) = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{f} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k_1+k_2}{m_1} & \frac{k_2}{m_1} & 0 & 0 \\ \frac{k_2}{m_2} & -\frac{k_2+k_3}{m_2} & 0 & 0 \end{bmatrix} \boldsymbol{y} + \begin{bmatrix} 0 \\ 0 \\ \frac{F_1}{m_1} \\ \frac{F_2}{m_2} \end{bmatrix} . \tag{2.48}$$

Suppose that $m_1 = m_2 = k_1 = k_2 = k_3 = 1$. If $\lambda$ is an eigenvalue of $\boldsymbol{A}$, then

$$0 = \begin{vmatrix} -\lambda & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 1 \\ -2 & 1 & -\lambda & 0 \\ 1 & -2 & 0 & -\lambda \end{vmatrix} = -\lambda \begin{vmatrix} -\lambda & 0 & 1 \\ 1 & -\lambda & 1 \\ -2 & 0 & -\lambda \end{vmatrix} + \begin{vmatrix} 0 & -\lambda & 1 \\ -2 & 1 & 0 \\ 1 & -2 & -\lambda \end{vmatrix}$$
$$= -\lambda(-\lambda^3 - 2\lambda) + (4 - 1 + 2\lambda^2) = \lambda^4 + 4\lambda^2 + 3 = (\lambda^2 + 3)(\lambda^2 + 1)$$

which implies that the eigenvalues of $\boldsymbol{A}$ are $\pm\sqrt{3}i$ and $\pm i$. Corresponding eigenvectors are

$$\pm\sqrt{3}i \leftrightarrow \Big[\pm \frac{1}{\sqrt{3}}i, \mp\frac{1}{\sqrt{3}}i, -1, 1\Big]^{\mathrm{T}}, \qquad \pm i \leftrightarrow \big[\mp i, \mp i, 1, 1\big]^{\mathrm{T}};$$

thus

$$\boldsymbol{A} = \begin{bmatrix} \frac{1}{\sqrt{3}}i & -\frac{1}{\sqrt{3}}i & -i & i \\ -\frac{1}{\sqrt{3}}i & \frac{1}{\sqrt{3}}i & -i & i \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3}i & & & \\ & -\sqrt{3}i & & \\ & & i & \\ & & & -i \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}}i & -\frac{1}{\sqrt{3}}i & -i & i \\ -\frac{1}{\sqrt{3}}i & \frac{1}{\sqrt{3}}i & -i & i \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}^{-1}.$$

Therefore,

$$e^{t\boldsymbol{A}} = \begin{bmatrix} \frac{1}{\sqrt{3}}i & -\frac{1}{\sqrt{3}}i & -i & i \\ -\frac{1}{\sqrt{3}}i & \frac{1}{\sqrt{3}}i & -i & i \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} e^{i\sqrt{3}t} & & & \\ & e^{-i\sqrt{3}t} & & \\ & & e^{it} & \\ & & & e^{-it} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}}i & -\frac{1}{\sqrt{3}}i & -i & i \\ -\frac{1}{\sqrt{3}}i & \frac{1}{\sqrt{3}}i & -i & i \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}^{-1}.$$

Using (2.45), we find that the general solution to (2.48) is given by

$$\boldsymbol{y}(t) = \boldsymbol{P} \begin{bmatrix} e^{i\sqrt{3}t} & & & \\ & e^{-i\sqrt{3}t} & & \\ & & e^{it} & \\ & & & e^{-it} \end{bmatrix} \boldsymbol{P}^{-1} \int \boldsymbol{P} \begin{bmatrix} e^{-i\sqrt{3}t} & & & \\ & e^{i\sqrt{3}t} & & \\ & & e^{-it} & \\ & & & e^{it} \end{bmatrix} \boldsymbol{P}^{-1} \boldsymbol{f}(t)\, dt$$
$$= \boldsymbol{P} \begin{bmatrix} e^{i\sqrt{3}t} & & & \\ & e^{-i\sqrt{3}t} & & \\ & & e^{it} & \\ & & & e^{-it} \end{bmatrix} \int \begin{bmatrix} e^{-i\sqrt{3}t} & & & \\ & e^{i\sqrt{3}t} & & \\ & & e^{-it} & \\ & & & e^{it} \end{bmatrix} \boldsymbol{P}^{-1} \boldsymbol{f}(t)\, dt\,,$$

where $\boldsymbol{P} = \begin{bmatrix} \frac{1}{\sqrt{3}}i & -\frac{1}{\sqrt{3}}i & -i & i \\ -\frac{1}{\sqrt{3}}i & \frac{1}{\sqrt{3}}i & -i & i \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{f}(t) = \begin{bmatrix} 0 \\ 0 \\ \frac{F_1}{m_1} \\ \frac{F_2}{m_2} \end{bmatrix}$.

**Example 2.32.** Consider the linear system $\boldsymbol{X}' = \boldsymbol{A}\boldsymbol{X}$, where

$$\boldsymbol{A} = \begin{bmatrix} 4 & -2 & 0 & 2 \\ 0 & 6 & -2 & 0 \\ 0 & 2 & 2 & 0 \\ 0 & -2 & 0 & 6 \end{bmatrix} = \begin{bmatrix} -2 & 0 & 0 & 1 \\ -2 & 1 & 1 & 0 \\ -2 & 2 & 1 & 0 \\ -2 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} -2 & 0 & 0 & 1 \\ -2 & 1 & 1 & 0 \\ -2 & 2 & 1 & 0 \\ -2 & 0 & 1 & 1 \end{bmatrix}^{-1}.$$

Using formula (2.45), the general solution to the given ODE is

$$\begin{aligned}\boldsymbol{X}(t) &= \begin{bmatrix}-2 & 0 & 0 & 1\\ -2 & 1 & 1 & 0\\ -2 & 2 & 1 & 0\\ -2 & 0 & 1 & 1\end{bmatrix} \exp\left(t\begin{bmatrix}4 & 1 & 0 & 0\\ 0 & 4 & 0 & 0\\ 0 & 0 & 4 & 0\\ 0 & 0 & 0 & 6\end{bmatrix}\right) \begin{bmatrix}-2 & 0 & 0 & 1\\ -2 & 1 & 1 & 0\\ -2 & 2 & 1 & 0\\ -2 & 0 & 1 & 1\end{bmatrix}^{-1} \begin{bmatrix}C_1\\ C_2\\ C_3\\ C_4\end{bmatrix}\\ &= \begin{bmatrix}-2 & 0 & 0 & 1\\ -2 & 1 & 1 & 0\\ -2 & 2 & 1 & 0\\ -2 & 0 & 1 & 1\end{bmatrix} \begin{bmatrix}e^{4t} & te^{4t} & 0 & 0\\ 0 & e^{4t} & 0 & 0\\ 0 & 0 & e^{4t} & 0\\ 0 & 0 & 0 & e^{6t}\end{bmatrix} \begin{bmatrix}\bar{C}_1\\ \bar{C}_2\\ \bar{C}_3\\ \bar{C}_4\end{bmatrix}\end{aligned}$$

for some constants $\bar{C}_1$, $\bar{C}_2$, $\bar{C}_3$ and $\bar{C}_4$.

## 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 (2.3)) with initial condition $\boldsymbol{y}(0) = \boldsymbol{y}_0$. Note that usually you need to write the IVP in a dimensionless form (under a proper choice of characteristic scale).

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

$$\frac{dp}{dt} = -0.16p + 0.08pq\,, \tag{2.49a}$$

$$\frac{dq}{dt} = 4.5q - 0.9pq\,, \tag{2.49b}$$

$$p(0) = 5, \quad q(0) = 3\,. \tag{2.49c}$$

numerically using matlab.

Let $\boldsymbol{y} = [p, q]^{\mathrm{T}}$, and $\boldsymbol{f}(t, \boldsymbol{y}) = \begin{bmatrix} -0.16p + 0.08pq \\ 4.5q - 0.9pq \end{bmatrix}$. 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);
```

(2.50)

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 recommend the reader to have (2.50) 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 (2.49). In the main window of matlab, type

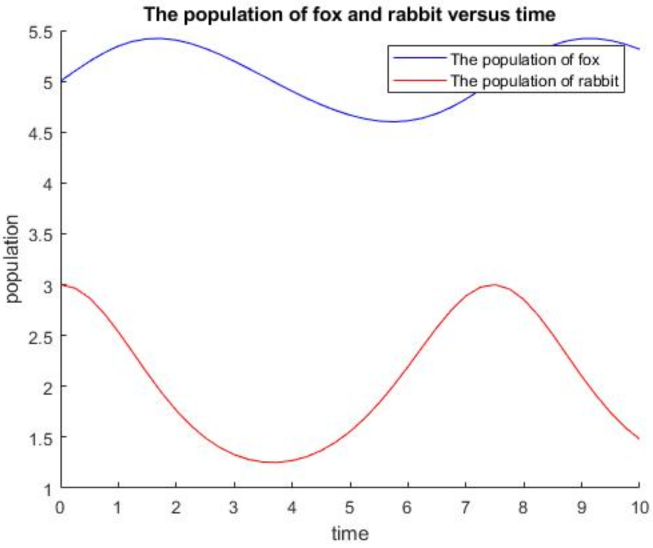
```
[t,y] = ode45(@ODE_RHS,[0,10],[5;3]);
```

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')
```

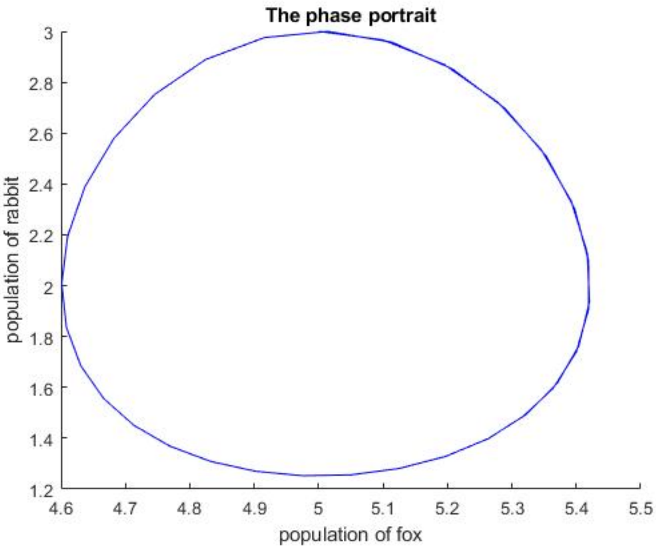
which outputs



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 2.34.** In this example we solve for the IVP (from the study of Kepler's laws of planetary motion)

$$-\frac{GMm}{r^2}\widehat{r} = m\boldsymbol{r}'' , \qquad \boldsymbol{r}(0) = \boldsymbol{r}_0, \quad \boldsymbol{r}'(0) = \boldsymbol{r}_1 ,$$

under the settings: $GM = 1$, $\boldsymbol{r}_0 = [1; 0]$ and $\boldsymbol{r}_1 = [0; 0.6]$. We note that the IVP above can be written as

$$-\frac{GM}{(x^2+y^2)^{1.5}}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x'' \\ y'' \end{bmatrix}, \qquad \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \boldsymbol{r}_0\,, \quad \begin{bmatrix} x'(0) \\ y'(0) \end{bmatrix} = \boldsymbol{r}_1\,.$$

In order to make use of the command "ode45", one needs to rewrite the equation into the first order form. Let $\boldsymbol{z} = [z_1; z_2; z_3; z_4] \equiv [x; y; x'; y']$. Then $\boldsymbol{z}$ satisfies

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} z_3 \\ z_4 \\ -\dfrac{z_1}{(z_1^2+z_2^2)^{1.5}} \\ -\dfrac{z_2}{(z_1^2+z_2^2)^{1.5}} \end{bmatrix}, \qquad \boldsymbol{z}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0.6 \end{bmatrix}.$$

Therefore, we execute the following codes (explained in the next page)

```
ODE_RHS = @(t,y) [y(3:4); -1/(norm(y(1:2))∧3)*y(1:2)];
[t,y] = ode45(@(t,y) ODE_RHS(t,y), [0,3], [1;0;0;0.6]);
plot(y(:,1),y(:,2),'b');
axis equal;
```

to produce the trajectory of the planet in the following figure.

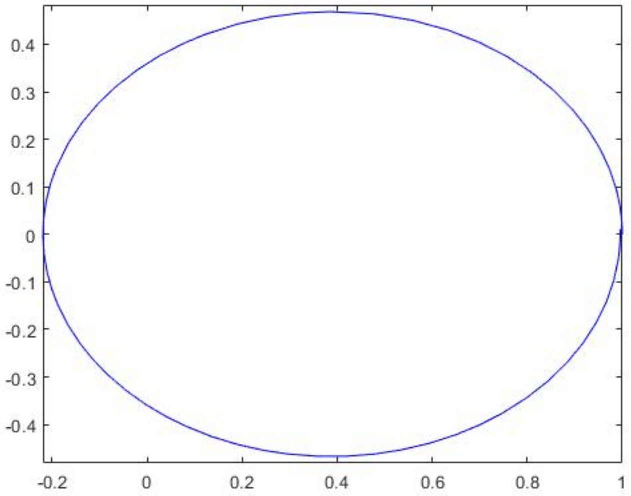


Figure 2.9: The trajectory of a planet is an ellipse in this example

**Explanation of the codes**:

1. If the right-hand side function is simple, sometimes we do not write a separate file for the function and can put it directly into the main file. The format is

   ```
   name of the function = @(list of variables) the output based on given variables
   ```

   This function can only be seen and used in this main file.

2. When using local function, in order to use “ode45” one needs to use

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

instead of

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

In fact, the new way of using “ode45” will be better because it allows that we have other parameters in the forcing function. For example, the original code can be modified as

```
ODE_RHS = @(t,y,G,M) [y(3:4); -G*M/(norm(y(1:2))∧3)*y(1:2)];
G = 1; M = 1;
[t,y] = ode45(@(t,y) ODE_RHS(t,y,G,M), [0,3], [1;0;0;0.6]);
plot(y(:,1),y(:,2),'b');
axis equal;
```

so that we can easily change the value of G and M. Here we emphasize that in order to use “ode45”, even if the right-hand side function has several input variables, you still have to use “@(t,y)” like the red part in the third line.

3. There is an even easier way of making use of “ode45” when the right-hand side function is simple.

```
G = 1; M = 1;
[t,y] = ode45(@(t,y) [y(3:4); -G*M/(norm(y(1:2))∧3)*y(1:2)], [0,3], [1;0;0;0.6]);
plot(y(:,1),y(:,2),'b');
axis equal;
```

**Example 2.35.** In this example we look for the position where the function $f(x, y) = xe^{-x^2-y^2}$ attains its global minimum or one of its local minimums. 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

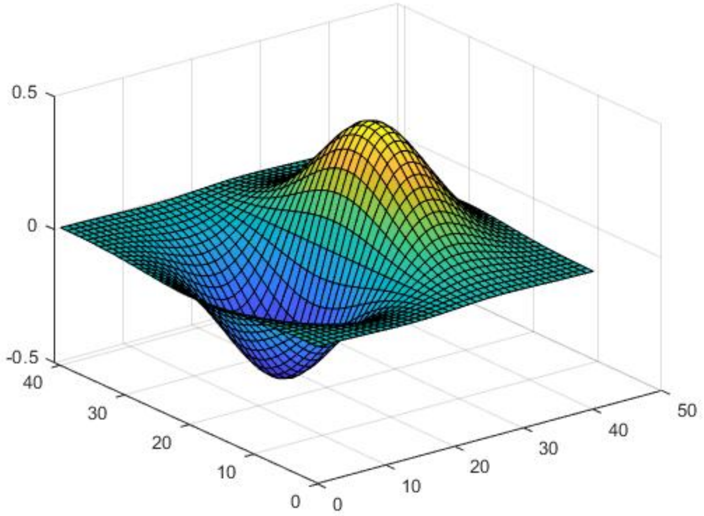


Figure 2.10: The graph of the function $f(x, y) = xe^{-x^2-y^2}$.

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 $\big(\pm\frac{1}{\sqrt{2}}, 0\big)$. 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

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = -(\nabla f)(x, y)\,, \tag{2.51a}$$

$$\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \tag{2.51b}$$

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 (2.51a) shows that the tangent direction of the integral curve should agree with the direction field.

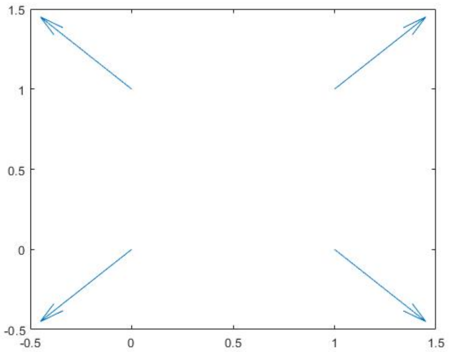
Let us visualize this by plotting first the vector field $-(\nabla f)$. To plots a vector $\boldsymbol{u} = (\text{x component}, \text{y component})$ at the point $p = (\text{x coordinate}, \text{y coordinate})$, we use the command "quiver" in the following way:

```
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));
```

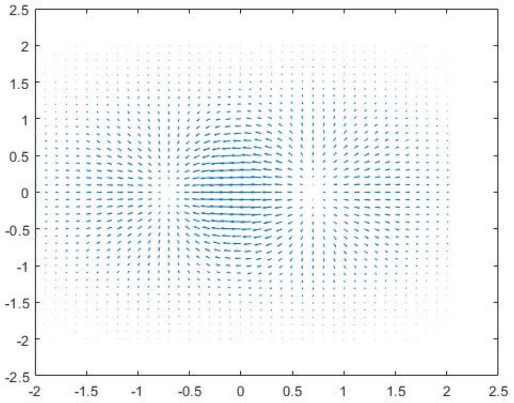
and the following figure will be produced:



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))
```

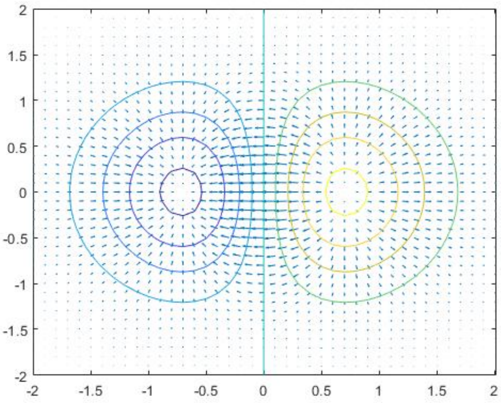
to produce the following figure of the vector field:



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

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

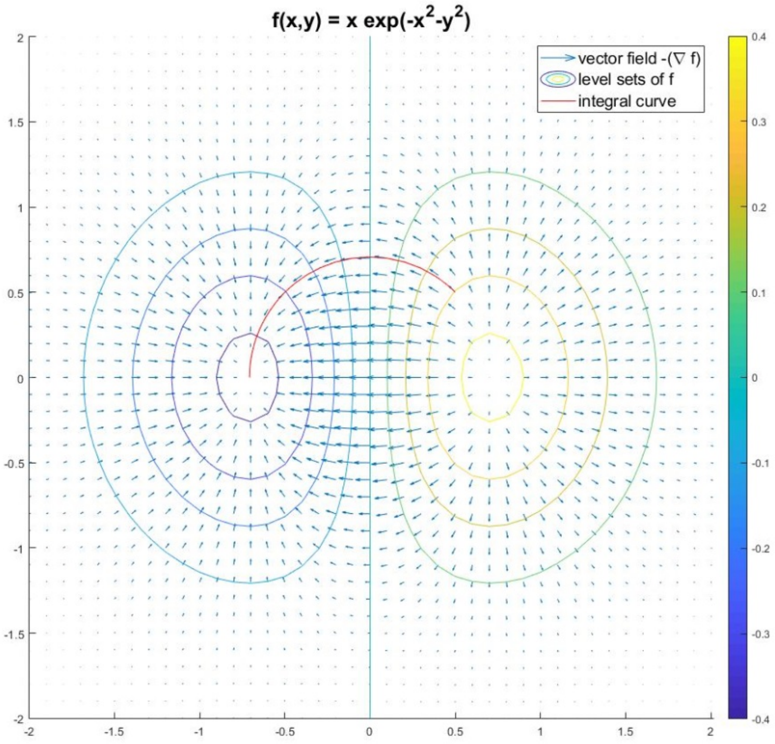
so that we obtain



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')
colorbar
```

## 2.4 Boundary Value Problems

In this section we only consider ODE of the form

$$y'' + p(x)y' + q(x)y = g(x)\,, \tag{2.52}$$

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 $\alpha$, $\beta$, $y_0$ 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 2.36.** 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 2.37.** Again we consider the ODE

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

in Example 1.14. 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

$$y'' + p(x)y' + q(x)y = g(x)\,, \qquad y(\alpha) = y_0, \quad y(\beta) = y_1\,. \tag{2.53}$$

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

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

where $G(x) = g(x) - p(x)\frac{y_0 - y_1}{\alpha - \beta} - q(x)\big(\frac{x-\alpha}{\beta-\alpha}y_1 + \frac{x-\beta}{\alpha-\beta}y_0\big)$. Therefore, in general we can assume the homogeneous boundary condition $y_0 = y_1 = 0$ in (2.53). Similarly, ODE (2.52) with the other three kinds of boundary conditions can also be rewritten as a BVP with homogeneous boundary condition.

**Remark 2.38.** Even though the initial value problem

$$y'' + p(t)y' + q(t)y = g(t)\,, \qquad y(t_0) = y_0\,, \quad y'(t_0) = y_1 \tag{2.54}$$

looks quite similar to the boundary value problem (2.53), they actually differ in some very important ways. For example, if $p, q, g$ are continuous, the initial value problem (2.54) always have a unique solution, while the boundary value problem (2.53) 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 (2.53) 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 $y(x) = \cos\sqrt{2}x - \cot\sqrt{2}\pi \sin\sqrt{2}x$.

Similar to Theorem 2.16, we have the following

**Theorem 2.39.** *Let $\alpha, \beta$ be real numbers and $\alpha < \beta$. Suppose that the function $f = f(t, y, p)$ is continuous on the set*

$$D = \big\{(x, y, p) \,\big|\, x \in [\alpha, \beta], y, p \in \mathbb{R}\big\}$$

*and the partial derivatives $f_y$ and $f_p$ are also continuous on $D$. If*

1. *$f_y(t, y, p) > 0$ for all $(t, y, p) \in D$, and*

2. *there exists a constant $M > 0$ such that*

$$\big|f_p(t, y, p)\big| \leqslant M \quad \forall\,(t, y, p) \in D\,,$$

*then the boundary value problem*

$$y'' = f(t, y, y') \qquad \forall\, x \in (\alpha, \beta),\; y(\alpha) = y(\beta) = 0$$

*has a unique solution.*

# 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 of the resulting equation through by $\Delta x$ 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 $\Delta t$ 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.

### 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 flowing into the section from outside. If there is a heat source $Q$, then the equation above has to be modified as

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

Dividing both sides by $\Delta t$ 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

$$\begin{aligned}\int_x^{x+\Delta x} \rho(y)s(y)\vartheta_t(y, t)\, dy &= \big[\kappa(x + \Delta x)\vartheta_x(x + \Delta x, t) - \kappa(x)\vartheta_x(x, t)\big] + \int_x^{x+\Delta x} Q(y, t)\, dy \\ &= \big[\kappa(y)\vartheta_x(y, t)\big]\Big|_{y=x}^{y=x+\Delta x} + \int_x^{x+\Delta x} Q(y, t)\, dy \qquad (3.1)\end{aligned}$$

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

$$\rho(x)s(x)\frac{\partial}{\partial t}\vartheta(x, t) = \frac{\partial}{\partial x}\big[\kappa(x)\vartheta_x(x, t)\big] + Q(x, t) \qquad 0 < x < L\,, \quad t > 0\,. \qquad (3.2)$$

Assuming uniform rod; that is, $\rho, s, \kappa$ are constant, then (3.2) reduces to that

$$\vartheta_t(x, t) = \alpha^2\vartheta_{xx}(x, t) + q(x, t)\,, \qquad 0 < x < L\,, \quad t > 0\,, \qquad (3.3a)$$

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

$$\vartheta(x, 0) = \vartheta_0(x) \qquad 0 < x < L \qquad (3.3b)$$

and a boundary condition.

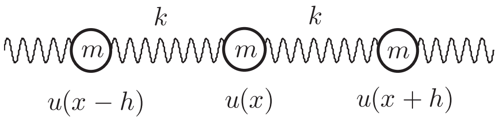
(a) Temperature on the end-points of the rod is fixed: $\vartheta(0, t) = T_1$ and $\vartheta(L, t) = T_2$.

(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 $u(x,t)$ measures the distance from the equilibrium of the mass situated at position $x$ and time $t$, then the forces exerted on the mass $m$ at the location $x$ are

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

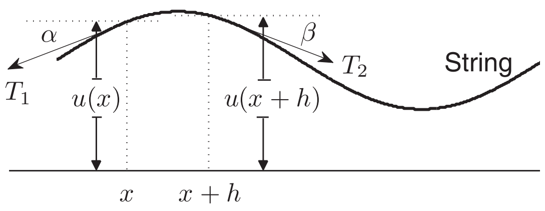
If the array of weights consists of $N$ weights spaced evenly over the length $L=(N+1)h$ of total mass $M=Nm$, and the total stiffness of the array $K=k/(N+1)$, then

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

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

$$u_{tt}(x,t) = c^2 u_{xx}(x,t)\,. \tag{3.4}$$

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

$$T_1\cos\alpha = T_2\cos\beta\,. \tag{3.5}$$

Noting that

$$\begin{aligned} \cos\alpha &= \frac{1}{\sec\alpha} = \frac{1}{\sqrt{1+\tan^2\alpha}} = \frac{1}{\sqrt{1+\tan^2(\pi+\alpha)}} = \frac{1}{\sqrt{1+u_x(x,t)^2}}\,, \\ \cos\beta &= \frac{1}{\sec\beta} = \frac{1}{\sqrt{1+\tan^2\beta}} = \frac{1}{\sqrt{1+\tan^2(2\pi-\beta)}} = \frac{1}{\sqrt{1+u_x(x+h,t)^2}}\,, \end{aligned}$$

identity (3.5) implies that the function $\dfrac{T(x,t)}{\sqrt{1+u_x(x,t)^2}}$ is constant in $x$ (but not necessary constant in $t$). Denote this constant as $\tau(t)$. Then by the fact that the vertical component of $T_1$ and $T_2$ induce the vertical motion, we obtain that

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

where $\mu$ denotes the density of the string, and the integral on the left-hand side is the total force due to the acceleration. Dividing both sides through by $h$ and passing to the limit as $h \to 0$, we find that

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

If there is an external forcing $f$ acting on the string, then (3.6) becomes

$$\mu(x)u_{tt}(x,t) = \tau(t)u_{xx}(x,t) + f(x,t)\,. \tag{3.7}$$

If $\mu$ is constant in $x$ and $\tau$ is constant in $t$ (which is a reasonable assumption if the vibration of the string is very small and uniform), then (3.7) reduces to

$$u_{tt}(x,t) = c^2 u_{xx}(x,t) + \frac{1}{\mu}f(x,t)\,. \tag{3.8}$$

To determined the state of the vibration, initial conditions and boundary conditions have to be imposed.

- **Initial conditions**: Since the second derivatives in $t$ is involved, similar to the case of ODE we need two initial conditions

$$u(x,0) = \varphi(x)\,, \qquad u_t(x,0) = \psi(x)\,,$$

where $\varphi$ and $\psi$ are given functions.

- **Boundary conditions**: Similar to the heat equation, one of the following three boundary conditions is imposed.

  (a) Vibration string with fixed ends: $u(0,t) = u(L,t) = 0$. The same as the case in the heat equation, this kind of boundary condition is also called **Dirichlet** boundary conditions.

  (b) Vibration string with free ends: $u_x(0,t) = u_x(L,t) = 0$. This kind of boundary condition is also called **Neumann** boundary conditions.

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

## 3.2 Solving PDE using matlab® - Part I

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

$$u_t = A(u) + f \qquad \text{or} \qquad u_{tt} = A(u) + f \tag{3.9}$$

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 (3.9) 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 $\xi$ 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 $\|\mathcal{P}\| = h = \dfrac{L}{n}$ 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

$$w'(x) \approx \frac{w(x+h) - w(x-h)}{2h} \quad \text{and} \quad w''(x) \approx \frac{w(x+h) - 2w(x) + w(x-h)}{h^2} \tag{3.10}$$

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

**Remark 3.2.** 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 3.3.** Let $\Delta_h$ 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).

### 3.2.1 The 1-dimensional heat equations

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

$$\vartheta_t - \kappa \vartheta_{xx} = f(x,t) \qquad \text{in} \quad (0,L) \times \mathbb{R}^+\,, \tag{3.11a}$$

$$\vartheta = \vartheta_0 \qquad \text{on} \quad (0,L) \times \{0\}\,, \tag{3.11b}$$

$$\vartheta(0,t) = a(t)\,,\ \vartheta(L,t) = b(t) \qquad \text{on} \quad \{0,L\} \times \mathbb{R}^+\,. \tag{3.11c}$$

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 (3.11) implies that

$$\begin{aligned} \frac{d\varphi_i}{dt} - \frac{\kappa}{h^2}(\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}) = f_i(t) + \mathcal{O}(h^2) \qquad & \text{for all } 1 \leqslant i \leqslant n \text{ and } t > 0\,, \\ \varphi_i(0) = \vartheta_0(x_i) \qquad & \text{for all } 1 \leqslant i \leqslant n\,, \\ \varphi_0(t) = a(t)\,,\ \varphi_{n+1}(t) = b(t) \qquad & \text{for all } t > 0\,, \end{aligned}$$

where $\vartheta_0$ 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 3.4.** 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 $\kappa$, $h$, 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 $\varphi$ 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;
```

### 3.2.2 The 1-dimensional wave equations

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

$$u_{tt} - c^2 u_{xx} = f(x,t) \qquad \text{in} \quad (0,L) \times \mathbb{R}^+ \,, \tag{3.12a}$$

$$u = u_0 \,, u_t = u_1 \qquad \text{on} \quad (0,L) \times \{0\} \,, \tag{3.12b}$$

$$u_x(0,t) = a(t) \,,\ u_x(L,t) = b(t) \qquad \text{on} \quad \{0,L\} \times \mathbb{R}^+ \,. \tag{3.12c}$$

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 (3.12a) and the central difference scheme (3.10) imply that

$$\frac{d^2 v_i}{dt^2} - c^2 \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} = f_i(t) + \mathcal{O}(h^2) \qquad \text{for all } 2 \leqslant i \leqslant n-1 \text{ and } t > 0\,. \tag{3.13}$$

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 (3.10), (3.12c) 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 (3.10), 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 3.2.1, naively we consider

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

with initial conditions

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

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

$$\frac{d}{dt}\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w} \\ \frac{c^2}{h^2} K \boldsymbol{v} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{f}(t) \end{bmatrix} = \begin{bmatrix} \mathrm{I}_n & 0 \\ 0 & \frac{c^2}{h^2} K \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{f}(t) \end{bmatrix}, \tag{3.14}$$

where $\mathrm{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{f} = (f_1, \cdots, f_n)^{\mathrm{T}}$.

Once (3.14) 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 $\{x_1, \cdots, x_n\}$.

### 3.2.3 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

$$u_t + u_x = q(x,t) \quad \text{in} \quad (0,L) \times (0,T)\,, \tag{3.15a}$$

$$u(x,0) = u_0(x) \quad \text{on} \quad (0,L) \times \{t = 0\}\,, \tag{3.15b}$$

$$u(0,t) = u(L,t) = 0 \quad \text{for all } t > 0\,. \tag{3.15c}$$

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 (3.15) 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 (3.10) to approximate $u_x(x_i,t)$, we find that

$$\frac{du_i}{dt} + \frac{u_{i+1}(t) - u_{i-1}(t)}{2h} = q(x_i,t) + \mathcal{O}(h^2) \qquad \text{for all } 1 \leqslant i \leqslant n \text{ and } t > 0$$

where $u_0(t) = u_{n+1}(t) = 0$. The ODE above motivates the following ODE

$$\frac{d}{dt}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \vdots \\ v_{n-2} \\ v_{n-1} \\ v_n \end{bmatrix} = \frac{1}{2h}\begin{bmatrix} 0 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & -1 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\ \vdots & & 0 & 1 & 0 & -1 & 0 \\ \vdots & & & 0 & 1 & 0 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \vdots \\ v_{n-2} \\ v_{n-1} \\ v_n \end{bmatrix} + \begin{bmatrix} q_1(t) \\ q_2(t) \\ q_3(t) \\ \vdots \\ \vdots \\ q_{n-2}(t) \\ q_{n-1}(t) \\ q_n(t) \end{bmatrix}$$

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 3.5.** 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 (3.10) 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

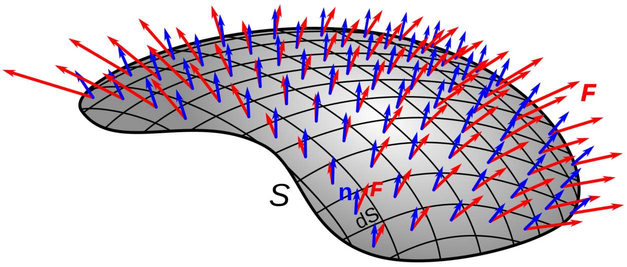
```
function output = source_q(x,t)
output = abs(x-5)-5;
```

and you will find that the numerical solution becomes a garbage immediately.

## 3.3 Models with Several Spatial Variables

### 3.3.1 Equation of continuity

Let $u$ be the density of concentration of some physical quantity ($u = u(x,t)$) in a domain $\Omega \subseteq \mathbb{R}^n$, where $n = 2$ or $n = 3$, and let $\boldsymbol{F}$ be the flux of the quantity; that is, $\boldsymbol{F} \cdot \boldsymbol{n}\, dS$ is the **flow rate** of the quantity that passes through an area $dS$ in the normal direction $\boldsymbol{n}$ of $dS$:



Then for a given open domain $\mathcal{O} \subset\subset \Omega$ so that $\partial\mathcal{O}$ is (piecewise) smooth,

1. the change of the total amount of the quantity in $\mathcal{O}$ from time $t$ to $t + \Delta t$ is
$$\int_{\mathcal{O}} \big[u(x, t + \Delta t) - u(x,t)\big]\, dx\,.$$

2. the flow rate of the quantity flowing out of $\mathcal{O}$ through $\partial\mathcal{O}$ at time $t$ is $\displaystyle\int_{\partial\mathcal{O}} (\boldsymbol{F}\cdot\boldsymbol{n})(t)\, dS$; thus the total amount of the quantity flows out of $\mathcal{O}$ through $\partial\mathcal{O}$ from time $t$ to $t+\Delta t$ is given by
$$\int_t^{t+\Delta t} \int_{\partial\mathcal{O}} (\boldsymbol{F}\cdot\boldsymbol{n})(x,t')\, dSdt'\,,$$
where $\boldsymbol{n}$ is the (almost everywhere defined) outward-pointing unit normal of $\partial O$.

3. if there is a source of the quantity, the total amount of the quantity in $\mathcal{O}$ produced by the source from time $t$ to $t + \Delta t$ is
$$\int_t^{t+\Delta t} \int_{\mathcal{O}} q(x,t') dxdt'\,,$$
where $q$ is the strength of sources for the quantity.

Therefore, the balance of the amount of the quantity in $\mathcal{O}$ implies that

$$\int_{\mathcal{O}} \big[u(x,t+\Delta t) - u(x,t)\big]\, dx = -\int_t^{t+\Delta t} \int_{\partial\mathcal{O}} (\boldsymbol{F}\cdot\boldsymbol{n})(x,t')\, dSdt' + \int_t^{t+\Delta t} \int_{\mathcal{O}} q(x,t') dxdt'$$

for all "good" subset $\mathcal{O} \subseteq \Omega$, here a "good" set refers to a set with piecewise smooth boundary. Dividing both sides of the equation above by $\Delta t$ and passing to the limit as $\Delta t \to 0$, we obtain that

$$\frac{d}{dt}\int_{\mathcal{O}} u\, dx = -\int_{\partial\mathcal{O}} \boldsymbol{F}\cdot\boldsymbol{n}\, dS + \int_{\mathcal{O}} q\, dx \qquad \text{for all "good" open subset } \mathcal{O} \subseteq \Omega\,. \tag{3.16}$$

If $u$ is smooth, by the divergence theorem we find that

$$\int_{\mathcal{O}} u_t\,dx = \int_{\mathcal{O}} (q - \mathrm{div}\boldsymbol{F})\,dx \qquad \text{for all "good" open subset } \mathcal{O} \subseteq \Omega\,,$$

or equivalently,

$$\int_{\mathcal{O}} \big[u_t + \mathrm{div}\boldsymbol{F} - q\big]\,dx = 0 \qquad \text{for all "good" open subset } \mathcal{O} \subseteq \Omega\,.$$

Since $\mathcal{O}$ is given arbitrarily in $\Omega$, we conclude that

$$u_t + \mathrm{div}\boldsymbol{F} = q \qquad \text{in} \quad \Omega \times (0, T)\,. \tag{3.17}$$

Equation (3.17) is called *the equation of continuity*.

- **The conservation of mass in fluid dynamics**

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

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

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

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

A vector field $\boldsymbol{u}$ satisfying $\mathrm{div}\boldsymbol{u} = 0$ everywhere inside the domain is said to be *solenoidal* or *divergence-free*.

### 3.3.2 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 $s(x)$, $\varrho(x)$, $\kappa(x)$ be the specific heat, density, and the inner thermal conductivity of the material body at $x$, and $Q(x,t)$ is the strength of the source of the heat energy. Then by the conservation of heat energy, similar to the derivation of Equation (3.1) and Equation (3.16) (with the heat flux $\boldsymbol{F} = -k\nabla\vartheta$ in mind) we obtain that for any "good" open set $\mathcal{O} \subset\subset \Omega$,

$$\frac{d}{dt}\int_{\mathcal{O}} s(x)\varrho(x)\vartheta(x,t)\,dx = \int_{\partial\mathcal{O}} \kappa(x)\nabla\vartheta(x,t)\cdot\boldsymbol{n}(x)\,dS + \int_{\mathcal{O}} Q(x,t)\,dx\,, \tag{3.20}$$

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

$$\int_{\mathcal{O}} s(x)\varrho(x)\vartheta_t(x,t)\,dx = \int_{\mathcal{O}} \mathrm{div}\big(\kappa(x)\nabla\vartheta(x,t)\big)\,dx + \int_{\mathcal{O}} Q(x,t)\,dx\,.$$

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

$$s(x)\varrho(x)\vartheta_t(x,t) - \mathrm{div}(\kappa(x)\nabla\vartheta(x,t)) = Q(x,t) \quad \forall\, x \in \Omega\,, t \in (0,T].$$

If the material body is uniform (that is, $s$, $\varrho$ and $\kappa$ are constants), then

$$\vartheta_t = \alpha^2 \Delta\vartheta + q(x,t) \qquad \forall\, x \in \Omega\,, t \in (0,T]\,, \tag{3.21}$$

where $\alpha^2 = \dfrac{\kappa}{s\varrho}$, $q = \dfrac{1}{s\varrho}Q$ and $\Delta$ is the Laplace operator (and $\Delta\vartheta$ reads laplacian theta) defined by

$$\Delta\vartheta \equiv \text{div}(\nabla\vartheta) = \sum_{i=1}^{n} \frac{\partial^2 \vartheta}{\partial x_i^2}\,.$$

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.

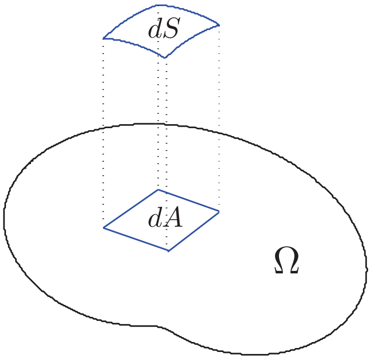
   (a) *Dirichlet condition*: $\vartheta(x,t) = g(x,t)$ for all $x \in \partial\Omega$ and $t \geqslant 0$, where $g$ is a given function.

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

### 3.3.3 The wave equations

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



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

**Idea**: The membrane stores certain energy $E(u)$ so that the deformation of the membrane changes the energy stored in the membrane which balances the work done by the external force $f$.

Suppose that an extra small external force $\vartriangle f = \vartriangle f(x_1, x_2)$ is suddenly added onto the membrane (so that the total force exerted on the membrane is $f + \vartriangle f$), and the membrane deforms to the surface $z = (u + \vartriangle u)(x_1, x_2)$ slowly (so the inertia does not have any effect). Then the extra energy needed to deform the membrane is $E(u+\vartriangle u) - E(u)$, while this extra work is done by the force $f + \vartriangle f$ given by

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

Therefore,

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

Even though we have assumed implicitly that $\vartriangle u$ is a function of $\vartriangle f$ (the deformation of the membrane is due to the change of external force), we can also assume that $\vartriangle f$ is a function of $\vartriangle u$ (so that we can modify $\vartriangle u$ independently). Let $\varphi$ be an "**admissible**" function (where "admissibility" means that $t\varphi$ can be used as $\vartriangle u$ for each $t \ll 1$) and $\vartriangle u = t\varphi$. Then if $t \neq 0$,

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

where, by the fact that $\vartriangle f \to 0$ as $\vartriangle u \to 0$, we have $\vartriangle f \to 0$ as $t \to 0$. Passing to the limit as $t \to 0$, we find that

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

Equation (3.22) always holds when considering time independent problems.

Suppose that the energy stored in the membrane is given by

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

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. Assuming that $u$ is a smooth function, then

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

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

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

thus (3.22) implies that

$$\int_\Omega \Big[\mathrm{div}\Big(\frac{\mathrm{T}\nabla u}{\sqrt{1 + |\nabla u|^2}}\Big) + f\Big]\varphi\, dA - \int_{\partial\Omega} \frac{\mathrm{T}}{\sqrt{1 + |\nabla u|^2}}\frac{\partial u}{\partial \mathbf{N}}\varphi\, ds = 0 \quad \text{for all admissible } \varphi\,. \tag{3.23}$$

In particular,

$$\int_\Omega \Big[\mathrm{div}\Big(\frac{\mathrm{T}\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)+f\Big]\varphi\,dA=0 \quad \text{for all admissible } \varphi \text{ that vanishes on } \partial\Omega\,. \tag{3.24}$$

The above identity implies that

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

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

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

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

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

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

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

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

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

**Remark 3.6.** 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). In fact, if $u=0$ on $\partial\Omega$, then all possible displacement $\vartriangle u$ should also satisfy that $\vartriangle u=0$ on $\partial\Omega$; thus each admissible $\varphi$ also has to vanish on $\partial\Omega$ in the derivation of (3.23) (and this is what the term "admissible" refers to in this case). In other words, if the membrane is constrained, instead of (3.23) we should obtain (3.24) directly.

**Remark 3.7.** By expanding the derivatives, we find that

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

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

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

thus if $|\nabla u| \ll 1$, (3.26) can be approximated by

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

while (3.27) can be approximated by

$$\left\{\begin{aligned} -\text{div}(\mathrm{T}\nabla u) &= f & &\text{in} \quad \Omega\,, \\ \frac{\partial u}{\partial \mathbf{N}} &= 0 & &\text{on} \quad \partial\Omega\,. \end{aligned}\right. \tag{N}$$

### • Equation for vibrating membrane

Let T be the tension, $\varrho$ be the density, and $f$ be the density of the external force which may depend on $x$ and $t$. For the case of vibrating membranes, part of $f$ induces the acceleration of the membrane which implies that

$$-\text{div}\Big(\frac{\mathrm{T}\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = f - \varrho u_{tt} \qquad \text{in} \quad \Omega \times (0,T]$$

or under the assumption that $|\nabla u| \ll 1$, the PDE above is simplified as

$$-\text{div}(\mathrm{T}\nabla u) = f - \varrho u_{tt} \qquad \text{in} \quad \Omega \times (0,T]\,.$$

This is in fact the **d'Alembert's principle** which states that the displacement $u$ satisfies that

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

for all $\varphi$ compatible with the existing constraints (or say, for all admissible $\varphi$).

Once the time derivative is involved in the PDEs, to fully determine the dynamics we need to impose initial conditions. The same as the case of the 1-dimensional wave equations, we need two initial conditions:

$$u(x,0) = \varphi(x)\,, \quad u_t(x,0) = \psi(x) \qquad \forall\, x \in \Omega\,,$$

where $\varphi$ and $\psi$ are given functions. Therefore, if $|\nabla u| \ll 1$,

1. Membrane fastened on the boundary:

$$\left\{\begin{aligned} \varrho u_{tt} - \text{div}(\mathrm{T}\nabla u) &= f & &\text{in} \quad \Omega \times (0,T]\,, \\ u &= g & &\text{on} \quad \partial\Omega \times (0,T]\,, \\ u(x,0) = \varphi(x)\,,\ u_t(x,0) &= \psi(x) & &\text{for all } x \in \Omega\,. \end{aligned}\right.$$

2. Membrane with free boundary:

$$\left\{\begin{aligned} \varrho u_{tt} - \text{div}(\mathrm{T}\nabla u) &= f & &\text{in} \quad \Omega \times (0,T]\,, \\ \frac{\partial u}{\partial \mathbf{N}} &= 0 & &\text{on} \quad \partial\Omega \times (0,T]\,, \\ u(x,0) = \varphi(x)\,,\ u_t(x,0) &= \psi(x) & &\text{for all } x \in \Omega\,. \end{aligned}\right.$$

### 3.3.4 The Navier-Stokes equations

In this section we derive the governing equation for fluid velocity in a fluid system. Let $\Omega$ be the fluid domain in which the fluid flows, and $\varrho$ and $\boldsymbol{u} = (u^1, u^2, u^3)$ be the density and the velocity of the fluid, respectively. 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 the 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 for all $\mathcal{O} \subset\subset \Omega$ with (piecewise) smooth boundary,

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

where $\boldsymbol{n}$ is the outward-pointing unit normal of $\partial\mathcal{O}$ (so that the first integral on the right-hand side is due to the momentum flux), $\boldsymbol{f}$ is the external force (such as the gravity) on the fluid system (so that the third integral on the right-hand side is the source of momentum), and $\boldsymbol{\sigma}$ is the stress (應力) exerted by the fluid due to the friction and the fluid pressure.

In the case of incompressible fluids, the stress is given by

$$\boldsymbol{\sigma} = 2\mu\text{Def}\boldsymbol{u}\boldsymbol{n} - -p\boldsymbol{n}\,,$$

where $\mu$ is called the dynamical viscosity (which may be a function of $\boldsymbol{u}$), $p$ is the fluid pressure, and Def$\boldsymbol{u}$, called the rate of strain tensor, is the symmetric part of the gradient of $\boldsymbol{u}$ given by

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

In other words, if $\boldsymbol{n} = (n_1, n_2, n_3)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$, then

$$\sigma_i = \mu \sum_{j=1}^{3} \Big(\frac{\partial u^i}{\partial x_j} + \frac{\partial u^j}{\partial x_i}\Big) n_j - p n_i\,. \tag{3.29}$$

Assuming the smoothness of the dependent variables, the application of the divergence theorem on (3.28) implies that for each $1 \leqslant i \leqslant 3$,

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

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

$$(\varrho\boldsymbol{u})_t + \text{div}(\varrho\boldsymbol{u}\otimes\boldsymbol{u}) + \nabla p = \text{div}(\mu\text{Def}\boldsymbol{u}) + \boldsymbol{f} \qquad \text{in} \quad \Omega\times(0,\infty)\,, \tag{3.30}$$

where $\boldsymbol{u}\otimes\boldsymbol{u} = [u^i u^j]$ and for a matrix $a = [a_{ij}]$, $(\text{div}a)_i \equiv \sum\limits_{j=1}^{3} \dfrac{\partial a_{ij}}{\partial x_j}$.

- **Newtonian and non-Newtonain fluids**

  1. Newtonian fluids: the viscosity $\mu$ is a constant.

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

Consider the Newtonian case. If the fluids under consideration is incompressible, we let $\varrho = 1$ so that (3.19) and (3.30) together imply the Navier-Stokes equations

$$\boldsymbol{u}_t + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} + \nabla p = \mu\Delta\boldsymbol{u} + \boldsymbol{f} \qquad \text{in} \quad \Omega\times(0,T)\,, \tag{3.31a}$$

$$\mathrm{div}\boldsymbol{u} = 0 \qquad \text{in} \quad \Omega\times(0,T)\,, \tag{3.31b}$$

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

$$(\mathrm{div}\boldsymbol{u}\otimes\boldsymbol{u})_i = \sum_{j=1}^{3}\frac{p}{px_j}(u^i u^j) = \sum_{j=1}^{3}\frac{\partial u^i}{\partial x_j}u^j + \sum_{j=1}^{3}u^i\frac{\partial u^j}{\partial x_j} = \sum_{j=1}^{3}u^j\frac{\partial u^i}{\partial x_j}$$

and

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

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

1. **No-slip boundary condition**: $\boldsymbol{u} = \mathbf{0}$ 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.

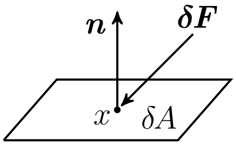
- **Some brief introduction about stress/traction**

- What is the stress/traction?

Let $\Sigma$ be a small piece of surface centered at $x$ with area $\delta A$ and $\boldsymbol{n}$ be a unit normal of $\Sigma$. The stress exerted by the fluid on the side toward which $\boldsymbol{n}$ points on the surface $\Sigma$ ($\boldsymbol{n}$ 方向所指的這一側的流體對曲面 $\Sigma$ 所施的應力) is defined as

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

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 $\boldsymbol{\sigma}(x, t, \mathbf{e}_j) = \tau_{1j}\mathbf{e}_1 + \tau_{2j}\mathbf{e}_2 + \tau_{3j}\mathbf{e}_3$ 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.

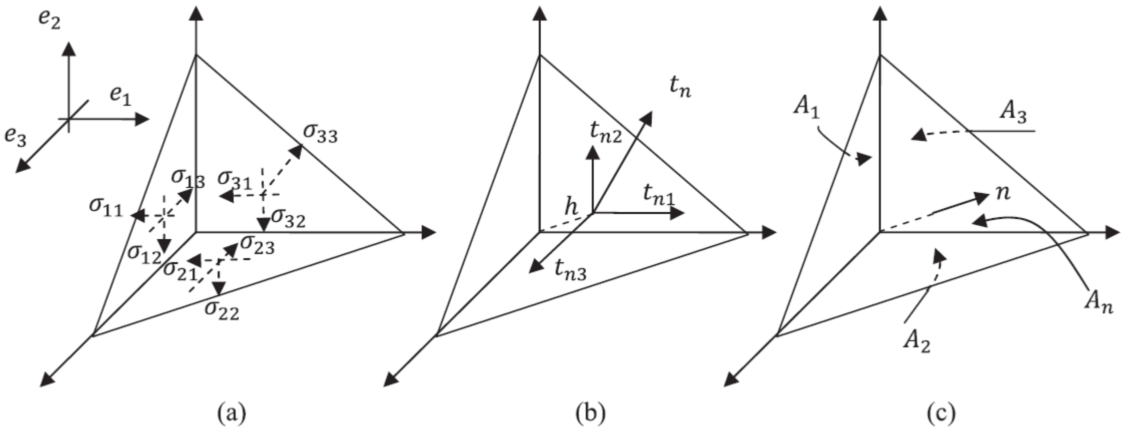


Figure 3.1: (a) On each side orthogonal to the coordinate axis, the stress is given by $\sigma(-\mathbf{e}_i) = \sum_{j=1}^{3} \sigma_{ij}\mathbf{e}_j$. (b) On the “slant” side, the stress is given by $\sigma(\boldsymbol{n}) = t_n = t_{n1}\mathbf{e}_1 + t_{n2}\mathbf{e}_2 + t_{n3}\mathbf{e}_3$. (c) By 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 $2\mu\mathrm{Def}\boldsymbol{u}\boldsymbol{n}$ appears in the expression of $\boldsymbol{\sigma}(\boldsymbol{n})$:

1. Suppose that $\Sigma$ 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 $\boldsymbol{\sigma}(\mathbf{n}) = \mu\dfrac{\partial u}{\partial x_3}\mathbf{e}_1$.
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 $\boldsymbol{\sigma}(\mathbf{n}) = \mu\dfrac{\partial(\boldsymbol{u} \cdot \widehat{\mathbf{e}}_1)}{\partial x_3}\widehat{\mathbf{e}}_1 = \mu\dfrac{\partial \boldsymbol{u}}{\partial x_3}$.
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 $\boldsymbol{\sigma}(\mathbf{n}) = \mu(\nabla\boldsymbol{u})\boldsymbol{n}$.
4. Since the stress tensor has to be symmetric, we have $\boldsymbol{\sigma}(\mathbf{n}) = 2\mu\mathrm{Def}\boldsymbol{u}\boldsymbol{n}$.

## 3.4 Solving PDE using matlab® - Part II

There is a built-in solver for PDE (with certain boundary conditions and probably initial condition) in matlab®. The PDE has to be of the form

$$m\frac{\partial^2 u}{\partial t^2} + d\,\frac{\partial u}{\partial t} - \operatorname{div}(c\nabla u) + au = f \qquad \text{for all } x \in \Omega\,, \tag{3.32}$$

where $\Omega$ is an open set in $\mathbb{R}^d$, $d = 2$ or 3, and either the Dirichlet, Neumann or mixed type boundary condition can be imposed. Moreover, the unknown $u$ can be a scalar or vector-valued function.

The main tool of solving PDE of form (3.32) in matlab® is the command "solvepde". The simplest code for PDE simulation in matlab® is for the PDE

$$\begin{aligned} -\Delta u &= 1 \qquad && \text{in} \quad \Omega\,, \\ u &= 0 \qquad && \text{on} \quad \partial\Omega\,, \end{aligned}$$

and is given as follows:

```
model = createpde();
geometryFromEdges(model,@lshapeg);
applyBoundaryCondition(model,’dirichlet’,’Edge’,1:model.Geometry.NumEdges,’u’,0);
specifyCoefficients(model,’m’,0,...
                          ’d’,0,...
                          ’c’,1,...
                          ’a’,0,...
                          ’f’,1);
generateMesh(model,’Hmax’,0.25);
results = solvepde(model);
```

here the domain $\Omega$ is an L-shape region (which has a built-in function named "lshapeg" for describing it).

- **Explanation of each line**

  1. the command "createpde" is to create a PDE model of the form (3.32), where the coefficients $m, d, c, a, f$ and the domain $\Omega$ will be specified later. In general, you need to specify the number of equations/unknowns N in the input (so that the line becomes "model = createpde(N)"); nevertheless, you do not need to specify N for a model when N = 1.

  2. the command "geometryFromEdges" is to create the domain $\Omega$. Usually one needs to write a separate function to specify the domain. The way of writing a function for your own domain will be described later.

3. the command “applyBoundaryCondition” is to assign boundary condition to the PDE. Three types of boundary conditions, ‘dirichlet’, ’neumenn’ and ’mixed’, can be specified.

4. the command “specifyCoefficients” is to assign the coefficients $m, d, c, a, f$. When they are constants, they can be assigned simply like what the code shown above. When they not constant, one usually needs to write a separate function to specify these non-constant functions.

5. the command “generateMesh” is to generate a mesh for the domain (for solving PDE using the finite element method).

6. the command “solvepde” is to solve the PDE based on the settings above.

We can make a slight modification of the codes above to solve

$$\begin{aligned} u_t - \Delta u = 1 \qquad & \text{in} \quad \Omega \times (0,\infty)\,, \\ u = u_0 \qquad & \text{on} \quad \Omega \times \{t=0\}\,, \\ u = 0 \qquad & \text{on} \quad \partial\Omega \times (0,\infty)\,, \end{aligned}$$

where $u_0(x,y) = x^2 + y^2$.

```
model = createpde();
geometryFromEdges(model,@lshapeg);
applyBoundaryCondition(model,'dirichlet','Edge',1:model.Geometry.NumEdges,'u',0);
setInitialConditions(model,@(location) location.x.∧2 + location.y.∧2;);
specifyCoefficients(model,'m',0,...
                          'd',1,...
                          'c',1,...
                          'a',0,...
                          'f',1);
generateMesh(model,'Hmax',0.25);
tlist = 0:0.1:1;
results = solvepde(model,tlist);
```

In the code above, we add the fourth line for specifying the initial condition, and in order to solve this time-dependent problem we need a list of time vector (given in “tlist”) for the PDE solver.

For the assignment of the initial data, the variable “location” is in fact a “structure” that the PDE solver will pass to the function, and location.x and location.y denote the x- and y- coordinate of the location. See the assignments of non-constant coefficients for more details.

• **Visualization of the domain**: Before talking about how to create the domain on which the PDEs satisfy, we first talk about how to plot the domain if the domain has been specified by a file (such as “lshapeg” in the example above). One way to plot the domain is as follows:

```
pdegplot(@filename)
```

where “filename” is the function used to describe the domain.

One can add additional parameters to show more information of the domain. For example, for 2-dimensional domain we can use the following command

```
pdegplot(@filename,’EdgeLabels’,’on’,’FaceLabels’,’on’)
```

or

```
pdegplot(model,’EdgeLabels’,’on’,’FaceLabels’,’on’)
```

to show the label of edges and faces.

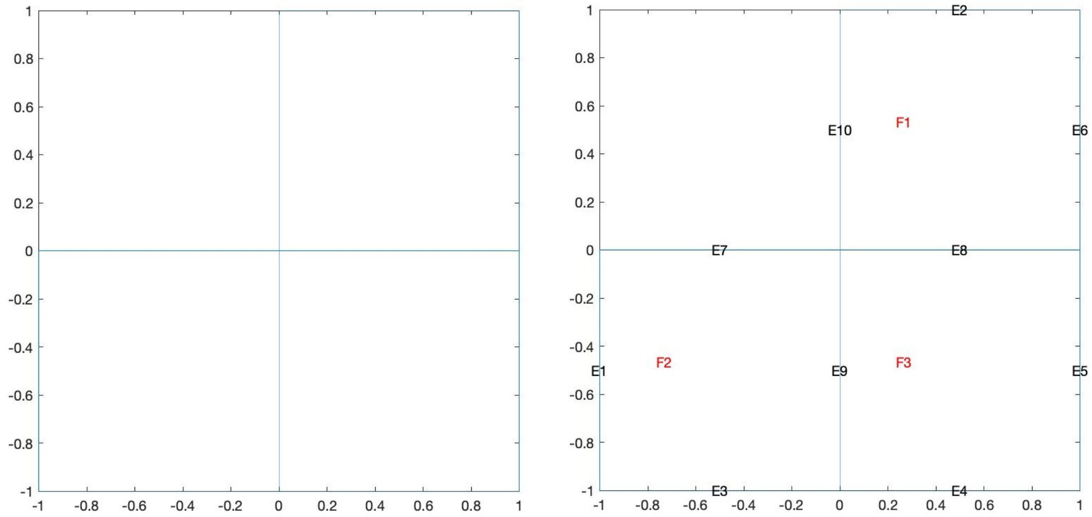


Figure 3.2: Difference between with and without ’EdgeLabels’ and ’FaceLabels’

You can also see the mesh (a type of discretization of the PDE domain) generated by the mesh generator “generateMesh” using

```
pdeplot(model)
```

• **Creation of domain using geometryFromEdges**

The domain on which PDEs are satisfied is described by a function. The function “lshapeg” is a built-in matlab® function which describes an L-shaped region shown in Figure 3.2, and the following codes

```
function [x,y] = Pacman_domain(bs,s)

switch nargin
  case 0
    x = 3;  % total number of curve parametrizations
  return

  case 1
    dl = [-3*pi/4  0  0  % start parameter values of each curve
           3*pi/4  1  1  % end parameter values of each curve
           1       0  1  % The region label on the LHS of each curve
           0       1  0]; % The region label on the RHS of each curve
    x = dl(:,bs);
    return

  case 2
    x = zeros(size(s));
    y = zeros(size(s));
    if numel(bs) == 1
      bs = bs*ones(size(s));
    end
    cbs = find(bs == 1);  % the following two lines describes the 1st curve
    x(cbs) = cos(s(cbs));  % s(cbs) is the arc-length parameter
    y(cbs) = sin(s(cbs));

    cbs = find(bs == 2);  % the following two lines describes the 2nd curve
    x(cbs) = cos(3*pi/4)*s(cbs);
    y(cbs) = sin(3*pi/4)*s(cbs);

    cbs = find(bs == 3);  % the following two lines describes the 3rd curve
    x(cbs) = cos(3*pi/4)*s(cbs);
    y(cbs) = -sin(3*pi/4)*s(cbs);
  return
end
```

describes the following domain

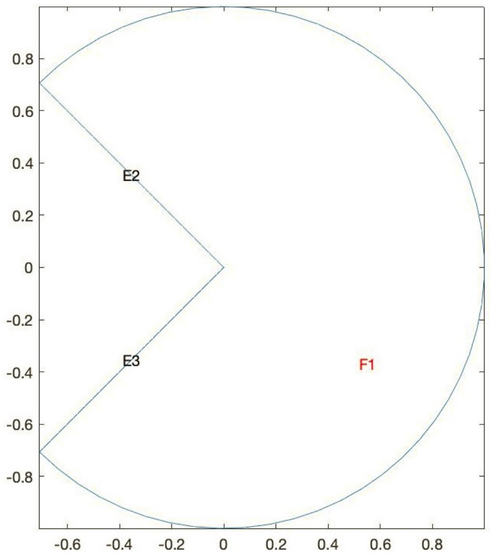


Figure 3.3: The Pacman-like domain

- **Assigning boundary conditions using applyBoundaryCondition**

The ‘mixed’ variable is to assign different boundary conditions to different component of the PDEs, and is seldom used. In the following, we only talk about how to impose Dirichlet and Neumann boundary conditions.

1. One needs to assign functions $h$ and $r$ to specify the Dirichlet boundary condition $hu = r$. If only $r$ is given, $h$ by default is the identity map. For example, one can use

   ```
   applyBoundaryCondition(model,'dirichlet','Edge',1:2,'r',0);
   ```

   or even

   ```
   applyBoundaryCondition(model,'dirichlet','Edge',1:2,'u',0);
   ```

   to impose the Dirichlet boundary condition $u = 0$ on Edge 1 and Edge 2.

2. One needs to assign functions $g$ and $q$ to specify the general Neumann boundary condition $c\dfrac{\partial u}{\partial \mathbf{N}} + qu = g$, where $p$ and $q$ are zero by default ($c$ is one of the coefficient in the PDE and will be specified using “specifyCoefficients”). For example, one can use

   ```
   applyBoundaryCondition(model,'neumann','Edge',3,'q',1,'g',1);
   ```

   to specify the Robin boundary condition $c\dfrac{\partial u}{\partial \mathbf{N}} + u = 1$ on Edge 3, and use

   ```
   applyBoundaryCondition(model,'neumann','Edge',3);
   ```

   to specify the Neumann boundary condition $c\dfrac{\partial u}{\partial \mathbf{N}} = 0$ on Edge 3.

- **Assigning coefficients using specifyCoefficients**: In the simple PDE example the domain is divided into three sub-domains. If some coefficients are constants but different in different sub-domains, one can assign the value of these coefficients by adding some additional switches. The following code is a typical example for assigning the forcing function in three sub-domains.

```
specifyCoefficients(model,’m’,0,’d’,0,’c’,1,’a’,0,’f’,1,’Face’,1);
specifyCoefficients(model,’m’,0,’d’,0,’c’,1,’a’,0,’f’,5,’Face’,2);
specifyCoefficients(model,’m’,0,’d’,0,’c’,1,’a’,0,’f’,-8,’Face’,3);
```

When the coefficients are non-constant, one needs to write separate functions for them. matlab® has a special requirement for these functions: these functions are of the form “**filename(location,state)**”, where “solvepde” passes the location and state **structures** to these functions automatically:

(a) location is a structure with these fields:

- location.x
- location.y
- location.z
- location.subdomain

The fields x, y, and z represent the x-, y-, and z- coordinates of points for which your function calculates coefficient values. The subdomain field represents the subdomain numbers, which currently apply only to 2-D models. The location fields are row vectors.

(b) state is a structure with these fields:

- state.u
- state.ux
- state.uy
- state.uz
- state.time

The state.u field represents the current value of the solution u. The state.ux, state.uy, and state.uz fields are estimates of the solution’s partial derivatives $u_x$, $u_y$ and $u_z$ at the corresponding points of the location structure. The solution and gradient estimates are N-by-Nr matrices, where Nr = length(location.x) (that is, the number of points to be evaluated). The state.time field is a scalar representing time for time-dependent models.

For example, to specify the forcing function

$$f = \begin{bmatrix} x - y + u_1 \\ 1 + \tanh\big(\frac{\partial u_1}{\partial x}\big) + \tanh\big(\frac{\partial u_3}{\partial y}\big) \\ (5 + u_3)\sqrt{x^2 + y^2}\,, \end{bmatrix},$$

one can use

```
function f = fcoeffunction(location,state)

N = 3; % Number of equations
Nr = length(location.x); % Number of columns
f = zeros(N,Nr); % Allocate f

% Now the particular functional form of f
f(1,:) = location.x - location.y + state.u(1,:);
f(2,:) = 1 + tanh(state.ux(1,:)) + tanh(state.uy(3,:));
f(3,:) = (5 + state.u(3,:)).*sqrt(location.x.∧2 + location.y.∧2);
```

and use the following command

```
specifyCoefficients(model,'f',@fcoeffunction,...)
```

to specify this coefficient.

Some requirement for other coefficients:

1. The $i$-th component of the vector $\operatorname{div}(c\nabla u)$, where the unknown $u = (u^1, \cdots, u^{\mathrm{N}})$ (here the superscript $i$ refer to the $i$-th component of $u$), is in general given by

$$\sum_{j=1}^{\mathrm{N}} \sum_{k,\ell=1}^{d} \frac{\partial}{\partial x_k}\Big(c_{ijk\ell}\frac{\partial u^j}{\partial x_\ell}\Big)\,.$$

   (a) If there exists $\nu$ such that $\operatorname{div}(c\nabla u)^i = \sum_{k=1}^{d} \frac{\partial}{\partial x_k}\big(\nu\frac{\partial u^i}{\partial x_k}\big)$ for all $1 \leqslant i \leqslant \mathrm{N}$, one can specify $c$ simply by

   ```
   specifyCoefficients(model,'c',ν,...)
   ```

   if $\nu$ is a constant, or

   ```
   specifyCoefficients(model,'c',@ccoeffunction,...)
   ```
   (3.33)

   if $\mu$ is a non-constant function, where "ccoeffunction" is a function of structures "location" and "state" whose output is a 1-by-Nr row vector, here again Nr = length(location.x) is the number of points to be evaluated.

(b) If there exists $\nu_1, \nu_2, \cdots, \nu_{\rm N}$ such that $\text{div}(c\nabla u)^i = \sum\limits_{k=1}^{d} \frac{\partial}{\partial x_k}\big(\nu_i \frac{\partial u^i}{\partial x_k}\big)$ for all $1 \leqslant i \leqslant {\rm N}$ (or equivalently, each equation has its own diffusion coefficient), one can specify $c$ simply by

$$\boxed{\text{specifyCoefficients(model,'c',}[\nu_1;\nu_2;\cdots;\nu_{\rm N}]\text{,...)}}$$

if all $\nu_j$'s are constants; that is, we assign $c$ as an N-Element column vector $[\nu_1; \nu_2; \cdots; \nu_{\rm N}]$, or using (3.33) if one of $\nu_j$ is a non-constant function, where "ccoeffunction" is a function of structures "location" and "state" whose output is an N-by-Nr matrix whose $i$-th row is the value of $\nu_i$ at point (location.x, location.y, location.z), similar to the one given in "fcoeffunction".

(c) When $c_{ijk\ell}$ is a general tensor elements, the assignment for $c$ is quite complicated and we will not discuss here. Check the manual in matlab® for the detail.

2. The $i$-th component of the vector $mu_{tt}$, $du_t$ and $au$ are in general of the form

$$(mu_{tt})^i = \sum_{j=1}^{N} m_{ij} \frac{\partial^2 u_j}{\partial t^2}\,, \quad (du_t)^i = \sum_{j=1}^{N} d_{ij} \frac{\partial u_j}{\partial t}\,, \quad (au)^i = \sum_{j=1}^{N} a_{ij} u_j\,.$$

(a) If $m_{ij} = \delta_{ij}\mu$ for some $\mu$ so that $(mu_{tt})^i = \mu u^i_{tt}$ for all $1 \leqslant i \leqslant N$, one can specify $m$ by

$$\boxed{\text{specifyCoefficients(model,'m',}\mu\text{,...)}}$$

if $\mu$ is a constant, or

$$\boxed{\text{specifyCoefficients(model,'m',@mcoeffunction,...)}} \tag{3.34}$$

if $\mu$ is a scalar function, where "mcoeffunction" is a function of structures "location" and "state" whose output is a 1-by-Nr row vector, here again Nr = length(location.x) is the number of points to be evaluated. Similar situation applies to $d$ and $a$.

(b) If $m_{ij} = \delta_{ij}\mu_i$ for some $\mu_1, \cdots, \mu_{\rm N}$ so that $(mu_{tt})^i = \mu_i u^i_{tt}$ for all $1 \leqslant i \leqslant N$, one can specify $m$ by

$$\boxed{\text{specifyCoefficients(model,'m',}[\mu_1;\mu_2;\cdots;\mu_{\rm N}]\text{,...)}}$$

if all $\mu_j$'s are constants; that is, assign $m$ as an N-Element column vector $[\mu_1; \mu_2; \cdots; \mu_{\rm N}]$, or using (3.34) if one of $\mu_j$ is a non-constant function, where "mcoeffunction" is a function of structures "location" and "state" whose output is an N-by-Nr matrix whose $i$-th row is the value of $\mu_i$ at point (location.x, location.y, location.z), similar to the one given in "fcoeffunction". Similar situation applies to $d$ and $a$.

(c) For general $m_{ij}$, transform the N × N matrix into one N$^2$-Element column vector in the way

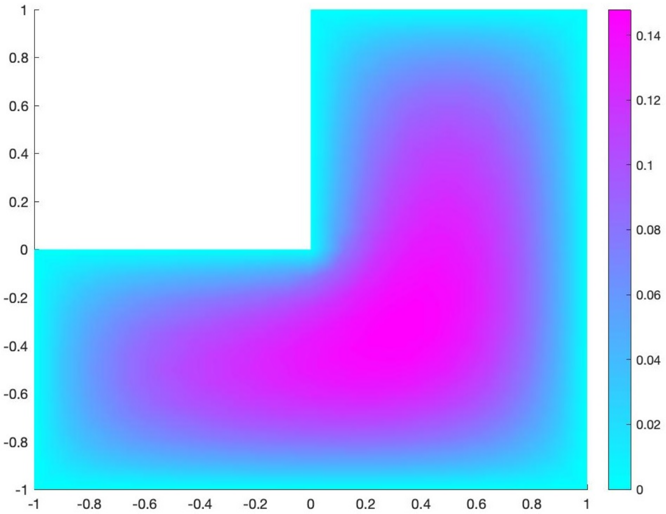
$$\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1\mathrm{N}} \\ m_{21} & m_{22} & \cdots & m_{2\mathrm{N}} \\ \vdots & & \ddots & \vdots \\ m_{\mathrm{N}1} & m_{\mathrm{N}2} & \cdots & m_{\mathrm{NN}} \end{bmatrix} \rightarrow \begin{bmatrix} m_{11} \\ m_{21} \\ \vdots \\ m_{\mathrm{N}1} \\ m_{12} \\ m_{22} \\ \vdots \\ m_{\mathrm{N}2} \\ m_{31} \\ m_{32} \\ \vdots \\ m_{\mathrm{N(N-1)}} \\ m_{\mathrm{NN}} \end{bmatrix}$$

and the output of the function “mcoeffunction” is an N$^2$-by-Nr matrix. Similar situation applies to $d$ and $a$.

• **Visualization of the solution**: The outcome “result” of the PDE solver is a structure that has several variables such as “NodalSolution”, “XGradients”, “YGradients” and “Mesh” stored inside the structure. In order to access these variables, we add “result.” in front of these variables. For example, for time-independent problems, you can use the command

```
pdeplot(model,‘XYData’,results.NodalSolution)
```

to visualize the solution, here “results.NodalSolution” is the value of “NodalSolution” in the results structure.



For time-dependent problems, “NodalSolution” is a matrix whose column represents the value of the solution at different time (given by “tlist”).

**Example 3.8.** Consider a possible extension of the Lotka-Volterra model (2.49)

$$
\begin{aligned}
p_t &= -0.16p + 0.08pq + \operatorname{div}(\kappa_1 \nabla p) && \text{in} \quad \Omega \times (0,T]\,,\\
q_t &= 4.5q - 0.9pq + \operatorname{div}(\kappa_2 \nabla q) && \text{in} \quad \Omega \times (0,T]\,,\\
\frac{\partial p}{\partial \mathbf{N}} &= \frac{\partial q}{\partial \mathbf{N}} = 0 && \text{on} \quad \partial\Omega \times (0,T]\,,\\
p &= p_0\,,\ \ q = q_0 && \text{on} \quad \Omega \times \{t=0\}\,,
\end{aligned}
$$

where the spatial dependence is included into the problem, and $p, q$ denote the population of the fox and the rabbit, respectively. It is reasonable to assume the $\kappa_1$ is a decreasing function of $q$ while $\kappa_2$ is an increasing function of $p$ (which indicates that foxes intend to stay in a place with more rabbits, but rabbits intend to leave a place with more foxes), so for a baby model we assume that

$$
\kappa_1 = \kappa_1(q) = 2 - \tanh(q) \qquad \text{and} \qquad \kappa_2 = \kappa_2(p) = 2 + \tanh(p)\,.
$$

The domain under consideration is given by the function "Pacman_domain" given on page 83. We assume the initial conditions

$$
p_0(x,y) = \left\{ \begin{array}{ll} 0 & \text{if } y \geqslant 0\,,\\ 5 & \text{if } y < 0\,, \end{array} \right. \qquad \text{and} \quad q_0(x,y) = \left\{ \begin{array}{ll} 3 & \text{if } y > 0\,,\\ 0 & \text{if } y \leqslant 0\,, \end{array} \right.
$$

which indicates that initially the foxes and the rabbits are separated into two regions. We also note that Neumann boundary condition are imposed so that the foxes and rabbits are not allowed to leave the region.

The following codes can be used to simulated the PDE described above.

```
model = createpde(2);
geometryFromEdges(model,@Pacman_domain);
initfun = @(location) [5*(location.y < 0); 3*(location.y > 0)];
setInitialConditions(model,initfun);
applyBoundaryCondition(model,’neumann’,’Edge’,1:3,’g’,[0;0]);
ccoeffunction = @(location,state) [2-tanh(state.u(2,:)); 2+tanh(state.u(1,:))];
acoeffunction = @(location,state) [0.16-0.08*state.u(2,:); 0.9*state.u(1,:)-4.5];
specifyCoefficients(model,’m’,0,...
                          ’d’,1,...
                          ’c’,ccoeffunction,...
                          ’a’,acoeffunction,...
                          ’f’,[0;0]);
generateMesh(model,’Hmax’,0.2);
tlist = 0:0.01:0.1;
results = solvepde(model,tlist);
```

# 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.

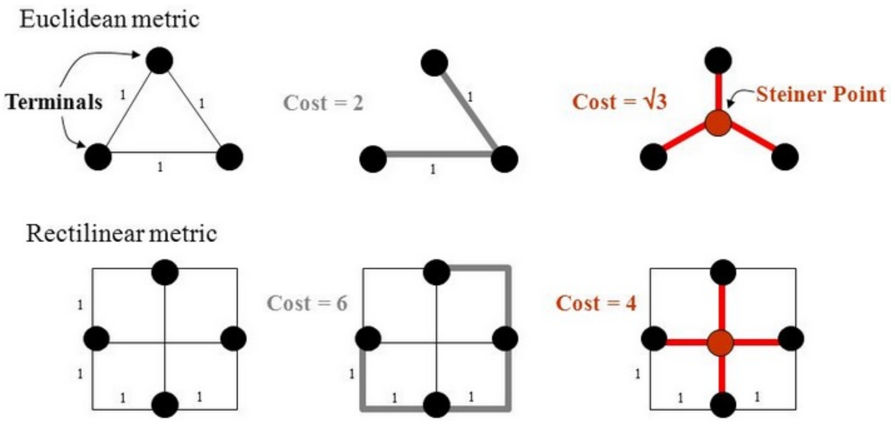


Figure 4.1

### 4.1.3 Separation problem (分群問題)

Suppose that we are given two types of points in $\mathbb{R}^n$: points of type A $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\cdots$, $\boldsymbol{x}_m$ and points of type B $\boldsymbol{x}_{m+1}$, $\boldsymbol{x}_{m+2}$, $\cdots$, $\boldsymbol{x}_{m+p}$. The goal of the separation problem is to find a ***linear separator***, a hyperplane of the form

$$H(\boldsymbol{\omega},\beta) \equiv \left\{\boldsymbol{x}\in\mathbb{R}^n \,\middle|\, \boldsymbol{\omega}\cdot\boldsymbol{x}+\beta = 0\right\}$$

for which points of type A and points of type B are on opposite sides of the hyperplane and the hyperplane is the "farthest" as possible from all points.

The margin of the separator is the distance of the separator from the closest point, as illustrated in Figure 4.2. In mathematics,

$$\text{margin} = \min_{1\leqslant i\leqslant m+p} \frac{|\boldsymbol{\omega}\cdot\boldsymbol{x}_i+\beta|}{\|\boldsymbol{\omega}\|_2}$$

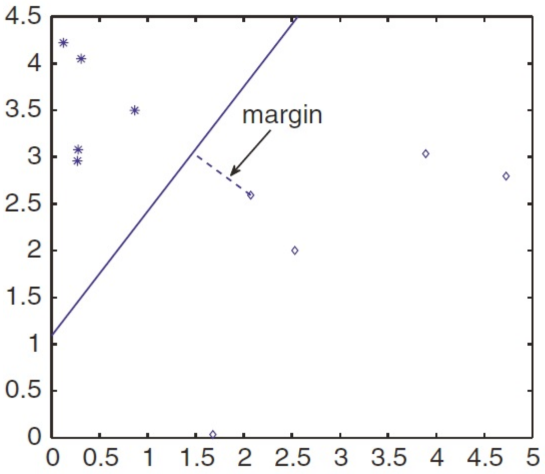


Figure 4.2: A linear separator that separates points of type $*$ and point of type $\diamond$

The separation problem will thus consist of finding the linear separator with the largest margin:

$$\max_{(\boldsymbol{\omega},\beta)\in\mathbb{R}^{n+1}} \left\{ \min_{1\leqslant i\leqslant m+p} \frac{|\boldsymbol{\omega}\cdot\boldsymbol{x}_i+\beta|}{\|\boldsymbol{\omega}\|_2} \right\} \text{ subject to } \left\{ \begin{array}{ll} \boldsymbol{\omega}\cdot\boldsymbol{x}_i+\beta<0 & \text{for } 1\leqslant i\leqslant m\,, \\ \boldsymbol{\omega}\cdot\boldsymbol{x}_i+\beta>0 & \text{for } m+1\leqslant i\leqslant m+p\,. \end{array}\right.$$

### 4.1.4 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 closed 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 perimeter $L$. We look for one $P$ in $\mathscr{P}_n$ which encloses the largest area. For a given set of points $B_1, \cdots, B_m$, let $[B_1, B_2, \cdots, B_m, B_1]$ denote the polygon with edges $\overline{B_1B_2}$, $\overline{B_2B_3}$, $\cdots$, $\overline{B_{m-1}B_m}$ and $\overline{B_mB_1}$. Suppose that

$$P_n = [A_1, A_2, \cdots, A_n, A_{n+1}, \cdots, A_{2n}, A_1]$$

is a polygon in $\mathscr{P}_n$ which encloses the largest area. We use the notion $A_j = A_k$ if $j = k$ (mod $2n$).

**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 optimization problem can be reformulated by looking for "minimizer" of a certain functional in the space of piecewise continuously differentiable closed curve. To be more precise, we look for curves $C$ that can be parameterized, using the arc-length, by vector-valued function $\boldsymbol{r}(s) = x(s)\mathbf{i} + y(s)\mathbf{j}$ in the set

$$\mathcal{A} = \Big\{\boldsymbol{r}(s) = x(s)\mathbf{i} + y(s)\mathbf{j}\,\Big|\, x, y \in \mathscr{D}^1([0,L];\mathbb{R}), \boldsymbol{r}(0) = \boldsymbol{r}(L), |\dot{\boldsymbol{r}}(s)|^2 = 1 \text{ for all } s \in [0,L]\Big\},$$

where $\mathscr{D}^1([0,L];\mathbb{R})$ consists of continuous, piecewise continuously differentiable functions on $[0,L]$, so that the functional

$$-\int_0^L \big[x(s)\dot{y}(s) - \dot{x}(s)y(s)\big]\,ds\,.$$

is minimized. The problem above is equivalent to the "minimization" problem

$$\max_{\boldsymbol{r}=x\mathbf{i}+y\mathbf{j}\in\mathcal{A}} \int_0^L \big[x(s)\dot{y}(s) - \dot{x}(s)y(s)\big]\,ds\,.$$

### 4.1.5 Minimal surface of revolution

This is a problem of finding a curve $C$ connecting two given points $(x_0, y_0)$ 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 $C = \big\{(x, y(x)) \,\big|\, y \in \mathscr{D}^1([x_0, x_1]; \mathbb{R}), y(x_0) = y_0, y(x_1) = y_1\big\}$ is given by

$$2\pi \int_{x_0}^{x_1} y(x)\sqrt{1 + y'(x)^2}\, dx\,.$$

Therefore, the problem of minimal surface of revolution is to find a function $y \in \mathcal{A} \equiv \big\{y \in \mathscr{D}^1([x_0, x_1]; \mathbb{R}) \,\big|\, y(x_0) = y_0, y(x_1) = y_1\big\}$ which minimizes the functional

$$I(y) = 2\pi \int_{x_0}^{x_1} y(x)\sqrt{1 + y'(x)^2}\, dx\,.$$

### 4.1.6 Newton's problem

The Newton problem is to find a curve $C$ connecting two given points $(x_0, y_0)$ 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 $v\mathbf{i}$).

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 $\varphi$, 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 $\mathbf{e}_2$ and $\mathbf{e}_3$ components all cancel 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 $y \in \mathcal{A} \equiv \big\{y \in \mathscr{D}^1([x_0, x_1]; \mathbb{R}) \,\big|\, y(x_0) = y_0, y(x_1) = y_1\big\}$ which minimizes $I(y)$".

**Newton's model**: $\varphi(u) = u^2$.

### 4.1.7 Brachistochrone problem (最速下降曲線問題)

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 points $(0,0)$ and $(a,b)$, where $a>0$ and $b<0$, what is the brachistochrone curve connecting $(0,0)$ and $(a,b)$?

Given a curve parameterized by $\big\{\big(x,y(x)\big)\,\big|\,x\in[0,a]\big\}$ for some function $y\in\mathscr{D}^1([0,a];\mathbb{R})$, the total time required to travel from $(0,0)$ to $(a,b)$ is given by

$$T(y)=\int_0^a\frac{\sqrt{1+y'(x)^2}}{\sqrt{-2gy(x)}}\,dx\,.$$

Therefore, the brachistochrone problem can be formulated as finding $y\in\mathcal{A}=\big\{y\in\mathscr{D}^1([0,a];\mathbb{R})\,\big|\,y(0)=0,y(a)=b\big\}$ such that $T(y)$ is minimized. In other words, the minimizer $\widehat{y}$ satisfies that

$$T(\widehat{y})=\inf_{y\in\mathcal{A}}\int_0^a\frac{\sqrt{1+y'(x)^2}}{\sqrt{-2gy(x)}}\,dx\,.$$

### 4.1.8 Plateau's problem - minimal surface problem (極小曲面問題)

The minimal surface problem is to find a (smooth) surface $\Sigma$ whose boundary is a given curve $C$ but has the minimal surface area. Consider the simplest case that the orthogonal projection from space onto the $xy$-plane is a bijection between the curve $C$ and the boundary of a simply connected region $\Omega$ on the $xy$-plane. In this case, there exists a continuous function $f:\partial\Omega\to\mathbb{R}$ so that

$$C=\big\{x\mathbf{i}+y\mathbf{j}+f(x,y)\mathbf{k}\,\big|\,(x,y)\in\partial\Omega\big\}\,.$$

The goal is then to find a (smooth) function $z=u(x,y)$ defined on $\Omega$ such that $u=f$ on $\partial\Omega$ and

$$\int_\Omega\sqrt{1+u_x(x,y)^2+u_y(x,y)^2}\,dA=\min_{v\in\mathcal{A}}\int_\Omega\sqrt{1+v_x(x,y)^2+v_y(x,y)^2}\,dA\,,$$

where $\mathcal{A}$ is the admissible set

$$\mathcal{A}=\big\{v:\bar{\Omega}\to\mathbb{R}\,\big|\,v\text{ is (piecewise) differentiable on }\Omega\text{ and }v=f\text{ on }\partial\Omega\big\}\,.$$

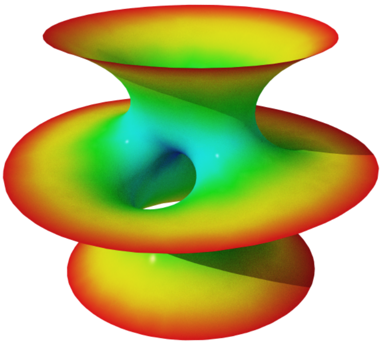


Figure 4.3: Costa's Minimal Surface - the minimal surface with three circles as prescribed boundaries.

### 4.1.9 Image processing

An image can often be viewed as a function defined on a square domain. In many problems in image processing, the goal is to recover an ideal image $u$ from an observation $f$, where $u$ is a perfect original image describing a real scene, $f$ is an observed image, which is a degraded version of $u$. The degradation can be due to:

1. Signal transmission: there can be some noise (random perturbation).
2. Defects of the imaging system: there can be some blur (deterministic perturbation).

The simplest modelization is the following:

$$f = Ku + n\,,$$

where $n$ is the noise, and $K$ is the blur, a linear operator (for example a convolution). The following assumptions are classical:

1. $K$ is known (but often not invertible);
2. Only some statistics (mean, variance, $\cdots$) are known of $n$.

A classical approach in the image processing problems consists in introducing a regularization term $L$ which admits a unique solution of the optimization problem

$$\inf_{u\in\mathcal{A}}\Big(\int_\Omega |f{-}Ku|^2\,dA + \lambda L(u)\Big)\,,$$

where $\mathcal{A}$ is an admissible set which describes the requirement for the real images, and $L$ is a non-negative function (with certain requirements that we will not explore here).

**Example 4.2.** Suppose that the polluted image $f$ is solely due to noise (so $K = \mathrm{Id}$, the identity map). The ROF model is a model for denoise which requires the minimization of the functional

$$\int_\Omega |f{-}u|^2\,dA + \lambda\int_\Omega |\nabla u|\,dA\,,$$

where $u$ should picked up in the admissible set

$$\mathcal{A} = \Big\{u:\Omega\to\mathbb{R}\,\Big|\,\int_\Omega |\nabla u|\,dA < \infty \text{ and } \frac{\partial u}{\partial \mathbf{N}} = 0 \text{ on } \partial\Omega\Big\}\,.$$

## 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 $y\in\mathscr{C}^1([a,b];\mathbb{R})$ or $\mathscr{D}^1([a,b];\mathbb{R})$, and $y$ satisfies the boundary condition $y(a) = A_0, y(b) = B_0$, where $\mathscr{C}^1([a,b];\mathbb{R})$ denotes the space of continuously differentiable real-valued functions defined on $[a,b]$, and $\mathscr{D}^1([a,b];\mathbb{R})$ denotes the space of continuous, piecewise continuously differentiable real-valued functions defined on $[a,b]$. In other words, with $\mathcal{A}$ denoting either the set $\big\{y \in \mathscr{C}^1([a,b];\mathbb{R}) \,\big|\, y(a) = A_0, y(b) = B_0\big\}$ or $\big\{y \in \mathscr{D}^1([a,b];\mathbb{R}) \,\big|\, y(a) = A_0, y(b) = B_0\big\}$, 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 $\underset{z\in\mathcal{A}}{\arg\min}\, I(z)$ denote the minimizer, if exists, of the minimization problem $\min\limits_{z\in\mathcal{A}} I(z)$. In other word, if $y = \underset{z\in\mathcal{A}}{\arg\min}\, I(z)$, then $y \in \mathcal{A}$ and

$$I(y) \leqslant I(z) \qquad \forall\, z \in \mathcal{A}\,.$$

**Remark 4.3.** Let

$$\begin{aligned}\mathcal{X} &= \big\{y \in \mathscr{C}^1([a,b];\mathbb{R}) \,\big|\, y(a) = A_0, y(b) = B_0\big\}\\ \mathcal{Y} &= \big\{y \in \mathscr{D}^1([a,b];\mathbb{R}) \,\big|\, y(a) = A_0, y(b) = B_0\big\}\,.\end{aligned}$$

Then $\underset{z\in\mathcal{X}}{\arg\min}\, I(z)$, if exists, equals $\underset{z\in\mathcal{Y}}{\arg\min}\, I(z)$. To see this, we first note that $\min\limits_{z\in\mathcal{X}} I(z) \geqslant \min\limits_{z\in\mathcal{Y}} I(z)$; thus for $\underset{z\in\mathcal{X}}{\arg\min}\, I(z) \neq \underset{z\in\mathcal{Y}}{\arg\min}\, I(z)$ to hold, we must have $\widehat{y} \in \mathcal{Y}\backslash\mathcal{X}$ such that $I(\widehat{y}) < \min\limits_{z\in\mathcal{X}} I(z)$. By smooth $\widehat{y}$ at corners, we obtain $\bar{y} \in \mathcal{X}$ such that $I(\bar{y}) < \min\limits_{z\in\mathcal{X}} I(z)$, a contradiction.

However, it is possible that there are only minimizers in $\mathscr{D}^1([a,b];\mathbb{R})$. See Example 4.17 for the detail.

### 4.2.1 First variation of $I$

Let $\mathcal{A} = \big\{y \in \mathscr{D}^1([a,b];\mathbb{R}) \,\big|\, y(a) = A_0, y(b) = B_0\big\}$ and $\mathcal{N} = \big\{\eta \in \mathscr{D}^1([a,b];\mathbb{R}) \,\big|\, \eta(a) = \eta(b) = 0\big\}$, 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

$$\frac{J(\epsilon) - J(0)}{\epsilon} = \frac{1}{\epsilon}\int_a^b \big[L(x, y(x) + \epsilon\eta(x), y'(x) + \epsilon\eta'(x)) - L(x, y(x), y'(x))\big]\, dx \quad \forall\, \epsilon \neq 0\,.$$

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 $\eta$.

**Theorem 4.4.** *If $y = \underset{z\in\mathcal{A}}{\arg\min}\, I(z)$ is a minimizer of $I$, then $\delta I(y;\eta) = 0$ for all $\eta \in \mathcal{N}$.*

**Definition 4.5.** 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 4.6.** *If $y \in \mathscr{C}([a,b];\mathbb{R})$ and $\int_a^b y(x)\eta(x)\,dx = 0$ for all $\eta \in \mathscr{C}([a,b];\mathbb{R})$, then $y \equiv 0$.*

*Proof.* If $y \in \mathscr{C}([a,b];\mathbb{R})$ and $\int_a^b y(x)\eta(x)\,dx = 0$ for all $\eta \in \mathscr{C}([a,b];\mathbb{R})$, then it is possible to plug in $\eta = y$ in the integral equation to obtain

$$\int_a^b |y(x)|^2\,dx = 0$$

which, by the continuity of $y$, shows that $y$ is the zero function. □

**Remark 4.7.** The conclusion in Lemma 4.6 holds true if the test function $\eta$ is chosen from $\mathscr{D}^1([a,b];\mathbb{R})$; that is, if $y \in \mathscr{C}([a,b];\mathbb{R})$, then

$$y \equiv 0 \qquad \Leftrightarrow \qquad \int_a^b y(x)\eta(x)\,dx = 0 \quad \forall\,\eta \in \mathscr{D}^1([a,b];\mathbb{R})\,.$$

The proof of this conclusion requires more tools in analysis, and we will not explore it here.

**Lemma 4.8.** *If $y \in \mathscr{C}([a,b];\mathbb{R})$ and $\int_a^b y(x)\eta'(x)\,dx = 0$ for all $\eta \in \mathscr{N}$, then $y \equiv c$ for some constant $c$.*

*Proof.* Let $\eta(x) = \int_a^x \big(y(t)-c\big)\,dt$, where the constant $c$ is chosen so that $\int_a^b \big(y(t)-c\big)\,dt = 0$. Then $\eta \in \mathscr{N}$ 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 4.9.** *If $y, z \in \mathscr{C}([a,b];\mathbb{R})$ satisfy*

$$\int_a^b \big[y(x)\eta(x) + z(x)\eta'(x)\big]\,dx = 0 \qquad \forall\,\eta \in \mathscr{N}\,, \tag{4.3}$$

*then $z \in \mathscr{C}^1([a,b];\mathbb{R})$ and $z'(x) = y(x)$ for all $x \in [a,b]$.*

*Proof.* Let $z_1(x) = \int_a^x y(t)\,dt$. 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

$$\int_a^b \big[z(x) - z_1(x)\big]\eta'(x)\,dx = 0 \qquad \forall\,\eta \in \mathscr{N}\,.$$

By Lemma 4.8, $z(x) - z_1(x) = C$ for some constant $C$. Therefore, $z(x) = C + \int_a^x y(t)\,dt$ which implies that $z \in \mathscr{C}^1([a,b];\mathbb{R})$ and $z'(x) = y(x)$. □

**Lemma 4.10.** *Suppose that $y, z \in \mathscr{C}([a,b];\mathbb{R})$ and $z$ is not a constant function. If*

$$\int_a^b y(x)\eta'(x)\,dx = 0 \qquad \forall\, \eta \in \mathscr{N} \text{ and } \eta \text{ satisfies } \int_a^b z(x)\eta'(x)\,dx = 0\,,$$

*then there are constants $\lambda, \mu \in \mathbb{R}$ such that $y(x) = \lambda z(x) + \mu$.*

*Proof.* Let $\eta(x) = \int_a^x \big(y(t) - \lambda z(t) - \mu\big)\,dt$, where $\lambda, \mu$ are chosen so that $\eta(b) = 0$ and $\int_a^b z(x)\eta'(x)\,dx = 0$; 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 $(\lambda, \mu)$. Since $\eta \in \mathscr{N}$ and satisfies $\int_a^b z(x)\eta'(x)\,dx = 0$, we have

$$\int_a^b \big|y(x) - \lambda z(x) - \mu\big|^2\,dx = \int_a^b \big(y(x) - \lambda z(x) - \mu)\eta'(x)\,dx = -\mu \int_a^b \eta'(x)\,dx = 0\,;$$

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 $\eta \in \mathscr{N}$.

**Theorem 4.11.** *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 4.4 and Lemma 4.9 to each interval on which $\widehat{y}$ is of class $\mathscr{C}^1$. □

**Definition 4.12.** Equation (4.4) is called (the ***strong form*** of) the Euler-Lagrange equation (associated with the minimization problem (4.2)).

**Remark 4.13.** 1. Theorem 4.11 is essentially due to Du Bois-Reymond, so (4.4) is also called the Du Bois-Reymond equation.

2. If $\widehat{y} \in \mathscr{C}^2([a,b];\mathbb{R})$ and $L_{px}, L_{yp}, L_{pp}$ are continuous, then $\widehat{y}$ satisfies the following second order ODE

$$\begin{aligned} &L_{pp}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}''(x) \\ &\qquad = L_y(x, \widehat{y}(x), \widehat{y}'(x)) - L_{px}(x, \widehat{y}(x), \widehat{y}'(x)) - L_{py}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}'(x)\,. \end{aligned}$$

This is the equation that Euler originally derived/obtained.

**Example 4.14.** Now we consider the brachistochrone problem. Making the change of variable $y \mapsto -y$ (and ignoring $\sqrt{2g}$ in the denominator), we rewritten the minimization problem as

$$\inf_{y\in\mathcal{A}} \int_0^a \frac{\sqrt{1+y'(x)^2}}{\sqrt{y(x)}}\,dx$$

where $\mathcal{A} = \big\{y \in \mathscr{D}^1([0,a];\mathbb{R}) \,\big|\, y(0) = 0, y(a) = -b\big\}$. Therefore, $L(x,y,p) = \dfrac{\sqrt{1+p^2}}{\sqrt{y}}$ which implies that the Euler-Lagrange equation for the brachistochrone problem is

$$\frac{d}{dx}\frac{y'}{\sqrt{y}\sqrt{1+y'^2}} = -\frac{\sqrt{1+y'^2}}{2y^{\frac{3}{2}}}\,.$$

Similarly, 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 4.15.** *Suppose that $\widehat{y} \in \mathscr{D}^1([a,b];\mathbb{R})$ 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 4.11 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,

$$\begin{aligned}&L_p(x+\epsilon,\widehat{y}(x+\epsilon),\widehat{y}'(x+\epsilon)) - L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon))\\ &\quad = L_p(x+\epsilon,\widehat{y}(x+\epsilon),\widehat{y}'(x+\epsilon)) - L_p(x,\widehat{y}(x+\epsilon),\widehat{y}'(x+\epsilon))\\ &\qquad + L_p(x,\widehat{y}(x+\epsilon),\widehat{y}'(x+\epsilon)) - L_p(x,\widehat{y}(x),\widehat{y}'(x+\epsilon))\\ &\quad = L_{px}(x+\theta_1\epsilon,\widehat{y}(x+\epsilon),\widehat{y}'(x+\epsilon))\epsilon\\ &\qquad + L_{py}(x,\widehat{y}(x)+\theta_2(\widehat{y}(x+\epsilon)-\widehat{y}(x)),\widehat{y}'(x+\epsilon))(\widehat{y}(x+\epsilon)-\widehat{y}(x))\end{aligned}$$

for some $\theta_1 = \theta_1(\epsilon, x)$ and $\theta_2 = \theta_2(\epsilon, x)$ satisfying $0 < \theta_1, \theta_2 < 1$. 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 the limit

$$\begin{aligned}&\lim_{\epsilon\to 0} \frac{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))}{\epsilon}\\ &\qquad = L_y(x, \widehat{y}(x), \widehat{y}'(x)) - L_{px}(x, \widehat{y}(x), \widehat{y}'(x)) - L_{py}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}'(x)\end{aligned} \tag{4.5}$$

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

$$\widehat{y}''(x) = \lim_{\epsilon\to 0} \frac{\widehat{y}'(x+\epsilon) - \widehat{y}'(x)}{\epsilon} = 0$$

exists.

Using (4.6) and the fact that $\widehat{y}''(x)$ does not exist, there exists a sequence $\{\epsilon_j\}_{j=1}^{\infty}$ with limit 0 such that $\widehat{y}(x+\epsilon_j) \neq \widehat{y}(x)$ for all $j \in \mathbb{N}$ and

$$\liminf_{j\to\infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j} < \limsup_{j\to\infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j}\,. \tag{4.7}$$

By the definition of $L_{pp}$ and the continuity of $\widehat{y}'$ at $x$,

$$\lim_{j\to\infty} \frac{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon_j)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))}{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)} = L_{pp}(x, \widehat{y}(x), \widehat{y}'(x))\,.$$

The condition $L_{pp}(x, \widehat{y}(x), \widehat{y}'(x)) \neq 0$ further implies that

$$\lim_{j\to\infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon_j)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))} = \frac{1}{L_{pp}(x, \widehat{y}(x), \widehat{y}'(x))}\,.$$

We then conclude from (4.5) that the limit

$$\begin{aligned}&\lim_{j\to\infty} \frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{\epsilon_j}\\ &\quad = \lim_{j\to\infty} \Big[\frac{\widehat{y}'(x+\epsilon_j) - \widehat{y}'(x)}{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon_j)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))} \frac{L_p(x, \widehat{y}(x), \widehat{y}'(x+\epsilon_j)) - L_p(x, \widehat{y}(x), \widehat{y}'(x))}{\epsilon_j}\Big]\\ &\quad = \frac{L_y(x, \widehat{y}(x), \widehat{y}'(x)) - L_{px}(x, \widehat{y}(x), \widehat{y}'(x)) - L_{py}(x, \widehat{y}(x), \widehat{y}'(x))\widehat{y}'(x)}{L_{pp}(x, \widehat{y}(x), \widehat{y}'(x))}\,.\end{aligned}$$

exist, a contradiction to (4.7). □

**Remark 4.16.** Let $\widehat{y} = \arg\min_{z\in\mathcal{A}} I(z)$. If $L_{px}$, $L_{py}$, $L_{pp}$ 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 in a neighborhood of $x$, then $\widehat{y}''$ exists in a neighborhood of $x$ and is continuous there.

**Example 4.17.** Let $\mathcal{A} = \big\{y \in \mathscr{D}^1([0,1];\mathbb{R}) \,\big|\, y(0) = y(1) = 0\big\}$. 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 $p^2 \neq \dfrac{1}{3}$ the minimizer $\widehat{y}$ satisfies

$$2\widehat{y}'^2\widehat{y}'' + (\widehat{y}'^2 - 1)\widehat{y}'' = 0$$

on points at which $\widehat{y}'$ is continuous. Therefore, $\widehat{y}''(3\widehat{y}'^2 - 1) = 0$ on points at which $\widehat{y}'$ is continuous if $\widehat{y}'^2 \neq \dfrac{1}{3}$. Therefore, $\widehat{y}'' = 0$ if $\widehat{y}'^2 \neq \dfrac{1}{3}$ 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 4.18** (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 $\mathcal{A} = \big\{y \in \mathscr{D}^2([a,b];\mathbb{R}) \,\big|\, y(a) = A_0, y(b) = B_0, y'(a) = A_1, y'(b) = B_1\big\}$. We note that the corresponding test function space is

$$\mathcal{N} = \big\{y \in \mathscr{D}^2([a,b];\mathbb{R}) \,\big|\, y(a) = y(b) = y'(a) = y'(b) = 0\big\}\,.$$

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 $L = L(x, y, p, q)$,

$$\begin{aligned}\int_a^b \Big[&L_y(x, \widehat{y}(x), \widehat{y}'(x), \widehat{y}''(x))\eta(x) + L_p(x, \widehat{y}(x), \widehat{y}'(x), \widehat{y}''(x))\eta'(x)\\ &+ L_q(x, \widehat{y}(x), \widehat{y}'(x), \widehat{y}''(x))\eta''(x)\Big]\, dx = 0 \qquad \forall\, \eta \in \mathcal{N}\,.\end{aligned}$$

2. **Free ends**: This is to consider the minimization problem

$$\inf_{y\in\mathscr{D}^1([a,b];\mathbb{R})} \int_a^b L(x, y(x), y'(x))\, dx\,.$$

In this case, the test function space is then $\mathscr{N} = \mathscr{D}^1([a,b];\mathbb{R})$. The same argument implies that

$$\int_a^b \big[L_y(x,\widehat{y}(x),\widehat{y}'(x))\eta(x) + L_p(x,\widehat{y}(x),\widehat{y}'(x))\eta'(x)\big]\,dx = 0 \qquad \forall\,\eta\in\mathscr{N} \tag{4.8}$$

if $\widehat{y}$ is a minimizer. In particular, the integral equation above holds for all $\eta \in \big\{y \in \mathscr{D}^1([a,b];\mathbb{R})\,\big|\,y(a)=y(b)=0\big\}$; thus Lemma 4.9 shows that if $L_y$ and $L_p$ are continuous, then

$$\frac{d}{dx}L_p(x,\widehat{y}(x),\widehat{y}'(x)) = L_y(x,\widehat{y}(x),\widehat{y}'(x))$$

for point $x$ at which $\widehat{y}'$ is continuous. Integrating-by-parts of (4.8) further implies that

$$L_p(b,\widehat{y}(b),\widehat{y}'(b))\eta(b) - L_p(a,\widehat{y}(a),\widehat{y}'(a))\eta(a) = 0 \qquad \forall\,\eta\in\mathscr{N}\,.$$

Choosing $\eta\in\mathscr{N}$ so that $\eta(a)=1$ and $\eta(b)=0$ (such $\eta$ always exists), we find that

$$L_p(a,\widehat{y}(a),\widehat{y}'(a)) = 0\,.$$

Similarly, the choice of $\eta\in\mathscr{N}$ satisfying $\eta(a)=0$ and $\eta(b)=1$ shows that

$$L_p(b,\widehat{y}(b),\widehat{y}'(b)) = 0\,.$$

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

$$\mathcal{A} = \big\{\boldsymbol{y} = (y_1,\cdots,y_n) : [a,b]\to\mathbb{R}^n\,\big|\,y_j\in\mathscr{D}^1([a,b];\mathbb{R}) \text{ for } 1\leqslant j\leqslant n\,,\\ \boldsymbol{y}(a)=\boldsymbol{A}_0,\,\boldsymbol{y}(b)=\boldsymbol{B}_0\big\}$$

or (when considering minimization problems with free ends)

$$\mathcal{A} = \big\{\boldsymbol{y} = (y_1,\cdots,y_n) : [a,b]\to\mathbb{R}^n\,\big|\,y_j\in\mathscr{D}^1([a,b];\mathbb{R}) \text{ for } 1\leqslant j\leqslant n\big\} \equiv \mathscr{D}^1([a,b];\mathbb{R}^n)\,,$$

and $L:[a,b]\times\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$. 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

$$\frac{d}{dx}L_{p_i}(x,\boldsymbol{y}(x),\boldsymbol{y}'(x)) = L_{y_i}(x,\boldsymbol{y}(x),\boldsymbol{y}'(x)) \qquad \text{for} \quad i=1,2,\cdots,n\,. \tag{4.9}$$

When considering free ends problem, natural boundary conditions

$$L_{p_i}(b,\widehat{\boldsymbol{y}}(b),\widehat{\boldsymbol{y}}'(b)) = L_{p_i}(a,\widehat{\boldsymbol{y}}(a),\widehat{\boldsymbol{y}}'(a)) = 0 \qquad \text{for} \quad i=1,2,\cdots,n \tag{4.10}$$

have to be imposed for the minimizer $\boldsymbol{y}$.

4. **Several independent variables**: Let $\Omega \subseteq \mathbb{R}^n$ be bounded open set, and $L : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ (here we write $L = L(x, y, p_1, \cdots, p_n)$) be continuous. Consider the minimization problem
$$\inf_{y \in \mathcal{A}} \int_\Omega L(x, y(x), \nabla y(x))\, dx\,,$$
where $\mathcal{A}$ could be

   (a) $\mathcal{A} = \big\{ y \in \mathscr{D}^1(\bar{\Omega};\mathbb{R}) \,\big|\, y = f \text{ on } \partial\Omega \big\}$ (with corresponding $\mathcal{N} = \big\{ \eta \in \mathscr{D}^1(\bar{\Omega};\mathbb{R}) \,\big|\, \eta = 0 \text{ on } \partial\Omega \big\}$) when considering the fixed-end problem, or

   (b) $\mathcal{A} = \mathscr{D}^1(\bar{\Omega};\mathbb{R})$ (with corresponding $\mathcal{N} = \mathscr{D}^1(\bar{\Omega};\mathbb{R})$) 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 $\epsilon \in \mathbb{R}$. The weak form of the Euler-Lagrange equation is $J'(0) = 0$:
$$\int_\Omega \Big[ L_y(x, \widehat{y}(x), \nabla\widehat{y}(x))\eta(x) + (\nabla_p L)(x, \widehat{y}(x), \nabla\widehat{y}(x)) \cdot \nabla_x \eta(x) \Big]\, dx = 0 \quad \forall\, \eta \in \mathcal{N}\,,$$
where $\nabla_p L = \Big( \dfrac{\partial L}{\partial p_1}, \dfrac{\partial L}{\partial p_2}, \cdots, \dfrac{\partial L}{\partial p_n} \Big)$ is the gradient of $L$ in $p$-variable. By the divergence theorem (Theorem A.51), the strong form of the Euler-Lagrange equation is
$$\operatorname{div}_x \big[ (\nabla_p L)(x, \widehat{y}(x), \nabla\widehat{y}(x)) \big] = L_y(x, \widehat{y}(x), \nabla\widehat{y}(x))\,. \tag{4.11}$$

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 4.20 and 4.22 for deriving the weak form of the Euler-Lagrange equation for minimizers.

**Example 4.19** (The minimal surface)**.** In this example we revisit Plateau's problem. 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 $\varphi \in \mathscr{C}^1(\overline{\Omega};\mathbb{R})$, 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 $u$ minimizes $A$, then $\delta A(u;\varphi) = 0$ for all $\varphi \in \mathscr{C}^1(\Omega;\mathbb{R})$ satisfying $\varphi = 0$ on $\partial\Omega$. Assuming that $u \in \mathscr{C}^2(\overline{\Omega};\mathbb{R})$, by the divergence theorem (Theorem A.51, or Green's Theorem in divergence form) 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***

$$(1+u_y^2)u_{xx} - 2u_xu_yu_{xy} + (1+u_x^2)u_{yy} = 0 \qquad \text{in} \quad \Omega\,. \tag{4.12}$$

**Example 4.20** (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

$$\mathcal{A} = \Big\{ \boldsymbol{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} \in \mathscr{D}^1([0,1];\mathbb{R}^2) \,\Big|\, \boldsymbol{r}(0) = \boldsymbol{r}(1)\,, \int_0^1 (x\dot{y} - y\dot{x})dt = 2\text{A} \Big\}$$

and $I(\boldsymbol{r}) = \inf\limits_{\boldsymbol{r}\in\mathcal{A}} \displaystyle\int_0^1 |\boldsymbol{r}'(t)|\,dt$. We would like to study the minimization problem $\inf\limits_{\boldsymbol{r}\in\mathcal{A}} I(\boldsymbol{r})$.

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 $\widehat{\boldsymbol{r}} = \widehat{x}\mathbf{i} + \widehat{y}\mathbf{j}$, we introduce a family of curves $\boldsymbol{r}(t;\epsilon) = x(t;\epsilon)\mathbf{i} + y(t;\epsilon)\mathbf{j} \in \mathcal{A}$, 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 $\epsilon$.

Denote $\delta\boldsymbol{r}(t) = \dfrac{d}{d\epsilon}\Big|_{\epsilon=0} \boldsymbol{r}(t;\epsilon) = \delta x(t)\mathbf{i} + \delta y(t)\mathbf{j}$. 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

$$\int_0^1 \big[(\delta x)\dot{\widehat{y}} + \widehat{x}(\dot{\delta y}) - (\delta y)\dot{\widehat{x}} - \widehat{y}(\dot{\delta x})\big]\,dt = 0\,. \tag{4.13}$$

For each possible minimizer $\widehat{\boldsymbol{r}}$, the relation above induces a linear vector space

$$\mathscr{N}_{\widehat{\boldsymbol{r}}} = \Big\{ \delta\boldsymbol{r} = \delta x\mathbf{i} + \delta y\mathbf{j} \in \mathscr{D}^1([0,1];\mathbb{R}^2) \,\Big|\, \int_0^1 \big[\widehat{x}(\dot{\delta y}) - \widehat{y}(\dot{\delta x})\big]\,dt = 0 \Big\}\,.$$

Now we look for a minimizer $\widehat{\boldsymbol{r}} \in \mathscr{C}^2([0,1];\mathbb{R}^2)$. We note that Remark 4.3 implies that if we are able to find a minimizer in $\mathscr{C}^2([0,1];\mathbb{R}^2)$ (thus a $\mathscr{C}^1$-minimizer), it must also be a minimizer in $\mathscr{D}^1([0,1];\mathbb{R}^2)$. Since $\widehat{\boldsymbol{r}} \in \mathscr{C}^2([0,1];\mathbb{R}^2)$ 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

$$\int_0^1 \frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\cdot(\delta\boldsymbol{r})'(t)\,dt = 0 \qquad \forall\,\delta\boldsymbol{r} \in \mathscr{N}_{\widehat{\boldsymbol{r}}}\,, \tag{4.14}$$

and Lemma 4.10 implies that there exists $\lambda_1, \lambda_2, \mu_1, \mu_2 \in \mathbb{R}$ such that

$$\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} = \big(\lambda_1\widehat{y}(t) + \mu_1\big)\mathbf{i} + \big(\lambda_2\widehat{x}(t) + \mu_2\big)\mathbf{j}\,. \tag{4.15}$$

Since $\widehat{\boldsymbol{r}} = (\widehat{x}, \widehat{y}) \in \mathscr{C}^2([0,1];\mathbb{R}^2)$, we differentiate the equation above and obtain that

$$\Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' = \lambda_1 \widehat{y}'(t)\mathbf{i} + \lambda_2 \widehat{x}'(t)\mathbf{j}\,.$$

Therefore, taking the inner product of the equation above with the unit tangent vector $\dfrac{\widehat{\boldsymbol{r}}'}{|\widehat{\boldsymbol{r}}'|}$, we find that

$$0 = \Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)\cdot\Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' = \big(\lambda_1 \widehat{y}'(t)\mathbf{i} + \lambda_2 \widehat{x}'(t)\mathbf{j}\big)\cdot\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} = (\lambda_2+\lambda_1)\frac{\widehat{x}'(t)\widehat{y}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} \quad \forall\, t\in[0,1]$$

which implies that $\lambda_2 = -\lambda_1 = \lambda$ (for otherwise $\widehat{x}'\widehat{y}' = 0$ which shows that the trajectory is a straight line); thus

$$\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} = (-\lambda\widehat{y}(t) + \mu_1)\mathbf{i} + \big(\lambda\widehat{x}(t) + \mu_2\big)\mathbf{j}\,.$$

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 $\widetilde{\boldsymbol{r}}$ denoting the vector

$$\widetilde{\boldsymbol{r}}(t) = \widetilde{x}(t)\mathbf{i} + \widetilde{y}(t)\mathbf{j} \equiv \Big(\widehat{x}(t) + \frac{\mu_2}{\lambda}\Big)\mathbf{i} + \Big(\widehat{y}(t) - \frac{\mu_1}{\lambda}\Big)\mathbf{j}\,,$$

we have

$$\frac{\widetilde{\boldsymbol{r}}'(t)}{|\widetilde{\boldsymbol{r}}'(t)|} = -\lambda\widetilde{y}(t)\mathbf{i} + \lambda\widetilde{x}(t)\mathbf{j}\,.$$

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.

**Remark 4.21.** The Dido problem can also be solved using the method of Lagrange multipliers. Let

$$F(x,y,\lambda) = \int_0^1 \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\,dt - \frac{\lambda}{2}\Big(\int_0^1 \big[x(t)\dot{y}(t) - y(t)\dot{x}(t)\big]\,dt - 2\mathrm{A}\Big)$$

and $\mathcal{A} = \big\{(x,y,\lambda)\,\big|\,x,y\in\mathscr{D}^1([0,1];\mathbb{R}), x(0)=x(1), y(0)=y(1), \lambda\in\mathbb{R}\big\}$. Define

$$L(t,x,y,p,q) = \sqrt{p^2+q^2} - \frac{\lambda}{2}(xq - yp - 2\mathrm{A})\,.$$

Then 3 in Remark 4.18 and the method of the Lagrange multipliers (which we did not prove in the case we need) imply that if $F$ attains its extremum at $(\widehat{x},\widehat{y},\lambda)\in\mathcal{A}$, then $\widehat{x},\widehat{y}$ satisfy the Euler-Lagrange equation

$$\frac{d}{dt}L_p(t,\widehat{x},\widehat{y},\dot{\widehat{x}},\dot{\widehat{y}}) = L_x(t,\widehat{x},\widehat{y},\dot{\widehat{x}},\dot{\widehat{y}})\,,$$
$$\frac{d}{dt}L_q(t,\widehat{x},\widehat{y},\dot{\widehat{x}},\dot{\widehat{y}}) = L_y(t,\widehat{x},\widehat{y},\dot{\widehat{x}},\dot{\widehat{y}})$$

and

$$0 = \frac{\partial F}{\partial \lambda}(\widehat{x}, \widehat{y}, \lambda) = \int_0^1 \big[x(t)\dot{y}(t) - y(t)\dot{x}(t)\big]\,dt - 2\mathrm{A}\,.$$

Define $\widehat{r}(t) = \widehat{x}(t)\mathbf{i} + \widehat{y}(t)\mathbf{j}$. By the fact that $|\dot{\boldsymbol{r}}(t)| = \sqrt{\dot{\widehat{x}}(t)^2 + \dot{\widehat{y}}(t)^2}$, we find that $\widehat{x}, \widehat{y}$ satisfy

$$\frac{d}{dt}\Big(\frac{\dot{\widehat{x}}(t)}{|\dot{\boldsymbol{r}}(t)|} + \frac{\lambda}{2}\,\widehat{y}(t)\Big) = -\frac{\lambda}{2}\dot{\widehat{y}}(t) \qquad \text{and} \qquad \frac{d}{dt}\Big(\frac{\dot{\widehat{y}}(t)}{|\dot{\boldsymbol{r}}(t)|} - \frac{\lambda}{2}\,\widehat{x}(t)\Big) = \frac{\lambda}{2}\dot{\widehat{x}}(t)\,.$$

The above equations shows that

$$\frac{d}{dt}\frac{\dot{\widehat{x}}(t)}{|\dot{\widehat{\boldsymbol{r}}}(t)|} = 2\lambda\dot{\widehat{y}}(t) \qquad \text{and} \qquad \frac{d}{dt}\frac{\dot{\widehat{y}}(t)}{|\dot{\widehat{\boldsymbol{r}}}(t)|} = -2\lambda\dot{\widehat{x}}(t)\,;$$

thus there exist constants $\mu_1, \mu_2$ such that

$$\frac{\dot{\widehat{x}}(t)}{|\dot{\widehat{\boldsymbol{r}}}(t)|} = -\lambda\widehat{y}(t) + \mu_1 \qquad \text{and} \qquad \frac{\dot{\widehat{y}}(t)}{|\dot{\widehat{\boldsymbol{r}}}(t)|} = \lambda\widehat{x}(t) + \mu_2\,.$$

The identities above can be combined and expressed as (4.15), and the rest of the argument of showing that $\boldsymbol{r}$ is the parametrization of a circle is identical.

We note that the method of the Lagrange multipliers only provides possible minimizers satisfying certain condition. Therefore, there might be minimizers that do not satisfy this condition, so one also has to show that there is no such minimizers. Indeed, a function violating this particular condition must represent a straight line, but we will not proceed further here.

**Example 4.22.** Consider finding the shortest path on the unit sphere connecting two points $A_0$ and $B_0$ (on the same sphere). In other words, we are interested in the minimization problem

$$\inf_{r\in\mathcal{A}}\int_0^1 \big|\boldsymbol{r}'(t)\big|\,dt\,,$$

where $\mathcal{A} = \big\{\boldsymbol{r} \in \mathscr{D}^1([0,1];\mathbb{R}^3)\,\big|\,\boldsymbol{r}(0) = A_0, \boldsymbol{r}(1) = B_0\,, |\boldsymbol{r}(t)| = 1 \text{ for all } t \in [0,1]\big\}$.

Similar to the previous example, we introduce a family of curves $\boldsymbol{r}(t;\epsilon)$, 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. $\boldsymbol{r}(0;\epsilon) = A_0$; 3. $\boldsymbol{r}(1;\epsilon) = B_0$; 4. $\boldsymbol{r}$ is also differentiable in $\epsilon$,

where $\widehat{\boldsymbol{r}}$ gives the shortest path connecting $A_0$ and $B_0$. Since $\widehat{\boldsymbol{r}} \in \mathcal{A}$, $\widehat{\boldsymbol{r}}'(t)\cdot\widehat{\boldsymbol{r}}(t) = 0$ whenever $\widehat{\boldsymbol{r}}'(t)$ exists. Therefore, we can assume that

$$\widehat{\boldsymbol{r}}(t), \widehat{\boldsymbol{r}}'(t), (\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}})(t) \text{ are linearly independent if } \widehat{\boldsymbol{r}}'(t) \neq \mathbf{0}\,. \tag{4.16}$$

Denote $\delta\boldsymbol{r}(t) = \frac{d}{d\epsilon}\Big|_{\epsilon=0}\boldsymbol{r}(t;\epsilon)$. Then the fact that $\boldsymbol{r} \in \mathcal{A}$ again implies that $\delta\boldsymbol{r}\cdot\widehat{\boldsymbol{r}} = 0$; thus we shall introduce $\mathscr{N}_{\widehat{\boldsymbol{r}}}$ as

$$\mathscr{N}_{\widehat{\boldsymbol{r}}} = \Big\{\delta\boldsymbol{r} \in \mathscr{C}^1([0,1];\mathbb{R}^3)\,\Big|\,\widehat{\boldsymbol{r}}(t)\cdot\delta\boldsymbol{r}(t) = 0 \text{ for all } t \in [0,1]\Big\}\,.$$

In particular, with (4.16) we conclude that

$$\mathscr{N}_{\widehat{\boldsymbol{r}}} = \operatorname{span}\big(\widehat{\boldsymbol{r}}', \widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}}\big) = \big\{a\widehat{\boldsymbol{r}}' + b(\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}}) \,\big|\, a, b \in \mathbb{R}\big\}\,. \tag{4.17}$$

Now suppose that $\widehat{\boldsymbol{r}} \in \mathscr{C}^2([0,1];\mathbb{R}^3)$. Similar to the derivation of (4.14), we obtain that

$$0 = \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int_0^1 \big|\boldsymbol{r}'(t;\epsilon)\big|\,dt = \int_0^1 \frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} \cdot (\delta\boldsymbol{r})'(t)\,dt \qquad \forall\,\delta\boldsymbol{r} \in \mathscr{N}_{\widehat{\boldsymbol{r}}}\,,$$

and integrating by parts further shows that for $\delta\boldsymbol{r} \in \mathscr{N}_{\widehat{\boldsymbol{r}}}$,

$$\begin{aligned} 0 &= \int_0^1 \frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} \cdot (\delta\boldsymbol{r})'(t)\,dt = \frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|} \cdot (\delta\boldsymbol{r})(t)\Big|_{t=0}^{t=1} - \int_0^1 \Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' \cdot (\delta\boldsymbol{r})(t)\,dt \\ &= -\int_0^1 \Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' \cdot (\delta\boldsymbol{r})(t)\,dt\,, \end{aligned}$$

where we have use the fact that $(\delta\boldsymbol{r})(0) = (\delta\boldsymbol{r})(1) = \mathbf{0}$ to eliminate the boundary contributions. Since $\Big(\dfrac{\widehat{\boldsymbol{r}}'}{|\widehat{\boldsymbol{r}}'|}\Big)' \cdot \widehat{\boldsymbol{r}}' = 0$, we conclude from (4.17) that

$$\int_0^1 b(t)\Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' \cdot (\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}})(t)\,dt = 0 \qquad \forall\, b \in \mathscr{C}([0,1];\mathbb{R})$$

which, by Lemma 4.6, shows that

$$\Big(\frac{\widehat{\boldsymbol{r}}'(t)}{|\widehat{\boldsymbol{r}}'(t)|}\Big)' \cdot (\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}})(t) = 0 \qquad \forall\, t \in [0,1]\,.$$

By the fact that $\widehat{\boldsymbol{r}}' \cdot (\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}}) = 0$, the identity above further shows that

$$\widehat{\boldsymbol{r}}''(t) \cdot (\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}})(t) = 0 \qquad \forall\, t \in [0,1]\,. \tag{4.18}$$

Now suppose that the parametrization of the shortest path satisfies that $|\widehat{\boldsymbol{r}}'(t)| = \text{constant}$; that is, the motion along the shortest path has constant speed. Then $\widehat{\boldsymbol{r}}'(t) \cdot \widehat{\boldsymbol{r}}''(t) = 0$ for all $t \in [0,1]$; thus (4.16) implies that

$$\widehat{\boldsymbol{r}}'' = c\widehat{\boldsymbol{r}} + d(\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}}) \quad \text{for some functions } c \text{ and } d \text{ of } t\,.$$

Identity (4.18) further shows that $d = 0$; thus $\widehat{\boldsymbol{r}}'' = c\widehat{\boldsymbol{r}}$ so that

$$(\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}})' = \widehat{\boldsymbol{r}}'' \times \widehat{\boldsymbol{r}} = c\widehat{\boldsymbol{r}} \times \widehat{\boldsymbol{r}} = \mathbf{0}\,.$$

As a consequence, $\widehat{\boldsymbol{r}}' \times \widehat{\boldsymbol{r}}$ is a constant vector $\boldsymbol{c}$; thus (4.16) implies that $\widehat{\boldsymbol{r}} \cdot \boldsymbol{c} = 0$. Therefore, the trajectory lies on a plane passing through the origin which shows that the shortest path connecting two points on the sphere must be part of a great circle.

# Appendix A

# Vector Calculus

## A.1 Vector Fields

**Definition A.1.** A (two-dimensional) vector field over a plane region $R$ is a vector-valued function $\boldsymbol{F}$ that assigns a vector $\boldsymbol{F}(x,y) \in \mathbb{R}^2$ to each point $(x,y)$ in $R$. A (three-dimensional) vector field over a solid region $Q$ is a vector-valued function $\boldsymbol{F}$ that assigns a vector $\boldsymbol{F}(x,y,z) \in \mathbb{R}^3$ to each point $(x,y,z)$ in $Q$.

In general, an $n$-dimensional vector field over a region $D \subseteq \mathbb{R}^n$ is a vector-valued function $\boldsymbol{F}$ that assigns a vector $\boldsymbol{F}(x_1,x_2,\cdots,x_n) \in \mathbb{R}^n$ to each point $\boldsymbol{x} = (x_1,x_2,\cdots,x_n)$ in $D$.

**Definition A.2** (旋度)**.** Let $Q$ be an open region in space, and $\boldsymbol{F}: Q \to \mathbb{R}^3$ be a vector field given by $\boldsymbol{F}(x,y,z) = M(x,y,z)\mathbf{i} + N(x,y,z)\mathbf{j} + P(x,y,z)\mathbf{k}$. The curl of $\boldsymbol{F}$, also called the vorticity of $\boldsymbol{F}$, is a vector field given by

$$\text{curl}\boldsymbol{F} = \Big(\frac{\partial P}{\partial y} - \frac{\partial N}{\partial z}\Big)\mathbf{i} - \Big(\frac{\partial P}{\partial x} - \frac{\partial M}{\partial z}\Big)\mathbf{j} + \Big(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y}\Big)\mathbf{k}\,.$$

If $\text{curl}\boldsymbol{F} = \mathbf{0}$, then $\boldsymbol{F}$ is said to be ***irrotational***.

Symbolically, the curl of $\boldsymbol{F}$ is given by

$$\text{curl}\boldsymbol{F} = \nabla \times \boldsymbol{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ M & N & P \end{vmatrix}\,.$$

**Remark A.3.** Let $\boldsymbol{F}$ be a three-dimensional vector field, and $F_i$ be the $i$-th component of $\boldsymbol{F}$; that is,

$$\boldsymbol{F} = F_1\mathbf{i} + F_2\mathbf{j} + F_3\mathbf{k} = \sum_{i=1}^{3} F_i\mathbf{e}_i\,.$$

Then using the permutation symbol $\varepsilon_{ijk}$, we have

$$(\text{curl}\boldsymbol{F})_i \equiv \text{the } i\text{-th component of curl}\boldsymbol{F} = \sum_{j,k=1}^{3} \varepsilon_{ijk}\frac{\partial F_k}{\partial x_j}\,. \tag{A.1}$$

**Remark A.4.** Let $\boldsymbol{F}$ be a two dimensional vector field given by $\boldsymbol{F}(x,y) = M(x,y)\mathbf{i} + N(x,y)\mathbf{j}$. We can also define the curl of $\boldsymbol{F}$ by treating $\boldsymbol{F}$ as a three-dimensional vector field

$$\widetilde{\boldsymbol{F}}(x,y,z) = M(x,y)\mathbf{i} + N(x,y)\mathbf{j} + 0\mathbf{k}$$

(which is a three-dimensional vector field independent of $z$) and define $\operatorname{curl}\boldsymbol{F}$ as the third component of $\operatorname{curl}\widetilde{\boldsymbol{F}}$ (for the first two components of $\operatorname{curl}\widetilde{\boldsymbol{F}}$ are zero). Therefore, the curl of a two dimensional vector field $\boldsymbol{F} = M\mathbf{i} + N\mathbf{j}$ is a scalar function given by

$$\operatorname{curl}\boldsymbol{F} = \frac{\partial N}{\partial x} - \frac{\partial M}{\partial y}\,.$$

Moreover, by defining the differential operator $\nabla^\perp = \left(-\frac{\partial}{\partial y}, \frac{\partial}{\partial x}\right)$ on plane we have the symbolic representation

$$\operatorname{curl}\boldsymbol{F} = \nabla^\perp \cdot \boldsymbol{F}\,.$$

**Definition A.5 (散度).** Let $R$ be an open region in the plane, and $\boldsymbol{F} : R \to \mathbb{R}^2$ be a vector field given by $\boldsymbol{F}(x,y) = M(x,y)\mathbf{i} + N(x,y)\mathbf{j}$. The divergence of $\boldsymbol{F}$ is a scalar function given by

$$\operatorname{div}\boldsymbol{F} = \frac{\partial M}{\partial x} + \frac{\partial N}{\partial y}\,.$$

Let $Q$ be an open region in space, and $\boldsymbol{F} : Q \to \mathbb{R}^3$ be a vector field given by $\boldsymbol{F}(x,y,z) = M(x,y,z)\mathbf{i} + N(x,y,z)\mathbf{j} + P(x,y,z)\mathbf{k}$. The divergence of $\boldsymbol{F}$ is a scalar function given by

$$\operatorname{div}\boldsymbol{F} = \frac{\partial M}{\partial x} + \frac{\partial N}{\partial y} + \frac{\partial P}{\partial z}\,.$$

In general, if $D$ is an open region in $\mathbb{R}^n$ and $\boldsymbol{F} : D \to \mathbb{R}^n$ be a vector field given by $\boldsymbol{F}(\boldsymbol{x}) = \big(F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), \cdots, F_n(\boldsymbol{x})\big)$, the divergence of $\boldsymbol{F}$ is a scalar function given by

$$\operatorname{div}\boldsymbol{F} = \sum_{i=1}^{n} \frac{\partial F_i}{\partial x_i}\,.$$

**Definition A.6.** A vector field $\boldsymbol{u} : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is called *solenoidal* or *divergence-free* if $\operatorname{div}\boldsymbol{u} = 0$ in $\Omega$.

## A.2 The Line Integrals

### A.2.1 Curves and parametric equations

**Definition A.7.** A subset $C$ in the plane (or space) is called a ***curve*** if $C$ is the image of an interval $I \subseteq \mathbb{R}$ under a continuous vector-valued function $\boldsymbol{r}$. The continuous function $\boldsymbol{r} : I \to \mathbb{R}^2$ (or $\mathbb{R}^3$) is called a ***parametrization*** of the curve, and the equation

$$(x,y) = \boldsymbol{r}(t)\,,\ t \in I \qquad \big(\text{or } (x,y,z) = \boldsymbol{r}(t),\ \ t \in I\big)$$

is called a ***parametric equation*** of the curve. A curve $C$ is called a ***plane curve*** if it is a subset in the plane.

Since a plane can be treated as a subset of space, in the following we always assume that the curve under discussion is a curve in space (so that the parametrization of the curve is given by $\boldsymbol{r} : I \to \mathbb{R}^3$).

**Definition A.8.** A curve $C$ is called ***simple*** if it has an injective parametrization; that is, there exists $\boldsymbol{r} : I \to \mathbb{R}^3$ such that $\boldsymbol{r}(I) = C$ and $\boldsymbol{r}(x) = \boldsymbol{r}(y)$ implies that $x = y$. A curve $C$ with parametrization $\boldsymbol{r} : I \to \mathbb{R}^3$ is called ***closed*** if $I = [a, b]$ for some closed interval $[a, b] \subseteq \mathbb{R}$ and $\boldsymbol{r}(a) = \boldsymbol{r}(b)$. A ***simple closed*** curve $C$ is a closed curve with parametrization $\boldsymbol{r} : [a, b] \to \mathbb{R}^3$ such that $\boldsymbol{r}$ is one-to-one on $(a, b)$. A ***smooth*** curve $C$ is a curve with continuously differentiable parametrization $\boldsymbol{r} : I \to \mathbb{R}^3$ such that $\boldsymbol{r}'(t) \neq \mathbf{0}$ for all $t \in I$.

**Example A.9.** The parabola $y = x^2 + 2$ on the plane is a simple smooth plane curve since $\boldsymbol{r} : \mathbb{R} \to \mathbb{R}^2$ given by $r(t) = t\mathbf{i} + (r^2 + 2)\mathbf{j}$ is an injective differentiable parametrization of this parabola. We note that $\tilde{\boldsymbol{r}} : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}^2$ given by $\tilde{\boldsymbol{r}}(t) = \tan t\mathbf{i} + (\sec^2 t + 1)\mathbf{j}$ is also an injective smooth parametrization of this parabola. In general, a curve usually has infinitely many parameterizations.

**Example A.10.** Let $I \subseteq \mathbb{R}$ be an interval, and $\boldsymbol{r} : I \to \mathbb{R}^2$ be defined by $\boldsymbol{r}(t) = \cos t\mathbf{i} + \sin t\mathbf{j}$. Since $\boldsymbol{r}$ is continuous and the co-domain is $\mathbb{R}^2$, the image of $I$ under $\boldsymbol{r}$, denoted by $C$, is a plane curve. We note that $C$ is part of the unit circle centered at the origin. Moreover, $C$ is a smooth curve since $\boldsymbol{r}'(t) \neq \mathbf{0}$ for all $t \in I$.

1. If $I = [a, b]$ and $|b - a| < 2\pi$, then $C$ is a simple curve.
2. If $I = [0, 2\pi]$, then $C$ is not a simple curve. However, $C$ a simple closed curve.

**Example A.11.** Let $\boldsymbol{r} : [0, 2\pi] \to \mathbb{R}^2$ be defined by $\boldsymbol{r}(t) = \sin t\mathbf{i} + \sin t \cos t\mathbf{j}$. The image $\boldsymbol{r}([0, 2\pi])$ is a curve called figure eight.

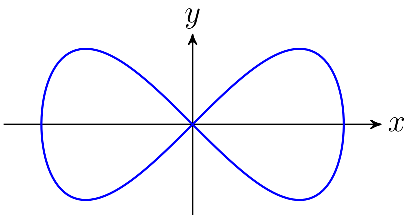


Figure A.1: Figure eight

**Example A.12.** Let $\boldsymbol{r} : \mathbb{R} \to \mathbb{R}^3$ be defined by $\boldsymbol{r}(t) = \cos t\mathbf{i} + \sin t\mathbf{j} + t\mathbf{k}$. Then the image $\boldsymbol{r}(\mathbb{R})$ is a simple smooth space curve. This curve is called a helix.

In the following, when a parametrization $\boldsymbol{r} : I \to \mathbb{R}^3$ of curves $C$ is mentioned, we always assume that "there is no overlap"; that is, there are no intervals $[a, b], [c, d] \subseteq I$ satisfying that $\boldsymbol{r}([a, b]) = \boldsymbol{r}([c, d])$. If in addition

1. $C$ is a simple curve, then $\boldsymbol{r}$ is injective, or

2. $C$ is closed, then $I = [a, b]$ and $\boldsymbol{r}(a) = \boldsymbol{r}(b)$, or
3. $C$ is simple closed, then $I = [a, b]$ and $\boldsymbol{r}$ is injective on $[a, b)$ and $\boldsymbol{r}(a) = \boldsymbol{r}(b)$.
4. $C$ is smooth, then $\boldsymbol{r}$ is continuously differentiable and $\boldsymbol{r}'(t) \neq \mathbf{0}$ for all $t \in I$.

**Theorem A.13.** *Let $C$ be a smooth curve parameterized by $\boldsymbol{r} : [a, b] \to \mathbb{R}^3$. Then the length of $C$ is given by*

$$\ell(C) = \int_a^b \|\boldsymbol{r}'(t)\|\, dt\,.$$

### A.2.2 Line integrals of scalar functions

In this section, we are concerned with the "integral" of a real-valued function $f$ defined on a curve $C$. It is motivated by calculating the total mass of a wire lying along a curve in space when the density of the wire at each point is given, or to find the work done by a given (variable) force acting along such a curve. We begin with the following

**Definition A.14.** Let $C$ be a curve in space. A ***partition*** of $C$ is a collection of curves $\{C_1, C_2, \cdots, C_n\}$ satisfying

1. $C = \bigcup\limits_{i=1}^{n} C_i$ (so that $C_i \subseteq C$);
2. If $i \neq j$, then $C_i \cap C_j$ contains at most two points.

Let $\mathcal{P} = \{C_1, C_2, \cdots, C_n\}$ be a partition of $C$. The ***norm*** of $\mathcal{P}$, denoted by $\|\mathcal{P}\|$, is the number

$$\|\mathcal{P}\| = \max\{\ell(C_1), \ell(C_2), \cdots, \ell(C_n)\}\,,$$

where $\ell(C_j)$ denotes the length of curve $C_j$. If $f : C \to \mathbb{R}$ is a real-valued function defined on $C$, a ***Riemann sum*** of $f$ for partition $\mathcal{P}$ is a sum of the form

$$\sum_{i=1}^{n} f(q_i)\ell(C_i)\,,$$

where $\{q_1, q_2, \cdots, q_n\}$ is a collection of points on $C$ satisfying $q_j \in C_j$ for all $1 \leqslant j \leqslant n$.

We note that **in order to define the norm of partitions, it is required that every sub-curve $C_j$ of $C$ has length**. This kind of curves is called ***rectifiable*** curves, and we can only consider line integrals along rectifiable curves. In particular, a piecewise continuously differentiable curve is rectifiable.

Similar to the Riemann integral, we consider the limit of Riemann sums

$$\lim_{\|\mathcal{P}\|\to 0} \sum_{j=1}^{n} f(q_j)\ell(C_j)\,.$$

The line integral of $f$ along $C$ is the limit above if the limit indeed exists. The precise definition of the limit is given below.

**Definition A.15.** Let $C$ be a rectifiable curve, and $f : C \to \mathbb{R}$ be a scalar function. The line integral of $f$ along $C$ is a real number $L$ such that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathcal{P} = \{C_1, C_2, \cdots, C_n\}$ is a partition of $C$ satisfying $\|\mathcal{P}\| < \delta$, then any Riemann sum of $f$ for $\mathcal{P}$ belongs to the interval $(L - \varepsilon, L + \varepsilon)$.

Whenever such an $L$ exists, it must be unique, and the number $L$ is denoted by $\displaystyle\int_C f\, ds$ $\big($and when $C$ is a closed curve, we use $\displaystyle\oint_C f\, ds$ to emphasize that the curve is closed$\big)$.

**Theorem A.16.** *Let $C$ be a (piecewise) smooth curve with (piecewise) continuously differentiable injective parametrization $\boldsymbol{r} : [a, b] \to \mathbb{R}^3$, and $f : C \to \mathbb{R}$ be a continuous function. Then the line integral of $f$ along $C$ exists and is given by*

$$\int_a^b (f \circ \boldsymbol{r})(t) \|\boldsymbol{r}'(t)\|\, dt\,.$$

**Example A.17.** Evaluate $\displaystyle\int_C (x^2 - y + 3z)\, ds$, where $C$ is the line segment connecting the points $(0, 0, 0)$ and $(1, 2, 1)$.

First we note that the line segment can be parameterized by

$$\boldsymbol{r}(t) = (1 - t)(0, 0, 0) + t(1, 2, 1) = t\mathbf{i} + 2t\mathbf{j} + t\mathbf{k} \qquad t \in [0, 1]\,.$$

Therefore, Theorem A.16 implies that

$$\int_C (x^2 - y + 3z)\, ds = \int_0^1 (t^2 - 2t + 3t)\|\mathbf{i} + 2\mathbf{j} + \mathbf{k}\|\, dt = \sqrt{6}\int_0^1 (t^2 + t)\, dt = \frac{5\sqrt{6}}{6}\,.$$

**Example A.18.** Evaluate $\displaystyle\int_C x\, ds$, where $C$ is the piecewise smooth curve starting from $(0, 0)$ to $(1, 1)$ along $y = x^2$ then from $(1, 1)$ to $(0, 0)$ along $y = x$.

Let $C_1$ be the piece of the curve connecting $(0, 0)$ and $(1, 1)$ along $y = x^2$, and $C_2$ be the piece of the curve connecting $(1, 1)$ and $(0, 0)$ along $y = x$. Then $C_1$ and $C_2$ can be parameterized by

$$\boldsymbol{r}_1(t) = t\mathbf{i} + t^2\mathbf{j} \quad t \in [0, 1] \qquad \text{and} \qquad \boldsymbol{r}_2(t) = t\mathbf{i} + t\mathbf{j} \quad t \in [0, 1]\,,$$

respectively. Since $C = C_1 \cup C_2$ and $C_1 \cap C_2$ has only two points,

$$\begin{aligned}\int_C x\, ds &= \int_{C_1} x\, ds + \int_{C_2} x\, ds = \int_0^1 t\|\mathbf{i} + 2t\mathbf{j}\|\, dt + \int_0^1 t\|\mathbf{i} + \mathbf{j}\|\, dt \\ &= \int_0^1 \big[t\sqrt{1 + 4t^2} + \sqrt{2}t\big]\, dt \\ &= \Big[\frac{1}{12}(1 + 4t^2)^{\frac{3}{2}} + \frac{\sqrt{2}t^2}{2}\Big]\Big|_{t=0}^{t=1} = \frac{1}{12}(5\sqrt{5} - 1) + \frac{\sqrt{2}}{2}\,.\end{aligned}$$

**Example A.19.** Let $C$ be the upper half part of the circle centered at the origin with radius $R > 0$ in the $xy$-plane. Evaluate the line integral $\displaystyle\int_C y\, ds$.

First, we parameterize $C$ by

$$\boldsymbol{r}(t) = R\cos t\,\mathbf{i} + R\sin t\,\mathbf{j} \qquad t \in [0, \pi]\,.$$

Then

$$\int_C y\,ds = \int_0^\pi R\sin t \big\| -R\sin t\,\mathbf{i} + R\cos t\,\mathbf{j}\big\|_{\mathbb{R}^2} dt = \int_0^\pi R^2 \sin t\,dt = 2R^2\,.$$

**Example A.20.** Find the mass of a wire lying along the first octant part of the curve of intersection of the elliptic paraboloid $z = 2 - x^2 - 2y^2$ and the parabolic cylinder $z = x^2$ between $(0, 1, 0)$ and $(1, 0, 1)$ if the density of the wire at position $(x, y, z)$ is $\varrho(x, y, z) = xy$.

Note that we can parameterize the curve $C$ by

$$\boldsymbol{r}(t) = t\mathbf{i} + \sqrt{1-t^2}\,\mathbf{j} + t^2\mathbf{k} \qquad t \in [0, 1]\,.$$

Therefore, the mass of the curve can be computed by

$$\begin{aligned}
\int_C \varrho\,ds &= \int_0^1 t\sqrt{1-t^2}\,\Big\|\mathbf{i} + \frac{-t}{\sqrt{1-t^2}}\mathbf{j} + 2t\mathbf{k}\Big\|_{\mathbb{R}^3} dt = \int_0^1 t\sqrt{1-t^2}\,\frac{\sqrt{1-t^2+t^2+4t^2(1-t^2)}}{\sqrt{1-t^2}}\,dt \\
&= \int_0^1 t\sqrt{2-(1-2t^2)^2}\,dt = \frac{1}{4}\int_{-1}^1 \sqrt{2-u^2}du = \frac{1}{4}\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} 2\cos^2\theta\,d\theta \\
&= \frac{1}{4}\Big[\theta + \frac{\sin(2\theta)}{2}\Big]\Big|_{\theta=-\frac{\pi}{4}}^{\theta=\frac{\pi}{4}} = \frac{\pi}{8} + \frac{1}{4}\,.
\end{aligned}$$

### A.2.3 Line integrals of vector fields

**Definition A.21.** An oriented curve is a curve on which a consistent tangent direction $\mathbf{T}$ is defined. In other words, an oriented curve is a (piecewise) smooth curve with a given parametrization $\boldsymbol{r} : I \to \mathbb{R}^3$ so that $\mathbf{T} \circ \boldsymbol{r} = \dfrac{\boldsymbol{r}'}{\|\boldsymbol{r}'\|}$ is defined (almost everywhere).

**Definition A.22.** Let $\boldsymbol{F}$ be a continuous vector field defined on a smooth oriented curve $C$ parameterized by $\boldsymbol{r}(t)$ for $t \in [a, b]$. The line integral of $F$ along $C$ is given by

$$\int_C \boldsymbol{F}\cdot\mathbf{T}\,ds\,.$$

**Remark A.23.** Note that since $\mathbf{T} \circ \boldsymbol{r} = \dfrac{\boldsymbol{r}'}{\|\boldsymbol{r}'\|}$, by Theorem A.16 we have

$$\int_C \boldsymbol{F}\cdot\mathbf{T}\,ds = \int_a^b (\boldsymbol{F}\circ\boldsymbol{r})(t)\cdot\frac{\boldsymbol{r}'(t)}{\|\boldsymbol{r}'(t)\|}\|\boldsymbol{r}'(t)\|\,dt = \int_a^b (\boldsymbol{F}\circ\boldsymbol{r})(t)\cdot\boldsymbol{r}'(t)\,dt\,.$$

Since $\boldsymbol{r}'(t)\,dt = d\boldsymbol{r}(t)$, sometimes we also use $\displaystyle\int_C \boldsymbol{F}\cdot d\boldsymbol{r}$ to denote the line integral of $\boldsymbol{F}$ along the oriented curve $C$ parameterized by $\boldsymbol{r}$.

**Remark A.24.** Given an oriented curve $C$ and $\boldsymbol{F} : C \to \mathbb{R}^3$, we sometimes use $\displaystyle\int_{-C} \boldsymbol{F}\cdot d\boldsymbol{r}$ to denote the line integral $\displaystyle\int_C \boldsymbol{F}\cdot(-\mathbf{T})\,ds$, where $-\mathbf{T}$ is the tangent direction opposite to the orientation of $C$.

**Example A.25.** Find the work done by the force field

$$\boldsymbol{F}(x, y, z) = -\frac{1}{2}x\mathbf{i} - \frac{1}{2}y\mathbf{j} + \frac{1}{4}\mathbf{k}$$

on a particle as it moves along the helix parameterized by

$$\boldsymbol{r}(t) = \cos t\,\mathbf{i} + \sin t\,\mathbf{j} + t\mathbf{k}$$

from the point $(1, 0, 0)$ to the point $(-1, 0, 3\pi)$. Note that such a helix is parameterized by $\boldsymbol{r}(t)$ with $t \in [0, 3\pi]$. Therefore,

$$\begin{aligned}\int_C \boldsymbol{F} \cdot d\boldsymbol{r} &= \int_0^{3\pi} \Big(-\frac{1}{2}\cos t\,\mathbf{i} - \frac{1}{2}\sin t\,\mathbf{j} + \frac{1}{4}\mathbf{k}\Big) \cdot \big(-\sin t\,\mathbf{i} + \cos t\,\mathbf{j} + \mathbf{k}\big)\,dt \\ &= \int_0^{3\pi} \Big(\frac{1}{2}\sin t\cos t - \frac{1}{2}\sin t\cos t + \frac{1}{4}\Big)\,dt = \frac{3\pi}{4}\,.\end{aligned}$$

**Example A.26.** Let $\boldsymbol{F}(x, y) = y^2\mathbf{i} + 2xy\mathbf{j}$. Evaluate the line integral $\int_C \boldsymbol{F} \cdot d\boldsymbol{r}$ from $(0, 0)$ to $(1, 1)$ along

1. the straight line $y = x$,
2. the curve $y = x^2$, and
3. the piecewise smooth path consisting of the straight line segments from $(0, 0)$ to $(0, 1)$ and from $(0, 1)$ to $(1, 1)$.

For the straight line case, we parameterize the path by $\boldsymbol{r}(t) = (t, t)$ for $t \in [0, 1]$. Then

$$\int_C \boldsymbol{F} \cdot d\boldsymbol{r} = \int_0^1 (t^2\mathbf{i} + 2t^2\mathbf{j}) \cdot (\mathbf{i} + \mathbf{j})dt = \int_0^1 3t^2 dt = 1\,.$$

For the case of parabola, we parameterize the path by $\boldsymbol{r}(t) = (t, t^2)$ for $t \in [0, 1]$. Then

$$\int_C \boldsymbol{F} \cdot d\boldsymbol{r} = \int_0^1 (t^4\mathbf{i} + 2t^3\mathbf{j}) \cdot (\mathbf{i} + 2t\mathbf{j})dt = \int_0^1 5t^4 dt = 1\,.$$

For the piecewise linear case, we let $C_1$ denote the line segment joining $(0, 0)$ and $(0, 1)$, and let $C_2$ denote the line segment joining $(0, 1)$ and $(1, 1)$. Note that we can parameterize $C_1$ and $C_2$ by

$$\boldsymbol{r}_1(t) = t\mathbf{j} \quad t \in [0, 1] \quad \text{and} \quad \boldsymbol{r}_2(t) = t\mathbf{i} + \mathbf{j} \quad t \in [0, 1]\,,$$

respectively. Therefore,

$$\int_C \boldsymbol{F} \cdot d\boldsymbol{r} = \int_{C_1} \boldsymbol{F} \cdot d\boldsymbol{r} + \int_{C_2} \boldsymbol{F} \cdot d\boldsymbol{r} = \int_0^1 t^2\mathbf{i} \cdot \mathbf{j}\,dt + \int_0^1 (\mathbf{i} + 2t\mathbf{j}) \cdot \mathbf{i}\,dt = 1\,.$$

We note that in this example the line integrals of $\boldsymbol{F}$ along three different paths joining $(0, 0)$ and $(1, 1)$ are identical.

**Example A.27.** Let $\boldsymbol{F}(x,y) = y\mathbf{i} - x\mathbf{j}$. Evaluate the line integral $\int_C \boldsymbol{F}\cdot d\boldsymbol{r}$ from $(1,0)$ to $(0,-1)$ along

1. the straight line segment joining these points, and
2. three-quarters of the circle of unit radius centered at the origin and traversed counterclockwise.

For the first case, we parameterize the path by $\boldsymbol{r}(t) = (1-t,-t)$ for $t \in [0,1]$. Then

$$\int_C \boldsymbol{F}\cdot d\boldsymbol{r} = \int_0^1 \big[-t\mathbf{i} + (t-1)\mathbf{j}\big]\cdot(-\mathbf{i}-\mathbf{j})\,dt = \int_0^1 1\,dt = 1\,.$$

For the second case, we parameterize the path by $\boldsymbol{r}(t) = \cos t\,\mathbf{i} + \sin t\,\mathbf{j}$ for $t \in \big[0,\frac{3\pi}{2}\big]$. Then

$$\int_C \boldsymbol{F}\cdot d\boldsymbol{r} = \int_0^{\frac{3\pi}{2}} (\sin t\,\mathbf{i} - \cos t\,\mathbf{j})\cdot(-\sin t\,\mathbf{i} + \cos t\,\mathbf{j})\,dt = \int_0^{\frac{3\pi}{2}} (-1)\,dt = -\frac{3\pi}{2}\,.$$

We note that in this example the line integrals of $\boldsymbol{F}$ along different paths joining $(1,0)$ and $(0,-1)$ can be different.

## A.3 The Green Theorem

Let $R \subseteq \mathbb{R}^2$ be a region enclosed by a simply closed curve $C$ and $\boldsymbol{F} = M\mathbf{i} + N\mathbf{j}$ be a vector fields on (an open set containing) $R$, where $C$ is ***oriented counterclockwise*** so that

$C$ is traversed once so that the region $R$ always lies to the left.

The line integral of $\boldsymbol{F}$ along an oriented curve $C$ sometimes is written as

$$\int_C Mdx + Ndy$$

since symbolically we have $d\boldsymbol{r} = dx\mathbf{i} + dy\mathbf{j}$ so that

$$\boldsymbol{F}\cdot d\boldsymbol{r} = (M\mathbf{i} + N\mathbf{j})\cdot(dx\mathbf{i} + dy\mathbf{j}) = Mdx + Ndy\,.$$

The right-hand side of the identity above is called a ***differential form***.

**Theorem A.28** (Green's Theorem)**.** *Let $R$ be a plane region enclosed by a closed curve $C$ oriented counterclockwise; that is, $C$ is traversed once so that the region $R$ always lies to the left. If $M$ and $N$ have continuous first partial derivatives in an open region containing $R$, then*

$$\oint_C Mdx + Ndy = \iint_R \Big(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y}\Big)(x,y)\,dA\,. \tag{A.2}$$

**Remark A.29.** If $\boldsymbol{F}$ is a two-dimensional vector field given by $\boldsymbol{F} = M\mathbf{i} + N\mathbf{j}$, then under the assumption of Green's Theorem,

$$\oint_C \boldsymbol{F}\cdot\mathbf{T}\,ds = \iint_R (\text{curl}\boldsymbol{F})(x,y)\,dA\,.$$

This is sometimes called **Green's Theorem in Tangential Form**. Moreover, by treating $\boldsymbol{F}$ as a three-dimensional vector field, then under the assumption of Green's Theorem,

$$\oint_C \boldsymbol{F}\cdot d\boldsymbol{r} = \iint_R (\text{curl}\boldsymbol{F}\cdot\mathbf{k})(x,y)\,dA\,.$$

**Remark A.30.** Let $R$ be a region enclosed by a smooth simply closed curve $C$ with **outward-pointing** unit normal $\mathbf{N}$ on $C$, and $\boldsymbol{F}$ be a smooth vector field defined on an open region containing $R$. We are interested in $\displaystyle\int_C \boldsymbol{F}\cdot\mathbf{N}ds$, the line integral of $\boldsymbol{F}\cdot\mathbf{N}$ along $C$.

Suppose that $\boldsymbol{F} = M\mathbf{i} + N\mathbf{j}$, and $C$ is parameterized by $\boldsymbol{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j}$, $t \in [a,b]$, so that $C$ is oriented counterclockwise. Define $\boldsymbol{G} = -N\mathbf{i} + M\mathbf{j}$. Then Green's Theorem implies that

$$\oint_C -Ndx + Mdy = \oint_C \boldsymbol{G}\cdot d\boldsymbol{r} = \iint_R \text{curl}\boldsymbol{G}\,dA = \iint_R \big(M_x + N_y\big)\,dA = \iint_R \text{div}\boldsymbol{F}\,dA\,.$$

On the other hand, if $\boldsymbol{r}$ is a counterclockwise parametrization of $C$, then

$$\mathbf{N}(\boldsymbol{r}(t)) = \frac{y'(t)}{\|\boldsymbol{r}'(t)\|}\mathbf{i} - \frac{x'(t)}{\|\boldsymbol{r}'(t)\|}\mathbf{j} \qquad \forall\, t \in [a,b]\,;$$

thus

$$\begin{aligned}\oint_C \boldsymbol{F}\cdot\mathbf{N}\,ds &= \int_a^b (\boldsymbol{F}\cdot\mathbf{N})(\boldsymbol{r}(t))\|\boldsymbol{r}'(t)\|\,dt = \int_a^b \boldsymbol{F}(\boldsymbol{r}(t))\cdot\mathbf{N}(\boldsymbol{r}(t))\|\boldsymbol{r}'(t)\|\,dt\\ &= \int_a^b \big[M(x(t),y(t))\mathbf{i} + N(x(t),y(t))\mathbf{j}\big]\cdot\Big[\frac{y'(t)}{\|\boldsymbol{r}'(t)\|}\mathbf{i} - \frac{x'(t)}{\|\boldsymbol{r}'(t)\|}\mathbf{j}\Big]\|\boldsymbol{r}'(t)\|\,dt\\ &= \int_a^b \big[M(x(t),y(t))y'(t) - N(x(t),y(t))x'(t)\big]\,dt\\ &= \oint_C -N\,dx + M\,dy = \oint_C \boldsymbol{G}\cdot d\boldsymbol{r} = \iint_R \text{div}\boldsymbol{F}\,dA\,.\end{aligned}$$

Therefore,

$$\oint_C \boldsymbol{F}\cdot\mathbf{N}\,ds = \iint_R \text{div}\boldsymbol{F}\,dA\,.$$

This is sometimes called **Green's Theorem in Normal Form or Divergence Form**.

**Example A.31.** Use Green's Theorem to evaluate the line integral $\displaystyle\oint_C y^3dx + (x^3 + 3xy^2)dy$, where $C$ is the path from $(0,0)$ to $(1,1)$ along the graph of $y = x^3$ and from $(1,1)$ to $(0,0)$ along the graph of $y = x$.

Let $R = \big\{(x,y)\,\big|\,0 \leqslant x \leqslant 1, x^3 \leqslant y \leqslant x\big\}$. Then Green's Theorem implies that

$$\begin{aligned}\oint_C y^3dx + (x^3 + 3xy^2)dy &= \iint_R \Big[\frac{\partial}{\partial x}(x^3 + 3xy^2) - \frac{\partial}{\partial y}y^3\Big]\,dA = \iint_R 3x^2\,dA\\ &= \int_0^1 \Big(\int_{x^3}^x 3x^2\,dy\Big)\,dx = \int_0^1 3x^2(x - x^3)\,dx = \Big(\frac{3}{4}x^4 - \frac{1}{2}x^6\Big)\Big|_{x=0}^{x=1} = \frac{1}{4}\,.\end{aligned}$$

**Example A.32.** Let $\mathcal{D} \subseteq \mathbb{R}^2$ be the annular region $\mathcal{D} = \{(x,y) \,|\, 1 < x^2 + y^2 < 4\}$, $\boldsymbol{F}(x,y) = \dfrac{y}{x^2+y^2}\mathbf{i} - \dfrac{x}{x^2+y^2}\mathbf{j}$, and $C \subseteq \mathcal{D}$ be a simple closed curve oriented counterclockwise so that the origin is inside the region enclosed by $C$. Find $\displaystyle\oint_C \boldsymbol{F} \cdot d\boldsymbol{r}$.

Choose $r > 1$ so that the circle centered at the origin with radius $r$ lies in the intersection of $\mathcal{D}$ and the region enclosed by $C$. Let $C_r$ denote this circle with clockwise orientation, and pick a line segment $B$ connecting $C$ and $C_r$ (with starting point on $C$ and end-point on $C_r$). Define $\Gamma$ as the oriented curve $B \cup C_r \cup (-B) \cup C$, where $-B$ denotes oriented curve $B$ with opposite orientation, and let $R$ be the region enclosed by $\Gamma$. Then $R \subseteq \mathcal{D}$ and $R$ is the region lies to the left of $\Gamma$. Therefore, Green's Theorem implies that

$$\int_\Gamma \boldsymbol{F} \cdot d\boldsymbol{r} = \iint_R \operatorname{curl}\boldsymbol{F}\, dA = 0\,.$$

On the other hand,

$$\int_\Gamma \boldsymbol{F} \cdot d\boldsymbol{r} = \int_B \boldsymbol{F} \cdot d\boldsymbol{r} + \int_{C_r} \boldsymbol{F} \cdot d\boldsymbol{r} + \int_{-B} \boldsymbol{F} \cdot d\boldsymbol{r} + \int_C \boldsymbol{F} \cdot d\boldsymbol{r}\,;$$

thus by the fact that $\displaystyle\int_{-B} \boldsymbol{F} \cdot d\boldsymbol{r} = -\int_B \boldsymbol{F} \cdot d\boldsymbol{r}$, we conclude that

$$\int_C \boldsymbol{F} \cdot d\boldsymbol{r} + \int_{C_r} \boldsymbol{F} \cdot d\boldsymbol{r} = \int_\Gamma \boldsymbol{F} \cdot d\boldsymbol{r} = 0$$

or equivalently,

$$\int_C \boldsymbol{F} \cdot d\boldsymbol{r} = -\int_{C_r} \boldsymbol{F} \cdot d\boldsymbol{r} = \int_{-C_r} \boldsymbol{F} \cdot d\boldsymbol{r}\,.$$

In other words, the line integral of $\boldsymbol{F}$ along $C$ is the same as the line integral of $\boldsymbol{F}$ along the circle $C_r$ with counterclockwise orientation. Since $-C_r$ can be parameterized by

$$\boldsymbol{r}(t) = r\cos t\,\mathbf{i} + r\sin t\,\mathbf{j} \qquad t \in [0, 2\pi]\,,$$

we find that

$$\int_C \boldsymbol{F} \cdot d\boldsymbol{r} = \int_0^{2\pi} \Big(\frac{r\sin t}{r^2}\mathbf{i} - \frac{r\cos t}{r^2}\mathbf{j}\Big) \cdot \big(-r\sin t\,\mathbf{i} + r\cos t\,\mathbf{j}\big)\, dt = \int_0^{2\pi} (-1)\, dt = -2\pi\,.$$

# A.4 The Surface Integrals

## A.4.1 Parametric surfaces

**Definition A.33** (Parametric Surfaces)**.** Let $X$, $Y$ and $Z$ be functions of $u$ and $v$ that are continuous on a domain $D$ in the $uv$-plane. The collection of points

$$\Sigma \equiv \Big\{\boldsymbol{r} \in \mathbb{R}^3 \,\Big|\, \boldsymbol{r} = X(u,v)\mathbf{i} + Y(u,v)\mathbf{j} + Z(u,v)\mathbf{k} \ \text{ for some } (u,v) \in D\Big\}$$

is called a parametric surface. The equations $x = X(u,v)$, $y = Y(u,v)$, and $z = Z(u,v)$ are the parametric equations for the surface, and $\boldsymbol{r} : D \to \mathbb{R}^3$ given by $\boldsymbol{r}(u,v) = X(u,v)\mathbf{i} + Y(u,v)\mathbf{j} + Z(u,v)\mathbf{k}$ is called a parametrization of $\Sigma$.

**Definition A.34** (Regular Surfaces)**.** A parametric surface

$$\Sigma \equiv \left\{ \boldsymbol{r} \in \mathbb{R}^3 \,\middle|\, \boldsymbol{r} = X(u,v)\mathbf{i} + Y(u,v)\mathbf{j} + Z(u,v)\mathbf{k} \text{ for some } (u,v) \in D \right\}$$

is said to be regular if $X$, $Y$, $Z$ are differentiable funcitons and

$$\boldsymbol{r}_u(u,v) \times \boldsymbol{r}_v(u,v) \neq \mathbf{0} \qquad \forall\,(u,v) \in D\,,$$

where $\boldsymbol{r}_u \equiv X_u\mathbf{i} + Y_u\mathbf{j} + Z_u\mathbf{k}$ and $\boldsymbol{r}_v \equiv X_v\mathbf{i} + Y_v\mathbf{j} + Z_v\mathbf{k}$.

**Example A.35.** Let $R$ be an open region in the plane, and $f : R \to \mathbb{R}$ be a continuous function. Then the graph of $f$ is a parametric surface. In fact,

$$\text{the graph of } f = \left\{ \boldsymbol{r} \in \mathbb{R}^3 \,\middle|\, \boldsymbol{r} = x\mathbf{i} + y\mathbf{j} + f(x,y)\mathbf{k}\big) \text{ for some } (x,y) \in R \right\}.$$

Therefore, a parametric surface can be viewed as a generalization of surfaces being graphs of functions.

**Example A.36.** Let $\mathbb{S}^2 = \left\{(x,y,z) \in \mathbb{R}^3 \,\middle|\, x^2 + y^2 + z^2 = 1\right\}$ be the unit sphere in $\mathbb{R}^3$. Consider

$$\boldsymbol{r}(\theta,\phi) = \cos\theta\sin\phi\,\mathbf{i} + \sin\theta\sin\phi\,\mathbf{j} + \cos\phi\,\mathbf{k}\,, \quad (\theta,\phi) \in D = [0,2\pi] \times [0,\pi].$$

Then $\boldsymbol{r} : D \to \mathbb{S}^2$ is a continuous bijection; thus $\mathbb{S}^2$ is a parametric surface.

**Example A.37.** Consider the torus shown below

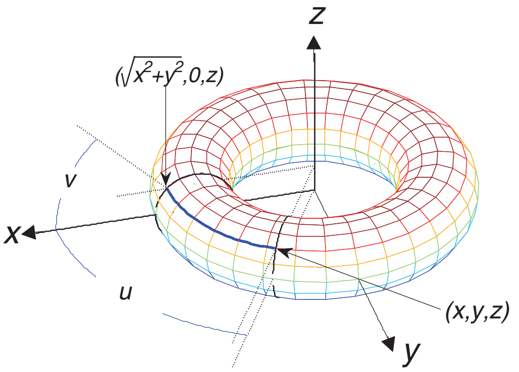


Figure A.2: Torus with parametrization $\boldsymbol{r}(u,v)$. (temporary picture)

Note that the torus has a parametrization

$$\boldsymbol{r}(u,v) = (a + b\cos v)\cos u\,\mathbf{i} + (a + b\cos v)\sin u\,\mathbf{j} + b\sin v\,\mathbf{k}\,, \quad (u,v) \in [0,2\pi] \times [0,2\pi]\,.$$

Therefore, the torus is a parametric surface.

### A.4.2 Surface area of parametric surfaces

**Theorem A.38.** *Let $D$ be an open region in the plane, and*

$$\Sigma \equiv \Big\{\boldsymbol{r} \in \mathbb{R}^3 \,\Big|\, \boldsymbol{r} = X(u,v)\mathbf{i} + Y(u,v)\mathbf{j} + Z(u,v)\mathbf{k} \;\; \textit{for some } (u,v) \in D\Big\}$$

*be a regular parametric surface so that $r$ is continuously differentiable; that is, $X_u$, $X_v$, $Y_u$, $Y_v$, $Z_u$, $Z_v$ are continuous. Then*

$$\textit{the surface area of } \Sigma = \iint_D \big\|\boldsymbol{r}_u(u,v) \times \boldsymbol{r}_v(u,v)\big\|\, d(u,v)\,.$$

**Remark A.39.** The theorem above provides one specific way of evaluating the surface integrals: if the surface $\Sigma$ is in fact a subset of the graph of a function $f : R \subseteq \mathbb{R}^2 \to \mathbb{R}$; that is, $\Sigma \subseteq \big\{x, y, f(x,y)) \,\big|\, (x,y) \in R\big\}$, then $\Sigma$ has a parametrization

$$\boldsymbol{r}(x,y) = x\mathbf{i} + y\mathbf{j} + f(x,y)\mathbf{k}\,, \qquad (x,y) \in R\,.$$

Then

$$\|\boldsymbol{r}_x(x,y) \times \boldsymbol{r}_y(x,y)\|_{\mathbb{R}^3}^2 = 1 + \Big|\frac{\partial f}{\partial x}(x,y)\Big|^2 + \Big|\frac{\partial f}{\partial y}(x,y)\Big|^2\,;$$

thus

$$\text{the surface area of } \Sigma = \iint_R \sqrt{1 + \Big|\frac{\partial f}{\partial x}(x,y)\Big|^2 + \Big|\frac{\partial f}{\partial y}(x,y)\Big|^2}\, dA\,.$$

**Example A.40.** Given the parametrization of the unit sphere $\mathbb{S}^2$ given in Example A.36, we find that

$$\begin{aligned} \boldsymbol{r}_\theta(\theta,\phi) &= -\sin\theta\sin\phi\mathbf{i} + \cos\theta\sin\phi\mathbf{j}\,, \\ \boldsymbol{r}_\phi(\theta,\phi) &= \cos\theta\cos\phi\mathbf{i} + \sin\theta\cos\phi\mathbf{j} - \sin\phi\mathbf{k} \end{aligned}$$

so that

$$\begin{aligned} (\boldsymbol{r}_u \times \boldsymbol{r}_v)(\theta,\phi) &= -\cos\theta\sin^2\phi\mathbf{i} - \sin\theta\sin^2\phi\mathbf{j} - \sin\phi\cos\phi\mathbf{k} \\ &= -\sin\phi\big(\cos\theta\sin\phi\mathbf{i} + \sin\theta\sin\phi\mathbf{j} + \cos\phi\mathbf{k}\big)\,. \end{aligned}$$

By Theorem A.38 the surface area of $\mathbb{S}^2$ is given by

$$\iint_{[0,2\pi]\times[0,\pi]} \big\|(\boldsymbol{r}_\theta \times \boldsymbol{r}_\phi)(\theta,\phi)\big\| d(\theta,\phi) = \int_0^\pi \Big(\int_0^{2\pi} \sin\phi\, d\theta\Big) d\phi = 4\pi\,.$$

**Example A.41.** Given the parametrization of the torus given in Example A.37, we find that

$$\begin{aligned} \boldsymbol{r}_u(u,v) &= -(a + b\cos v)\sin u\mathbf{i} + (a + b\cos v)\cos u\mathbf{j}\,, \\ \boldsymbol{r}_v(u,v) &= -b\sin v\cos u\mathbf{i} - b\sin v\sin u\mathbf{j} + b\cos v\mathbf{k}\,; \end{aligned}$$

thus

$$\begin{aligned}(\boldsymbol{r}_u \times \boldsymbol{r}_v)(u,v) &= b(a+b\cos v)\cos u\cos v\mathbf{i} + b(a+b\cos v)\cos v\sin u\mathbf{j} + b(a+b\cos v)\sin v\mathbf{k} \\ &= b(a+b\cos v)\big(\cos u\cos v\mathbf{i} + \sin u\cos v\mathbf{j} + \sin v\mathbf{k}\big)\,.\end{aligned}$$

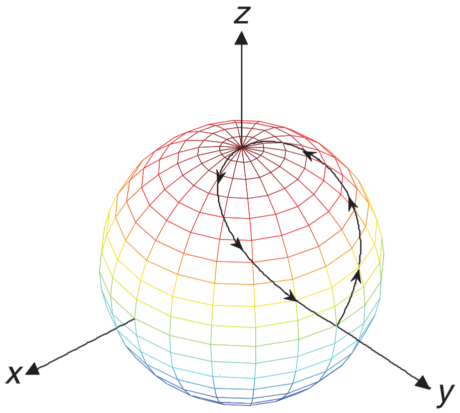
By Theorem A.38 the surface area of the torus is given by

$$\iint_{[0,2\pi]\times[0,2\pi]} b(a+b\cos v)\,d(u,v) = \int_0^{2\pi}\Big(\int_0^{2\pi}(ab+b^2\cos v)\,du\Big)dv = 4\pi^2 ab\,.$$

**Example A.42.** Let $C$ be a smooth curve parameterized by

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

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



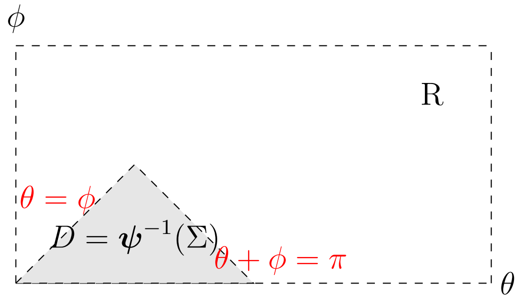
To compute the surface area of $\Sigma$, we need to find a way to parameterize $\Sigma$. Naturally we try to parameterize $\Sigma$ using the spherical coordinate. In other words, let $\mathrm{R} = (0,2\pi)\times(0,\pi)$ and $\boldsymbol{\psi} : \mathrm{R} \to \mathbb{R}^3$ be defined by

$$\boldsymbol{\psi}(\theta,\phi) = \cos\theta\sin\phi\mathbf{i} + \sin\theta\sin\phi\mathbf{j} + \cos\phi\mathbf{k}\,,$$

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

Suppose that $\gamma(t) = \big(\theta(t), \varphi(t)\big)$, $t \in \big[-\frac{\pi}{2}, \frac{\pi}{2}\big]$, is a curve in R such that $(\boldsymbol{\psi}\circ\gamma)(t) = \boldsymbol{r}(t)$. Then for $t \in \big[0, \frac{\pi}{2}\big]$, 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 $t \in \big[-\frac{\pi}{2}, 0\big]$, 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

$$\begin{aligned}\big\|(\boldsymbol{\psi}_\theta \times \boldsymbol{\psi}_\phi)(\theta,\phi)\big\|^2_{\mathbb{R}^3} &= \big\|(-\sin\theta\sin\phi\mathbf{i} + \cos\theta\sin\phi\mathbf{j}) \times (\cos\theta\cos\phi\mathbf{i} + \sin\theta\cos\phi\mathbf{j} - \sin\phi\mathbf{k})\big\|^2_{\mathbb{R}^3} \\ &= \big\|-\cos\theta\sin^2\phi\mathbf{i} - \sin\theta\sin^2\phi\mathbf{j} - (\sin^2\theta + \cos^2\theta)\sin\phi\cos\phi\mathbf{k}\big\|^2_{\mathbb{R}^3} \\ &= (\cos^2\theta + \sin^2\theta)\sin^4\phi + \sin^2\phi\cos^2\phi = \sin^2\phi\,,\end{aligned}$$

by Theorem A.38 the area of the desired surface can be computed by

$$\int_0^{\frac{\pi}{2}}\int_\phi^{\pi-\phi} \sin\phi\, d\theta d\phi = \int_0^{\frac{\pi}{2}} (\pi - 2\phi)\sin\phi\, d\phi = \big(-\pi\cos\phi + 2\phi\cos\phi - 2\sin\phi\big)\Big|_{\phi=0}^{\phi=\frac{\pi}{2}} = \pi - 2\,.$$

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

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

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

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

Therefore, with $f$ denoting the function $f(x,y) = \sqrt{1-x^2-y^2}$, Remark A.39 implies that the surface area of $\Sigma$ can be computed by

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

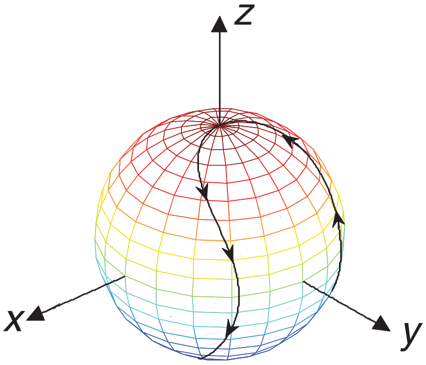
thus making a change of variable $y = \tan^2\theta$ we conclude that

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

**Example A.43.** Let $C$ be a smooth curve parameterized by

$$\boldsymbol{r}(t)=\cos(\sin t)\sin t\,\mathbf{i}+\sin(\sin t)\sin t\,\mathbf{j}+\cos t\,\mathbf{k}\,,\qquad t\in[0,2\pi]\,.$$

Then the curve $C$ is a closed curve on $\mathbb{S}^2$, and divide $\mathbb{S}^2$ into two parts. Let $\Sigma$ denote the part with smaller area.



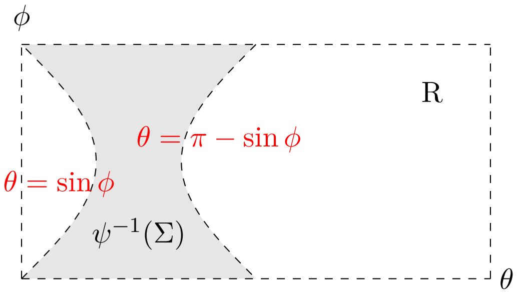
As in Example A.42, we would like to find the area of $\Sigma$. First we need to parameterize $\Sigma$. As in Example A.42, we look for $\boldsymbol{\gamma}(t)=\big(\theta(t),\phi(t)\big)$, $t\in[0,2\pi]$, on the $r\theta$-plane such that $\boldsymbol{\psi}\big(\boldsymbol{\gamma}(t)\big)=\boldsymbol{r}(t)$, where $\boldsymbol{\psi}:\mathrm{R}\equiv(0,2\pi)\times(0,\pi)$ is given by $\boldsymbol{\psi}(\theta,\phi)=\cos\theta\sin\phi\,\mathbf{i}+\sin\theta\sin\phi\,\mathbf{j}+\cos\phi\,\mathbf{k}$.

For $t\in(0,\pi)$, since $\cos t=\cos\phi(t)$ and $\phi(t)\in(0,\pi)$, we must have $\phi(t)=t$; thus the two identities $\cos(\sin t)\sin t=\cos\theta(t)\sin\phi(t)$ and $\sin(\sin t)\sin t=\sin\theta(t)\sin\phi(t)$ further imply that $\theta(t)=\sin t$. Therefore, the curve $\boldsymbol{r}\big((0,\pi)\big)$ corresponds to $\theta=\sin\phi$, $\phi\in(0,\pi)$, on $\mathrm{R}$.

On the other hand, for $t\in(\pi,2\pi)$, the identity $\cos\phi(t)=\cos t$ implies that $\phi(t)=2\pi-t$. The two identities $\cos(\sin t)\sin t=\cos\theta(t)\sin\phi(t)$ and $\sin(\sin t)\sin t=\sin\theta(t)\sin\phi(t)$ further imply that

$$\cos(\sin t)=-\cos\theta(t)\quad\text{and}\quad\sin(\sin t)=-\sin\theta(t)\qquad t\in(\pi,2\pi)\,.$$

Therefore, $\theta(t)=\pi+\sin t$ which implies that the curve $\boldsymbol{r}\big((\pi,2\pi)\big)$ corresponds to $\theta=\pi-\sin\phi$, $\phi\in(0,\pi)$, on $\mathrm{R}$.



Therefore, by Theorem A.38 the surface area of $\Sigma$ is

$$\int_0^\pi\int_{\sin\phi}^{\pi-\sin\phi}\sin\phi\,d\theta d\phi=\int_0^\pi(\pi-2\sin\phi)\sin\phi\,d\phi=-\Big(\pi\cos\phi+\phi-\frac{\sin(2\phi)}{2}\Big)\Big|_{\phi=0}^{\phi=\pi}=\pi\,.$$

### A.4.3 Surface integrals of scalar functions

Let $\Sigma \subseteq \mathbb{R}^3$ be a regular surface and $f : \Sigma \to \mathbb{R}$ be a real-valued function. We partition $\Sigma$ into small pieces $\Sigma_1, \Sigma_2, \cdots, \Sigma_n$ so that $\Sigma_i \cap \Sigma_j$ has zero area if $i \neq j$ and $\Sigma = \bigcup\limits_{k=1}^{n} \Sigma_k$. A Riemann sum of $f$ for partition $\{\Sigma_1, \cdots, \Sigma_n\}$ (of $\Sigma$) takes the form

$$\sum_{k=1}^{n} f(p_k)\nu(\Sigma_k),$$

where $p_1, \cdots, p_n$ are points on $\Sigma$ satisfying $p_k \in \Sigma_k$, and $\nu(\Sigma_k)$ denotes the surface area of $\Sigma_k$. The limit of Riemann sums as $\max\big\{\text{diam}(\Sigma_1), \text{diam}(\Sigma_2), \cdots, \text{diam}(\Sigma_n)\big\}$ approaches zero, if exists, is called the surface integral of $f$ on $\Sigma$, and is denoted by $\displaystyle\int_\Sigma f\, dS$.

**Theorem A.44.** *Let $D$ be an open region in the plane, and*

$$\Sigma \equiv \Big\{\boldsymbol{r} \in \mathbb{R}^3 \,\Big|\, \boldsymbol{r} = X(u,v)\mathbf{i} + Y(u,v)\mathbf{j} + Z(u,v)\mathbf{k} \ \text{ for some } (u,v) \in D\Big\}$$

*be a regular parametric surface so that $\boldsymbol{r}$ is continuously differentiable, and $f : \Sigma \to \mathbb{R}$ be a continuous function. Then the surface integral of $f$ on $\Sigma$ exists and is given by*

$$\iint_D (f \circ \boldsymbol{r})(u,v)\big\|(\boldsymbol{r}_u \times \boldsymbol{r}_v)(u,v)\big\|\, dA\,.$$

**Remark A.45.** If the surface $\Sigma$ is the graph of a function $f : R \subseteq \mathbb{R}^2 \to \mathbb{R}$; that is, $\Sigma = \big\{x\mathbf{i} + y\mathbf{j} + f(x,y)\mathbf{k} \,\big|\, (x,y) \in R\big\}$, then for a continuous function $g : \Sigma \to \mathbb{R}$, we have

$$\int_\Sigma g\, dS = \iint_R g\big(x, y, f(x,y)\big)\sqrt{1 + f_x(x,y)^2 + f_y(x,y)^2}\, dA\,. \tag{A.3}$$

**Example A.46.** Evaluate the surface integral

$$\int_\Sigma (y^2 + 2yz) dS\,,$$

where $\Sigma$ is the first-octant portion of the plane $2x + y + 2z = 6$.

First, we note that $\Sigma$ can be parameterized by

$$\Sigma = \Big\{x\mathbf{i} + y\mathbf{j} + \frac{6 - 2x - y}{2}\mathbf{k} \,\Big|\, (x,y) \in R\Big\}\,,$$

where $R$ is the triangle $\big\{(x,y) \,\big|\, x \in [0,3], 0 \leqslant y \leqslant 6 - 2x\big\}$. Therefore, using (A.3) and Fubini's Theorem we find that

$$\begin{aligned}\int_\Sigma (y^2 + 2yz)\, dS &= \iint_R \big(y^2 + 2y \cdot \frac{6 - 2x - y}{2}\big)\sqrt{1 + (-1)^2 + \big(-\frac{1}{2}\big)^2}\, dA \\ &= \int_0^3 \Big(\int_0^{6-2x} \frac{3}{2}(6y - 2xy)\, dy\Big) dx = \int_0^3 \Big(\int_0^{6-2x} (9y - 3xy)\, dy\Big) dx \\ &= \int_0^3 \frac{(9 - 3x)y^2}{2}\Big|_{y=0}^{y=6-2x} dx = \int_0^3 6(3 - x)^3\, dx = -\frac{6(3 - x)^4}{4}\Big|_{x=0}^{x=3} = \frac{243}{2}\,.\end{aligned}$$

**Example A.47.** Evaluate the surface integral

$$\int_{\Sigma} \sqrt{x(1+2z)}\, dS\,,$$

where $\Sigma$ is the portion of the cylinder $z = \dfrac{y^2}{2}$ over the triangular region

$$R \equiv \big\{(x,y)\,\big|\, x \geqslant 0, y \geqslant 0, x+y \leqslant 1\big\}$$

in the $xy$-plane.

We compute the surface integral using (A.3) and Fubini's Theorem and obtain that

$$\begin{aligned}\int_{\Sigma} \sqrt{x(1+2z)}\, dS &= \iint_R \sqrt{x(1+y^2)}\sqrt{1+0^2+y^2}\, dA = \int_0^1 \Big(\int_0^{1-x} \sqrt{x}(1+y^2)\, dy\Big)\, dx \\ &= \int_0^1 \sqrt{x}\Big(y + \frac{y^3}{3}\Big)\Big|_{y=0}^{y=1-x} dx = \int_0^1 \sqrt{x}\Big(1-x+\frac{(1-x)^3}{3}\Big)\, dx \\ &= \frac{1}{3}\int_0^1 \Big(4x^{\frac{1}{2}} - 6x^{\frac{3}{2}} + 3x^{\frac{5}{2}} - x^{\frac{7}{2}}\Big)\, dx \\ &= \frac{1}{3}\Big(\frac{8}{3}x^{\frac{3}{2}} - \frac{12}{5}x^{\frac{5}{2}} + \frac{6}{7}x^{\frac{7}{2}} - \frac{2}{9}x^{\frac{9}{2}}\Big)\Big|_{x=0}^{x=1} = \frac{284}{945}\,.\end{aligned}$$

**Example A.48.** Evaluate the surface integral

$$\int_{\Sigma} z\, dS\,,$$

where $\Sigma$ is the surface given in Example A.42.

As already shown in Example A.42, $\Sigma$ can be parameterized by

$$\Sigma = \Big\{\boldsymbol{r}(\theta,\phi) = \cos\theta\sin\phi\,\mathbf{i} + \sin\theta\sin\phi\,\mathbf{j} + \cos\phi\,\mathbf{k}\,\Big|\, 0 \leqslant \phi \leqslant \frac{\pi}{2}, \phi \leqslant \theta \leqslant \pi-\phi\Big\}\,.$$

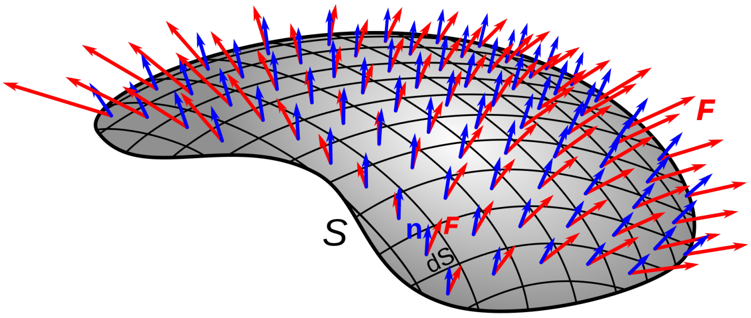
Therefore,

$$\begin{aligned}\int_{\Sigma} z\, dS &= \int_0^{\frac{\pi}{2}} \Big(\int_{\phi}^{\pi-\phi} \cos\phi \big\|(\boldsymbol{r}_\theta \times \boldsymbol{r}_\phi)(\theta,\phi)\big\|\, d\theta\Big) d\phi = \int_0^{\frac{\pi}{2}} \Big(\int_{\phi}^{\pi-\phi} \cos\phi\sin\phi\, d\theta\Big) d\phi \\ &= \frac{1}{2}\int_0^{\frac{\pi}{2}} (\pi-2\phi)\sin(2\phi)\, d\phi = \frac{1}{2}\Big[(\pi-2\phi)\frac{-\cos(2\phi)}{2}\Big|_{\phi=0}^{\phi=\frac{\pi}{2}} - \int_0^{\frac{\pi}{2}} \frac{\cos(2\phi)}{2}\, d\phi\Big] \\ &= \frac{1}{2}\Big(\frac{\pi}{2} - \frac{\sin(2\phi)}{4}\Big|_{\phi=0}^{\phi=\frac{\pi}{2}}\Big) = \frac{\pi}{4}\,.\end{aligned}$$

## A.5 The Flux Integrals

Let $S \subseteq \mathbb{R}^3$ be a regular parametric surface with a continuous normal vector field $\boldsymbol{n} : S \to \mathbb{R}^3$ (sometimes this is termed "$S$ is oriented by $\boldsymbol{n}$"). For a bounded continuous vector-valued function $\boldsymbol{F} : S \to \mathbb{R}^3$, the flux integral of $\boldsymbol{F}$ across $S$ (in direction $\boldsymbol{n}$) is the surface integral of $\boldsymbol{F}\cdot\boldsymbol{n}$ on $S$; that is,

$$\text{the flux integral of } \boldsymbol{F} \text{ across } S \text{ (in direction } \boldsymbol{n}) = \int_S \boldsymbol{F}\cdot\boldsymbol{n}\, dS\,.$$



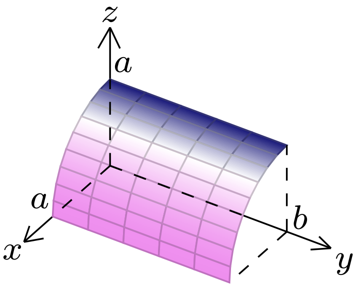
### A.5.1 Physical interpretation

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 $\boldsymbol{n}$, 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 $\boldsymbol{n}$ by the fluid in a time period of $\Delta t$ is

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

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

**Example A.49.** Find the flux integral of the vector field $\boldsymbol{F}(x, y, z) = (x, y^2, z)$ upward through the first octant part $\Sigma$ of the cylindrical surface $x^2 + z^2 = a^2$, $0 < y < b$.



First, we parameterize $\Sigma$ by

$$\boldsymbol{r}(u, v) = u\mathbf{i} + v\mathbf{j} + \sqrt{a^2 - u^2}\mathbf{k}\,, \quad (u, v) \in D = (0, a) \times (0, b)$$

so that $\|(\boldsymbol{r}_u \times \boldsymbol{r}_v)(u, v)\|^2_{\mathbb{R}^3} = \dfrac{a^2}{a^2 - u^2}$, and the upward-pointing unit normal is $\mathbf{N}(x, y, z) = \left(\dfrac{x}{a}, 0, \dfrac{z}{a}\right)$. Therefore,

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

### A.5.2 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 continuously differentiable vector field. Suppose that $\mathcal{O}$ is a bounded open set whose boundary is piecewise smooth so that an outward-pointing unit normal vector field $\mathbf{N} = (\mathrm{N}_1, \mathrm{N}_2, \mathrm{N}_3)$ can be defined on $\partial\mathcal{O}$ except on some curves. Then the flux integral of $\boldsymbol{u}$ on $\partial\mathcal{O}$ in the direction $\mathbf{N}$ is

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

Consider a special case that $\mathcal{O} = (a_1, a_2) \times (b_1, b_2) \times (c_1, c_2)$ be an open cube so that $\partial\mathcal{O} = \bigcup\limits_{k=1}^{3} \Sigma_k$, where $\Sigma_1 = \{a_1, a_2\} \times [b_1, b_2] \times [c_1, c_2]$, $\Sigma_2 = [a_1, a_2] \times \{b_1, b_2\} \times [c_1, c_2]$ and $\Sigma_3 = [a_1, a_2] \times [b_1, b_2] \times \{c_1, c_2\}$. Then

$$\int_{\partial\mathcal{O}} \boldsymbol{u} \cdot \mathbf{N}\, dS = \sum_{k=1}^{3} \int_{\Sigma_k} \boldsymbol{u} \cdot \mathbf{N}\, dS\,.$$

Since on $\Sigma_3$ the outward-pointing normal $\mathbf{N}$ is given by

$$\mathbf{N}(x, y, z) = \begin{cases} -\mathbf{k} & \text{if } (x, y, z) \in [a_1, a_2] \times [b_1, b_2] \times \{c_1\}\,, \\ \ \ \mathbf{k} & \text{if } (x, y, z) \in [a_1, a_2] \times [b_1, b_2] \times \{c_2\}\,, \end{cases}$$

we find that

$$\begin{aligned} \int_{\Sigma_3} \boldsymbol{u} \cdot \mathbf{N}\, dS &= \iint_{[a_1,a_2]\times[b_1,b_2]} u_3(x, y, c_2)\, dA - \iint_{[a_1,a_2]\times[b_1,b_2]} u_3(x, y, c_1)\, dA \\ &= \iint_{[a_1,a_2]\times[b_1,b_2]} u_3(x, y, z)\Big|_{x=c_1}^{x=c_2} dA \\ &= \iint_{[a_1,a_2]\times[b_1,b_2]} \Big( \int_{[c_1,c_2]} \frac{\partial u_3}{\partial z}(x, y, z)\, dz \Big)\, dA = \iiint_{\mathcal{O}} \frac{\partial u_3}{\partial z}\, dV\,, \end{aligned}$$

where the last equality is established by Fubini's Theorem. Similarly,

$$\int_{\Sigma_1} \boldsymbol{u} \cdot \mathbf{N}\, dS = \iiint_{\mathcal{O}} \frac{\partial u_1}{\partial x}\, dV \quad \text{and} \quad \int_{\Sigma_2} \boldsymbol{u} \cdot \mathbf{N}\, dS = \iiint_{\mathcal{O}} \frac{\partial u_2}{\partial y}\, dV\,;$$

thus

$$\int_{\partial\mathcal{O}} \boldsymbol{u} \cdot \mathbf{N}\, dS = \iiint_{\mathcal{O}} \Big( \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} + \frac{\partial u_3}{\partial z} \Big)\, dV = \iiint_{\mathcal{O}} \mathrm{div}\boldsymbol{u}\, dV\,. \tag{A.4}$$

**Remark A.50.** Let $\Omega \subseteq \mathbb{R}^3$ be an open set, and $\boldsymbol{u} : \bar{\Omega} \to \mathbb{R}^3$ be a continuously differentiable vector field, and $\mathcal{O}(\boldsymbol{a}, r)$ denote a cube centered at $\boldsymbol{a} \in \Omega$ with side length $r$. Using (A.4), the continuity of div$\boldsymbol{u}$ implies that

$$\lim_{r\to 0} \frac{1}{|\mathcal{O}(\boldsymbol{a}, r)|} \int_{\partial\mathcal{O}(\boldsymbol{a},r)} \boldsymbol{u} \cdot \mathbf{N}\, dS = (\mathrm{div}\boldsymbol{u})(\boldsymbol{a}) \qquad \forall\, \boldsymbol{a} \in \Omega\,.$$

In other words, div$\boldsymbol{u}$ at a point $\boldsymbol{x}$ is the instantaneous amount (per volume) of material (with concentration 1) carried outside an infinitesimal cube centered at $\boldsymbol{x}$.

## A.6 The Divergence Theorem

**Theorem A.51 (The Divergence Theorem).** *Let $\Omega \subseteq R^3$ be a bounded domain such that $\partial\Omega$ is piecewise smooth with outward pointing normal $\mathbf{N}$, and $\boldsymbol{w} : \bar{\Omega} \to \mathbb{R}^3$ be continuously differentiable vector field. Then*

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

**Remark A.52.** Similar to Green's Theorem in Divergence Form, the Divergence Theorem states that "一向量場在一區域的邊界上的某種有方向性的和（積分）等於該向量場某種微分的樣子（即散度）在該區域上的和（積分）":

$$\text{一向量場在一區域的邊界上的某種具方向性的和} = \int_{\partial\Omega} \boldsymbol{w}\cdot\mathbf{N}\,dS.$$

$$\text{該向量場某種微分的樣子在該區域上的和} = \iiint_{\Omega} \operatorname{div}\boldsymbol{w}\,dV.$$

You should try to compare with the Fundamental Theorem of Calculus

$$\int_a^b f'(x)\,dx = f(b) - f(a) \text{ ``$=$'' } \int_{\partial\,[a,b]} f$$

and get the idea of the Divergence Theorem.

Letting $\boldsymbol{w}$ be the product of a scalar function $\varphi$ and a vector field $\boldsymbol{v}$ in Theorem A.51, using the identity

$$\operatorname{div}(\varphi\boldsymbol{v}) = \sum_{i=1}^{3} \frac{\partial \varphi v^i}{\partial x_i} = \sum_{i=1}^{3} \Big(\frac{\partial\varphi}{\partial x_i} v^i + \varphi\frac{\partial v^i}{\partial x_i}\Big) = \nabla\varphi\cdot\boldsymbol{v} + \varphi\operatorname{div}\boldsymbol{v}\,,$$

we conclude the following

**Corollary A.53.** *Let $\Omega \subseteq \mathbb{R}^3$ be a bounded domain such that $\partial\Omega$ is piecewise smooth with outward-pointing unit normal $\mathbf{N}$, $\boldsymbol{v} : \bar{\Omega} \to \mathbb{R}^3$ be a continuously differentiable vector field, and $\varphi : \bar{\Omega} \to \mathbb{R}$ be continuously differentiable. Then*

$$\iiint_{\Omega} \varphi\operatorname{div}\boldsymbol{v}\,dV = \int_{\partial\Omega} (\boldsymbol{v}\cdot\mathbf{N})\varphi\,dS - \iiint_{\Omega} \boldsymbol{v}\cdot\nabla\varphi\,dV\,.$$

Letting $\boldsymbol{v} = f\mathbf{e}_i$ for some continuously differentiable function $f : \bar{\Omega} \to \mathbb{R}$ in the corollary above, we obtain the following

**Corollary A.54.** *Let $\Omega \subseteq \mathbb{R}^3$ be a bounded domain such that $\partial\Omega$ is piecewise smooth with outward-pointing normal $\mathbf{N} = (\mathrm{N}_1, \mathrm{N}_2, \mathrm{N}_3)$, and $f, \varphi : \bar{\Omega} \to \mathbb{R}$ be continuously differentiable functions. Then*

$$\iiint_{\Omega} \varphi\,\frac{\partial f}{\partial x_i}\,dV = \int_{\partial\Omega} f\varphi\mathrm{N}_i\,dS - \iiint_{\Omega} f\,\frac{\partial\varphi}{\partial x_i}\,dV\,.$$

**Example A.55.** Let $\Omega$ be the the first octant part bounded by the cylindrical surface $x^2 + z^2 = a^2$ and the plane $y = b$, and $\boldsymbol{F} : \Omega \to \mathbb{R}^3$ be a vector-valued function defined by $\boldsymbol{F}(x, y, z) = (x, y^2, z)$.

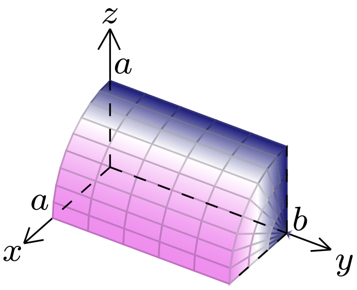


Figure A.3: The domain $\Omega$ and its five pieces of boundaries

First we compute the flux integral $\displaystyle\int_{\partial\Omega} \boldsymbol{F} \cdot \mathbf{N}\, dS$. We note that the boundary of $\Omega$ has five parts: $\Sigma$ as given in Example A.49, two rectangles $\mathrm{R}_1 = \{x = 0\} \times [0, b] \times [0, a]$, $\mathrm{R}_2 = [0, a] \times [0, b] \times \{z = 0\}$, and two quarter disc $\mathrm{D}_1 = \big\{(x, 0, z) \in \mathbb{R}^3 \,\big|\, x^2 + z^2 \leqslant a^2, x, z \geqslant 0\big\}$ and $\mathrm{D}_2 = \big\{(x, b, z) \in \mathbb{R}^3 \,\big|\, x^2 + z^2 \leqslant a^2, x, z \geqslant 0\big\}$. Therefore,

$$\begin{aligned}
\int_{\mathrm{R}_1} \boldsymbol{F} \cdot \mathbf{N}\, dS &= \int_0^a \int_0^b (0, y^2, z) \cdot (-1, 0, 0)\, dydz = 0\,, \\
\int_{\mathrm{R}_2} \boldsymbol{F} \cdot \mathbf{N}\, dS &= \int_0^a \int_0^b (x, y^2, 0) \cdot (0, 0, -1)\, dydx = 0\,, \\
\int_{\mathrm{D}_1} \boldsymbol{F} \cdot \mathbf{N}\, dS &= \int_0^a \int_0^{\sqrt{a^2 - x^2}} (x, 0, z) \cdot (0, -1, 0)\, dzdx = 0\,,
\end{aligned}$$

and

$$\int_{\mathrm{D}_1} \boldsymbol{F} \cdot \mathbf{N}\, dS = \int_0^a \int_0^{\sqrt{a^2 - x^2}} (x, b^2, z) \cdot (0, 1, 0)\, dzdx = b^2 \int_0^a \int_0^{\sqrt{a^2 - x^2}} dzdx = \frac{\pi a^2 b^2}{4}\,.$$

Together with the result in Example A.49, we find that

$$\int_{\partial\Omega} \boldsymbol{F} \cdot \mathbf{N}\, dS = \Big( \int_{\Sigma} + \int_{\mathrm{R}_1} + \int_{\mathrm{R}_2} + \int_{\mathrm{D}_1} + \int_{\mathrm{D}_2} \Big) \boldsymbol{F} \cdot \mathbf{N}\, dS = \frac{\pi a^2 b^2}{4} + \frac{\pi a^2 b}{2} = \frac{\pi a^2 (b^2 + 2b)}{4}\,.$$

On the other hand,

$$\begin{aligned}
\iiint_{\Omega} \operatorname{div} \boldsymbol{F}\, dV &= \int_0^a \int_0^b \int_0^{\sqrt{a^2 - x^2}} (2 + 2y)\, dzdydx = (b^2 + 2b) \int_0^a \int_0^{\sqrt{a^2 - x^2}} dzdx \\
&= \frac{\pi a^2 (b^2 + 2b)}{4}\,.
\end{aligned}$$

Therefore, $\displaystyle\iiint_{\Omega} \operatorname{div} \boldsymbol{F}\, dV = \int_{\partial\Omega} \boldsymbol{F} \cdot \mathbf{N}\, dS$.

# Index