

## Chapter 1

# Mathematical Foundations

### 1.1 Big- $O$ Notations

In the description of algorithmic complexity, we often have to use the order notations, often in terms of big  $O$  and small  $o$ . Loosely speaking, for two functions  $f(x)$  and  $g(x)$ , if

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} \rightarrow K, \quad (1.1)$$

where  $K$  is a finite, non-zero limit, we write

$$f = O(g). \quad (1.2)$$

The big  $O$  notation means that  $f$  is asymptotically equivalent to the order of  $g(x)$ . If the limit is unity or  $K = 1$ , we say  $f(x)$  is order of  $g(x)$ . In this special case, we write

$$f \sim g, \quad (1.3)$$

which is equivalent to  $f/g \rightarrow 1$  and  $g/f \rightarrow 1$  as  $x \rightarrow x_0$ . Obviously,  $x_0$  can be any value, including 0 and  $\infty$ . The notation  $\sim$  does not necessarily mean  $\approx$  in general, though it might give the same results, especially in the case when  $x \rightarrow 0$ . For example,  $\sin x \sim x$  and  $\sin x \approx x$  if  $x \rightarrow 0$ .

When we say  $f$  is order of 100 (or  $f \sim 100$ ), this does not mean  $f \approx 100$ , but it can mean that  $f$  could be between about 50 and 150. The small  $o$  notation is often used if the limit tends to 0. That is

$$\lim_{x \rightarrow x_0} \frac{f}{g} \rightarrow 0, \quad (1.4)$$

or

$$f = o(g). \quad (1.5)$$

If  $g > 0$ ,  $f = o(g)$  is equivalent to  $f \ll g$ . For example, for  $\forall x \in \mathcal{R}$ , we have  $e^x \approx 1 + x + O(x^2) \approx 1 + x + \frac{x^2}{2} + o(x)$ .

---

**Example 1.1:** A classic example is Stirling's asymptotic series for factorials

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51480n^3} - \dots\right),$$

which can demonstrate the fundamental difference between asymptotic series and the standard approximate expansions. For standard power expansions, the error  $R_k(h^k) \rightarrow 0$ , but for an asymptotic series, the error of the truncated series  $R_k$  decreases compared with the leading term [here  $\sqrt{2\pi n}(n/e)^n$ ]. However,  $R_n$  does not necessarily tend to zero. In fact,

$$R_2 = \frac{1}{12n} \cdot \sqrt{2\pi n}(n/e)^n,$$

is still very large as  $R_2 \rightarrow \infty$  if  $n \gg 1$ . For example, for  $n = 100$ , we have  $n! = 9.3326 \times 10^{157}$ , while the leading approximation is  $\sqrt{2\pi n}(n/e)^n = 9.3248 \times 10^{157}$ . The difference between these two values is  $7.7740 \times 10^{154}$ , which is still very large, though three orders smaller than the leading approximation.

---

## 1.2 Vector and Vector Calculus

Vector analysis is an important part of computational mathematics. Many quantities such as force, velocity, and deformation in sciences are vectors which have both a magnitude and a direction. Vectors are a special class of matrices. In this chapter, we will briefly review the basic concepts in linear algebra.

A vector  $\mathbf{u}$  is a set of ordered numbers  $\mathbf{u} = (u_1, u_2, \dots, u_n)$ , where its components  $u_i (i = 1, \dots, n) \in \mathfrak{R}$  are real numbers. All these vectors form an  $n$ -dimensional vector space  $\mathcal{V}^n$ . A simple example is the position vector  $\mathbf{p} = (x, y, z)$  where  $x, y, z$  are the 3-D Cartesian coordinates.

To add any two vectors  $\mathbf{u} = (u_1, \dots, u_n)$  and  $\mathbf{v} = (v_1, \dots, v_n)$ , we simply add their corresponding components,

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, \dots, u_n + v_n), \quad (1.6)$$

and the sum is also a vector. The addition of vectors has commutability ( $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ ) and associativity  $[(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})]$ . This is because each of the components is obtained by simple addition which means it has the same properties.

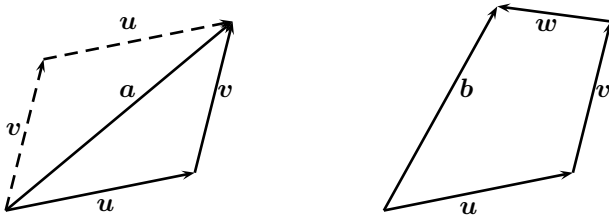


Fig. 1.1 Addition of vectors: (a) parallelogram  $\mathbf{a} = \mathbf{u} + \mathbf{v}$ ; (b) vector polygon  $\mathbf{b} = \mathbf{u} + \mathbf{v} + \mathbf{w}$ .

The zero vector  $\mathbf{0}$  is a special case where all its components are zeros. The multiplication of a vector  $\mathbf{u}$  with a scalar or constant  $\alpha \in \Re$  is carried out by the multiplication of each component,

$$\alpha \mathbf{u} = (\alpha u_1, \alpha u_2, \dots, \alpha u_n). \quad (1.7)$$

Thus, we have

$$-\mathbf{u} = (-u_1, -u_2, \dots, -u_n). \quad (1.8)$$

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

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \dots + x_n y_n = \sum_{i=1}^n x_i y_i, \quad (1.9)$$

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

$$|\mathbf{x}| = \|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^n x_i^2}. \quad (1.10)$$

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

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}, \quad \mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}, \quad (1.11)$$

and

$$(\alpha \mathbf{u}) \cdot (\beta \mathbf{v}) = (\alpha \beta) \mathbf{u} \cdot \mathbf{v}, \quad (1.12)$$

where  $\alpha, \beta \in \Re$  are constants.

The angle  $\theta$  between two vectors  $\mathbf{u}$  and  $\mathbf{v}$  can be calculated using their dot product

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos(\theta), \quad 0 \leq \theta \leq \pi, \quad (1.13)$$

which leads to

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \quad (1.14)$$

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

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

$$\|\mathbf{u} \cdot \mathbf{v}\| \leq \|\mathbf{u}\| \|\mathbf{v}\|. \quad (1.15)$$

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

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

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

$$= (x_2y_3 - x_3y_2)\mathbf{i} + (x_3y_1 - x_1y_3)\mathbf{j} + (x_1y_2 - x_2y_1)\mathbf{k}. \quad (1.16)$$

In fact, the norm of  $\|\mathbf{x} \times \mathbf{y}\|$  is the area of the parallelogram formed by  $\mathbf{x}$  and  $\mathbf{y}$ . We have

$$\|\mathbf{x} \times \mathbf{y}\| = \|\mathbf{x}\| \|\mathbf{y}\| \sin \theta, \quad (1.17)$$

where  $\theta$  is the angle between the two vectors. In addition, the vector  $\mathbf{u} = \mathbf{x} \times \mathbf{y}$  is perpendicular to both  $\mathbf{x}$  and  $\mathbf{y}$ , following a right-hand rule.

It is straightforward to check that the cross product has the following properties:

$$\mathbf{x} \times \mathbf{y} = -\mathbf{y} \times \mathbf{x}, \quad (\mathbf{x} + \mathbf{y}) \times \mathbf{z} = \mathbf{x} \times \mathbf{z} + \mathbf{y} \times \mathbf{z}, \quad (1.18)$$

and

$$(\alpha\mathbf{x}) \times (\beta\mathbf{y}) = (\alpha\beta)\mathbf{x} \times \mathbf{y}, \quad \alpha, \beta \in \mathfrak{R}. \quad (1.19)$$

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

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}, \quad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \quad \mathbf{k} \times \mathbf{i} = \mathbf{j}. \quad (1.20)$$

**Example 1.2:** For two 3-D vectors  $\mathbf{u} = (4, 5, -6)$  and  $\mathbf{v} = (2, -2, 1/2)$ , their dot product is

$$\mathbf{u} \cdot \mathbf{v} = 4 \times 2 + 5 \times (-2) + (-6) \times 1/2 = -5.$$

As their moduli are

$$\|\mathbf{u}\| = \sqrt{4^2 + 5^2 + (-6)^2} = \sqrt{77},$$

$$\|\mathbf{v}\| = \sqrt{2^2 + (-2)^2 + (1/2)^2} = \sqrt{33}/2,$$

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

$$\cos \theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} = \frac{-5}{\sqrt{77} \times \sqrt{33}/2} = -\frac{10}{11\sqrt{21}},$$

or

$$\theta = \cos^{-1}\left(-\frac{10}{11\sqrt{21}}\right) \approx 101.4^\circ.$$

Their cross product is

$$\begin{aligned} \mathbf{w} &= \mathbf{u} \times \mathbf{v} \\ &= (5 \times 1/2 - (-2) \times (-6), (-6) \times 2 - 4 \times 1/2, 4 \times (-2) - 5 \times 2) \\ &= (-19/2, -14, -18). \end{aligned}$$

Similarly, we have

$$\mathbf{v} \times \mathbf{u} = (19/2, 14, 18) = -\mathbf{u} \times \mathbf{v}.$$

The norm of the cross product is

$$\|\mathbf{w}\| = \sqrt{\left(\frac{-19}{2}\right)^2 + (-14)^2 + (-18)^2} \approx 24.70,$$

while

$$\|\mathbf{u}\| \|\mathbf{v}\| \sin \theta = \sqrt{77} \times \frac{\sqrt{33}}{2} \times \sin(101.4^\circ) \approx 24.70 = \|\mathbf{w}\|.$$

It is easy to verify that

$$\mathbf{u} \cdot \mathbf{w} = 4 \times (-19/2) + 5 \times (-14) + (-6) \times (-18) = 0,$$

and

$$\mathbf{v} \cdot \mathbf{w} = 2 \times (-19/2) + (-2) \times (-14) + 1/2 \times (-18) = 0.$$

Indeed, the vector  $\mathbf{w}$  is perpendicular to both  $\mathbf{u}$  and  $\mathbf{v}$ .

---

Any vector  $\mathbf{v}$  in an  $n$ -dimensional vector space  $\mathcal{V}^n$  can be written as a combination of a set of  $n$  independent basis vectors or orthogonal spanning vectors  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ , so that

$$\mathbf{v} = v_1 \mathbf{e}_1 + v_2 \mathbf{e}_2 + \dots + v_n \mathbf{e}_n = \sum_{i=1}^n v_i \mathbf{e}_i, \quad (1.21)$$

where the coefficients/scalars  $v_1, v_2, \dots, v_n$  are the components of  $\mathbf{v}$  relative to the basis  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ . The most common basis vectors are the orthogonal unit vectors. In a three-dimensional case, they are  $\mathbf{i} = (1, 0, 0)$ ,  $\mathbf{j} = (0, 1, 0)$ ,  $\mathbf{k} = (0, 0, 1)$  for  $x$ -,  $y$ -,  $z$ -axis, respectively. Thus, we have

$$\mathbf{x} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}. \quad (1.22)$$

The three unit vectors satisfy  $\mathbf{i} \cdot \mathbf{j} = \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{i} = 0$ .

Two non-zero vectors  $\mathbf{u}$  and  $\mathbf{v}$  are said to be linearly independent if  $\alpha\mathbf{u} + \beta\mathbf{v} = \mathbf{0}$  implies that  $\alpha = \beta = 0$ . If  $\alpha, \beta$  are not all zeros, then these two vectors are linearly dependent. Two linearly dependent vectors are parallel ( $\mathbf{u} \parallel \mathbf{v}$ ) to each other. Similarly, any three linearly dependent vectors  $\mathbf{u}, \mathbf{v}, \mathbf{w}$  are in the same plane.

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

$$\mathbf{P}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}, \quad (1.23)$$

we can write its velocity as

$$\mathbf{v} = \frac{d\mathbf{P}}{dt} = \dot{x}(t)\mathbf{i} + \dot{y}(t)\mathbf{j} + \dot{z}(t)\mathbf{k}, \quad (1.24)$$

and acceleration as

$$\mathbf{a} = \frac{d^2\mathbf{P}}{dt^2} = \ddot{x}(t)\mathbf{i} + \ddot{y}(t)\mathbf{j} + \ddot{z}(t)\mathbf{k}, \quad (1.25)$$

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

$$\mathbf{P} = \int_0^t \mathbf{v} dt + \mathbf{p}_0, \quad (1.26)$$

where  $\mathbf{p}_0$  is a vector constant or the initial position at  $t = 0$ .

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

$$\frac{d(\alpha\mathbf{a})}{dt} = \alpha \frac{d\mathbf{a}}{dt}, \quad \frac{d(\mathbf{a} \cdot \mathbf{b})}{dt} = \frac{d\mathbf{a}}{dt} \cdot \mathbf{b} + \mathbf{a} \cdot \frac{d\mathbf{b}}{dt}, \quad (1.27)$$

and

$$\frac{d(\mathbf{a} \times \mathbf{b})}{dt} = \frac{d\mathbf{a}}{dt} \times \mathbf{b} + \mathbf{a} \times \frac{d\mathbf{b}}{dt}. \quad (1.28)$$

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

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

$$\frac{\partial \phi}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla \phi = \frac{\partial \phi}{\partial x} \cos(\alpha) + \frac{\partial \phi}{\partial y} \cos(\beta) + \frac{\partial \phi}{\partial z} \cos(\gamma), \quad (1.29)$$

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

$$\text{grad} \phi = \nabla \phi = \frac{\partial \phi}{\partial x} \mathbf{i} + \frac{\partial \phi}{\partial y} \mathbf{j} + \frac{\partial \phi}{\partial z} \mathbf{k}. \quad (1.30)$$

This is the same as the application of the del operator  $\nabla$  to the scalar function  $\phi$

$$\nabla = \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k}. \quad (1.31)$$

The direction of the gradient operator on a scalar field gives a vector field.

As the gradient operator is a linear operator, it is straightforward to check that it has the following properties:

$$\nabla(\alpha\psi + \beta\phi) = \alpha\nabla\psi + \beta\nabla\phi, \quad \nabla(\psi\phi) = \psi\nabla\phi + \phi\nabla\psi, \quad (1.32)$$

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

For a vector field

$$\mathbf{u}(x, y, z) = u(x, y, z)\mathbf{i} + v(x, y, z)\mathbf{j} + w(x, y, z)\mathbf{k}, \quad (1.33)$$

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

$$\text{div } \mathbf{u} \equiv \nabla \cdot \mathbf{u} = \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} + \frac{\partial u_3}{\partial z}, \quad (1.34)$$

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

$$\text{curl } \mathbf{u} \equiv \nabla \times \mathbf{u} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ u_1 & u_2 & u_3 \end{vmatrix}. \quad (1.35)$$

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

$$\nabla^2 \phi = \nabla \cdot (\nabla \phi) = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2}, \quad (1.36)$$

for Laplace's equation

$$\Delta\phi \equiv \nabla^2\phi = 0. \quad (1.37)$$

Some important theorems are often rewritten in terms of the above three operators, especially in fluid dynamics and finite element analysis. For example, Gauss's theorem connects the integral of divergence with the related surface integral

$$\iiint_{\Omega} (\nabla \cdot \mathbf{Q}) d\Omega = \iint_S \mathbf{Q} \cdot \mathbf{n} dS. \quad (1.38)$$

### 1.3 Matrices and Matrix Decomposition

Matrices are widely used in computational mathematics, especially in the implementation of many algorithms. A matrix is a table or array of numbers or functions arranged in rows and columns. The elements or entries of a matrix  $\mathbf{A}$  are often denoted as  $a_{ij}$ . For a matrix  $\mathbf{A}$  with  $m$  rows and  $n$  columns,

$$\mathbf{A} \equiv [a_{ij}] = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2n} \\ \vdots & \vdots & & & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix}, \quad (1.39)$$

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

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} e^x & \sin x \\ -i \cos x & e^{i\theta} \end{pmatrix}, \quad (1.40)$$

and

$$\mathbf{u} = \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \quad (1.41)$$

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

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

$$\mathbf{C} = \mathbf{A} + \mathbf{B}, \quad c_{ij} = a_{ij} + b_{ij}, \quad (1.42)$$

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

$$c_{ij} = (\mathbf{AB})_{ij} = \sum_{k=1}^n a_{ik}b_{kj}. \quad (1.43)$$

If  $\mathbf{A}$  is a square matrix, then we have  $\mathbf{A}^n = \overbrace{\mathbf{AA}\dots\mathbf{A}}^n$ . The multiplications of matrices are generally not commutative, i.e.,  $\mathbf{AB} \neq \mathbf{BA}$ . However, the multiplication has associativity  $\mathbf{A}(\mathbf{uv}) = (\mathbf{Au})\mathbf{v}$  and  $\mathbf{A}(\mathbf{u}+\mathbf{v}) = \mathbf{Au}+\mathbf{Av}$ .

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

$$(\mathbf{A}^T)^T = \mathbf{A}, \quad (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \quad (1.44)$$

The differentiation and integral of a matrix are carried out over each of its members or elements. For example, for a  $2 \times 2$  matrix

$$\frac{d\mathbf{A}}{dt} = \dot{\mathbf{A}} = \begin{pmatrix} \frac{da_{11}}{dt} & \frac{da_{12}}{dt} \\ \frac{da_{21}}{dt} & \frac{da_{22}}{dt} \end{pmatrix}, \quad (1.45)$$

and

$$\int \mathbf{A}dt = \begin{pmatrix} \int a_{11}dt & \int a_{12}dt \\ \int a_{21}dt & \int a_{22}dt \end{pmatrix}. \quad (1.46)$$

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

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & d_n \end{pmatrix}. \quad (1.47)$$

For example, the matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (1.48)$$

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

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}. \quad (1.49)$$

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

There are three important matrices: lower (upper) triangular matrix, tridiagonal matrix, and augmented matrix, and they are important in the solution of linear equations. A tridiagonal matrix often arises naturally from the finite difference and finite volume discretization of partial differential equations, and it can in general be written as

$$\mathbf{Q} = \begin{pmatrix} b_1 & c_1 & 0 & 0 & \dots & 0 & 0 \\ a_2 & b_2 & c_2 & 0 & \dots & 0 & 0 \\ 0 & a_3 & b_3 & c_3 & \dots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 0 & 0 & 0 & 0 & \dots & a_n & b_n \end{pmatrix}. \quad (1.50)$$

An augmented matrix is formed by two matrices with the same number of rows. For example, the follow system of linear equations

$$\begin{aligned} a_{11}u_1 + a_{12}u_2 + a_{13}u_3 &= b_1, \\ a_{21}u_1 + a_{22}u_2 + a_{23}u_3 &= b_2, \\ a_{31}u_1 + a_{32}u_2 + a_{33}u_3 &= b_3, \end{aligned} \quad (1.51)$$

can be written in a compact form in terms of matrices

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \quad (1.52)$$

or

$$\mathbf{A}\mathbf{u} = \mathbf{b}. \quad (1.53)$$

This can in turn be written as the following augmented form

$$[\mathbf{A}|\mathbf{b}] = \left( \begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array} \right). \quad (1.54)$$

The augmented form is widely used in Gauss-Jordan elimination and linear programming.

A lower (upper) triangular matrix is a square matrix with all the elements above (below) the diagonal entries being zeros. In general, a lower triangular matrix can be written as

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{12} & l_{22} & \dots & 0 \\ & & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix}, \quad (1.55)$$

while the upper triangular matrix can be written as

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ & & \ddots & \\ 0 & 0 & \dots & u_{nn} \end{pmatrix}. \quad (1.56)$$

Any  $n \times n$  square matrix  $\mathbf{A} = [a_{ij}]$  can be decomposed or factorized as a product of an  $\mathbf{L}$  and a  $\mathbf{U}$ , that is

$$\mathbf{A} = \mathbf{LU}, \quad (1.57)$$

though some decomposition is not unique because we have  $n^2 + n$  unknowns:  $n(n + 1)/2$  coefficients  $l_{ij}$  and  $n(n + 1)/2$  coefficients  $u_{ij}$ , but we can only provide  $n^2$  equations from the coefficients  $a_{ij}$ . Thus, there are  $n$  free parameters. The uniqueness of decomposition is often achieved by imposing either  $l_{ii} = 1$  or  $u_{ii} = 1$  where  $i = 1, 2, \dots, n$ .

Other LU variants include the LDU and LUP decompositions. An LDU decomposition can be written as

$$\mathbf{A} = \mathbf{LDU}, \quad (1.58)$$

where  $\mathbf{L}$  and  $\mathbf{U}$  are lower and upper matrices with all the diagonal entries being unity, and  $\mathbf{D}$  is a diagonal matrix. On the other hand, the LUP decomposition can be expressed as

$$\mathbf{A} = \mathbf{LUP}, \quad \text{or} \quad \mathbf{A} = \mathbf{PLU}, \quad (1.59)$$

where  $\mathbf{P}$  is a permutation matrix which is a square matrix and has exactly one entry 1 in each column and each row with 0's elsewhere. However, most numerical libraries and software use the following LUP decomposition

$$\mathbf{PA} = \mathbf{LU}. \quad (1.60)$$

which makes it easier to decompose some matrices. However, the requirement for LU decompositions is relatively strict. An invertible matrix  $\mathbf{A}$

has an LU decomposition provided that the determinants of all its diagonal minors or leading submatrices are not zeros.

A simpler way of decomposing a square matrix  $\mathbf{A}$  for solving a system of linear equations is to write

$$\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}, \quad (1.61)$$

where  $\mathbf{D}$  is a diagonal matrix.  $\mathbf{L}$  and  $\mathbf{U}$  are the strictly lower and upper triangular matrix without diagonal elements. This decomposition is much simpler to implement than the LU decomposition because there is no multiplication involved here.

**Example 1.3:** The following  $3 \times 3$  matrix

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

can be decomposed as  $\mathbf{A} = \mathbf{L}\mathbf{U}$ . That is

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 0 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix},$$

which becomes

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1 & 5 \\ 4 & -4 & 5 \\ 5 & 2 & -5 \end{pmatrix}. \end{aligned}$$

This leads to  $u_{11} = 2$ ,  $u_{12} = 1$  and  $u_{13} = 5$ . As  $l_{21}u_{11} = 4$ , so  $l_{21} = 4/u_{11} = 2$ . Similarly,  $l_{31} = 2.5$ . From  $l_{21}u_{12} + u_{22} = -4$ , we have  $u_{22} = -4 - 2 \times 1 = -6$ . From  $l_{21}u_{13} + u_{23} = 5$ , we have  $u_{23} = 5 - 2 \times 5 = -5$ .

Using  $l_{31}u_{12} + l_{32}u_{22} = 2$ , or  $2.5 \times 1 + l_{32} \times (-6) = 2$ , we get  $l_{32} = 1/12$ . Finally,  $l_{31}u_{13} + l_{32}u_{23} + u_{33} = -5$  gives  $u_{33} = -5 - 2.5 \times 5 - 1/12 \times (-5) = -205/12$ . Therefore, we now have

$$\begin{pmatrix} 2 & 1 & 5 \\ 4 & -4 & 5 \\ 5 & 2 & -5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 5/2 & 1/12 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 5 \\ 0 & -6 & -5 \\ 0 & 0 & -205/12 \end{pmatrix}.$$

The  $L+D+U$  decomposition can be written as

$$\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & -5 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 5 & 2 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 0 \end{pmatrix}.$$

## 1.4 Determinant and Inverse

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

$$\det(\mathbf{A}) = |\mathbf{A}| = \sum_{j=1}^n (-1)^{k+j} a_{kj} M_{kj}, \quad (1.62)$$

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

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \quad (1.63)$$

The determinants of matrices have the following properties:

$$|\alpha\mathbf{A}| = \alpha|\mathbf{A}|, \quad |\mathbf{A}^T| = |\mathbf{A}|, \quad |\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|, \quad (1.64)$$

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

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

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \dots + a_{nn}. \quad (1.65)$$

The rank of a matrix  $\mathbf{A}$  is the number of linearly independent vectors forming the matrix. Generally speaking, the rank of  $\mathbf{A}$  satisfies

$$\text{rank}(\mathbf{A}) \leq \min(m, n). \quad (1.66)$$

For an  $n \times n$  square matrix  $\mathbf{A}$ , it is nonsingular if  $\text{rank}(\mathbf{A}) = n$ .

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

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \quad (1.67)$$

More generally,

$$\mathbf{A}_l^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}_r^{-1} = \mathbf{I}, \quad (1.68)$$

where  $\mathbf{A}_l^{-1}$  is the left inverse while  $\mathbf{A}_r^{-1}$  is the right inverse. If  $\mathbf{A}_l^{-1} = \mathbf{A}_r^{-1}$ , we say that the matrix  $\mathbf{A}$  is invertible and its inverse is simply denoted by  $\mathbf{A}^{-1}$ . It is worth noting that the unit matrix  $\mathbf{I}$  has the same size as  $\mathbf{A}$ .

The inverse of a square matrix exists if and only if  $\mathbf{A}$  is nonsingular or  $\det(\mathbf{A}) \neq 0$ . From the basic definitions, it is straightforward to prove that the inverse of a matrix has the following properties

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}, \quad (\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T, \quad (1.69)$$

and

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \quad (1.70)$$

The inverse of a lower (upper) triangular matrix is also a lower (upper) triangular matrix. The inverse of a diagonal matrix

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & d_n \end{pmatrix}, \quad (1.71)$$

can simply be written as

$$\mathbf{D}^{-1} = \begin{pmatrix} 1/d_1 & 0 & \dots & 0 \\ 0 & 1/d_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1/d_n \end{pmatrix}, \quad (1.72)$$

where  $d_i \neq 0$ . If any of these elements  $d_i$  is zero, then the diagonal matrix is not invertible as it becomes singular. For a  $2 \times 2$  matrix, its inverse is simply

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \quad (1.73)$$

**Example 1.4:** For two matrices,

$$\mathbf{A} = \begin{pmatrix} 4 & 5 & 0 \\ -2 & 2 & 5 \\ 2 & -3 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2 & 3 \\ 0 & -2 \\ 5 & 2 \end{pmatrix},$$

their transpose matrices are

$$\mathbf{A}^T = \begin{pmatrix} 4 & -2 & 2 \\ 5 & 2 & -3 \\ 0 & 5 & 1 \end{pmatrix}, \quad \mathbf{B}^T = \begin{pmatrix} 2 & 0 & 5 \\ 3 & -2 & 2 \end{pmatrix}.$$

Let  $\mathbf{D} = \mathbf{AB}$  be their product; we have

$$\mathbf{AB} = \mathbf{D} = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \\ D_{31} & D_{32} \end{pmatrix}.$$

The first two entries are

$$D_{11} = \sum_{j=1}^3 A_{1j}B_{j1} = 2 \times 4 + 5 \times 0 + 0 \times 5 = 8,$$

and

$$D_{12} = \sum_{j=1}^3 A_{1j}B_{j2} = 4 \times 3 + 5 \times (-2) + 0 \times 2 = 2.$$

Similarly, the other entries are:

$$D_{21} = 21, \quad D_{22} = 0, \quad D_{31} = 9, \quad D_{33} = 14.$$

Therefore, we get

$$\mathbf{AB} = \begin{pmatrix} 4 & 5 & 0 \\ -2 & 2 & 5 \\ 2 & -3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 0 & -2 \\ 5 & 2 \end{pmatrix} = \mathbf{D} = \begin{pmatrix} 8 & 2 \\ 21 & 0 \\ 9 & 14 \end{pmatrix}.$$

However, the product  $\mathbf{BA}$  does not exist, though

$$\mathbf{B}^T \mathbf{A}^T = \begin{pmatrix} 8 & 21 & 9 \\ 2 & 0 & 14 \end{pmatrix} = \mathbf{D}^T = (\mathbf{AB})^T.$$

The inverse of  $\mathbf{A}$  is

$$\mathbf{A}^{-1} = \frac{1}{128} \begin{pmatrix} 17 & -5 & 25 \\ 12 & 4 & -20 \\ 2 & 22 & 18 \end{pmatrix},$$

and the determinant of  $\mathbf{A}$  is

$$\det(\mathbf{A}) = 128.$$

It is straightforward to verify that

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For example, the first entry is obtained by

$$\sum_{j=1}^3 A_{1j}A_{j1}^{-1} = 4 \times \frac{17}{128} + 5 \times \frac{12}{128} + 0 \times \frac{2}{128} = 1.$$

Other entries can be verified similarly. Finally, the trace of  $\mathbf{A}$  is

$$\text{tr}(\mathbf{A}) = A_{11} + A_{22} + A_{33} = 4 + 2 + 1 = 7.$$

The algorithmic complexity of most algorithms for obtaining the inverse of a general square matrix is  $O(n^3)$ . That is why most modern algorithms try to avoid the direct inverse of a large matrix. Solution of a large matrix system is instead carried out either by partial inverse via decomposition or by iteration (or a combination of these two methods). If the matrix

can be decomposed into triangular matrices either by LU factorization or direction decomposition, the aim is then to invert a triangular matrix, which is simpler and more efficient.

For a triangular matrix, the inverse can be obtained using algorithms of  $O(n^2)$  complexity. Similarly, the solution of a linear system with a lower (upper) triangular matrix  $\mathbf{A}$  can be obtained by forward (back) substitutions. In general, for a lower triangular matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & 0 & \dots & 0 \\ \alpha_{12} & \alpha_{22} & \dots & 0 \\ & & \ddots & \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{pmatrix}, \quad (1.74)$$

the forward substitutions for the system  $\mathbf{A}\mathbf{u} = \mathbf{b}$  can be carried out as follows:

$$\begin{aligned} u_1 &= \frac{b_1}{\alpha_{11}}, \\ u_2 &= \frac{1}{\alpha_{22}}(b_2 - \alpha_{21}u_1), \\ u_i &= \frac{1}{\alpha_{ii}}\left(b_i - \sum_{j=1}^{i-1} \alpha_{ij}u_j\right), \end{aligned} \quad (1.75)$$

where  $i = 2, \dots, n$ . We see that it takes 1 division to get  $u_1$ , 3 floating point calculations to get  $u_2$ , and  $(2i - 1)$  to get  $u_i$ . So the total algorithmic complexity is  $O(1 + 3 + \dots + (2n - 1)) = O(n^2)$ . Similar arguments apply to the upper triangular systems.

The inverse  $\mathbf{A}^{-1}$  of a lower triangular matrix can in general be written as

$$\mathbf{A}^{-1} = \begin{pmatrix} \beta_{11} & 0 & \dots & 0 \\ \beta_{12} & \beta_{22} & \dots & 0 \\ & & \ddots & \\ \beta_{n1} & \beta_{n2} & \dots & \beta_{nn} \end{pmatrix} = \mathbf{B} = (\mathbf{B}_1 \ \mathbf{B}_2 \ \dots \ \mathbf{B}_n), \quad (1.76)$$

where  $\mathbf{B}_j$  are the  $j$ -th column vector of  $\mathbf{B}$ . The inverse must satisfy  $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$  or

$$\mathbf{A}(\mathbf{B}_1 \ \mathbf{B}_2 \ \dots \ \mathbf{B}_n) = \mathbf{I} = (\mathbf{e}_1 \ \mathbf{e}_2 \ \dots \ \mathbf{e}_n), \quad (1.77)$$

where  $\mathbf{e}_j$  is the  $j$ -th unit vector of size  $n$  with the  $j$ -th element being 1 and all other elements being zero. That is  $\mathbf{e}_j^T = (0 \ 0 \ \dots \ 1 \ 0 \ \dots \ 0)$ . In order to obtain  $\mathbf{B}$ , we have to solve  $n$  linear systems

$$\mathbf{A}\mathbf{B}_1 = \mathbf{e}_1, \quad \mathbf{A}\mathbf{B}_2 = \mathbf{e}_2, \quad \dots, \quad \mathbf{A}\mathbf{B}_n = \mathbf{e}_n. \quad (1.78)$$

As  $\mathbf{A}$  is a lower triangular matrix, the solution of  $\mathbf{A}\mathbf{B}_j = \mathbf{e}_j$  can easily be obtained by direct forward substitutions discussed earlier in this section.

## 1.5 Matrix Exponential

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

$$e^{\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!} \mathbf{A}^n = \mathbf{I} + \mathbf{A} + \frac{1}{2} \mathbf{A}^2 + \dots, \quad (1.79)$$

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

---

**Example 1.5:** For a simple matrix

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

its exponential is simply

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

For a more complicated matrix

$$\mathbf{B} = \begin{pmatrix} t & a \\ a & t \end{pmatrix},$$

we have

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

---

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

$$e^{t\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!} (t\mathbf{A})^n = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2} \mathbf{A}^2 + \dots, \quad (1.80)$$

$$\ln(\mathbf{I} + \mathbf{A}) \equiv \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n!} \mathbf{A}^n = \mathbf{A} - \frac{1}{2} \mathbf{A}^2 + \frac{1}{3} \mathbf{A}^3 + \dots, \quad (1.81)$$

$$e^{\mathbf{A}} e^{\mathbf{B}} = e^{\mathbf{A} + \mathbf{B}} \quad (\text{if } \mathbf{AB} = \mathbf{BA}), \quad (1.82)$$

$$\frac{d}{dt} e^{t\mathbf{A}} = \mathbf{A} e^{t\mathbf{A}} = e^{t\mathbf{A}} \mathbf{A}, \quad (1.83)$$

$$(e^{\mathbf{A}})^{-1} = e^{-\mathbf{A}}, \quad \det(e^{\mathbf{A}}) = e^{\text{tr} \mathbf{A}}. \quad (1.84)$$

## 1.6 Hermitian and Quadratic Forms

The matrices we have discussed so far are real matrices because all their elements are real. In general, the entries or elements of a matrix can be complex numbers, and the matrix becomes a complex matrix. For a matrix  $\mathbf{A}$ , its complex conjugate  $\mathbf{A}^*$  is obtained by taking the complex conjugate of each of its elements. The Hermitian conjugate  $\mathbf{A}^\dagger$  is obtained by taking the transpose of its complex conjugate matrix. That is to say, for

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots \\ a_{21} & a_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix}, \quad (1.85)$$

we have

$$\mathbf{A}^* = \begin{pmatrix} a_{11}^* & a_{12}^* & \dots \\ a_{21}^* & a_{22}^* & \dots \\ \dots & \dots & \dots \end{pmatrix}, \quad (1.86)$$

and

$$\mathbf{A}^\dagger = (\mathbf{A}^*)^T = (\mathbf{A}^T)^* = \begin{pmatrix} a_{11}^* & a_{21}^* & \dots \\ a_{12}^* & a_{22}^* & \dots \\ \dots & \dots & \dots \end{pmatrix}. \quad (1.87)$$

A square matrix  $\mathbf{A}$  is called orthogonal if and only if  $\mathbf{A}^{-1} = \mathbf{A}^T$ . If a square matrix  $\mathbf{A}$  satisfies  $\mathbf{A}^* = \mathbf{A}$ , it is called an Hermitian matrix. It is an anti-Hermitian matrix if  $\mathbf{A}^* = -\mathbf{A}$ . If the Hermitian matrix of a square matrix  $\mathbf{A}$  is equal to the inverse of the matrix (or  $\mathbf{A}^\dagger = \mathbf{A}^{-1}$ ), it is called a unitary matrix.

**Example 1.6:** For a complex matrix

$$\mathbf{A} = \begin{pmatrix} 2 + 3i\pi & 1 + 9i & 0 \\ e^{i\pi} & -2i & i \sin \theta \end{pmatrix},$$

its complex conjugate  $\mathbf{A}^*$  is

$$\mathbf{A}^* = \begin{pmatrix} 2 - 3i\pi & 1 - 9i & 0 \\ e^{-i\pi} & 2i & -i \sin \theta \end{pmatrix}.$$

The Hermitian conjugate of  $\mathbf{A}$  is

$$\mathbf{A}^\dagger = \begin{pmatrix} 2 - 3i\pi & e^{-i\pi} \\ 1 - 9i & 2i \\ 0 & -i \sin \theta \end{pmatrix} = (\mathbf{A}^*)^T.$$

For the rotation matrix

$$\mathbf{A} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix},$$

its inverse and transpose are

$$\mathbf{A}^{-1} = \frac{1}{\cos^2 \theta + \sin^2 \theta} \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

and

$$\mathbf{A}^T = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

Since  $\cos^2 \theta + \sin^2 \theta = 1$ , we have  $\mathbf{A}^T = \mathbf{A}^{-1}$ . Therefore, the original rotation matrix  $\mathbf{A}$  is orthogonal.

A very useful concept in computational mathematics and computing is quadratic forms. For a real vector  $\mathbf{q}^T = (q_1, q_2, q_3, \dots, q_n)$  and a real symmetric square matrix  $\mathbf{A}$ , a quadratic form  $\psi(\mathbf{q})$  is a scalar function defined by

$$\psi(\mathbf{q}) = \mathbf{q}^T \mathbf{A} \mathbf{q} = (q_1 \ q_2 \ \dots \ q_n) \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix}, \quad (1.88)$$

which can be written as

$$\psi(\mathbf{q}) = \sum_{i=1}^n \sum_{j=1}^n q_i A_{ij} q_j. \quad (1.89)$$

Since  $\psi$  is a scalar, it should be independent of the coordinates.

In the case of a square matrix  $\mathbf{A}$ ,  $\psi$  might be more easily evaluated in certain intrinsic coordinates  $Q_1, Q_2, \dots, Q_n$ . An important result concerning the quadratic form is that it can always be written through appropriate transformations as

$$\psi(\mathbf{q}) = \sum_{i=1}^n \lambda_i Q_i^2 = \lambda_1 Q_1^2 + \lambda_2 Q_2^2 + \dots + \lambda_n Q_n^2, \quad (1.90)$$

where  $\lambda_i$  are the eigenvalues of the matrix  $\mathbf{A}$  determined by

$$\det |\mathbf{A} - \lambda \mathbf{I}| = 0, \quad (1.91)$$

and  $Q_i$  are the intrinsic components along directions of the eigenvectors in this case.

The natural extension of quadratic forms is the Hermitian form which is the quadratic form for a complex Hermitian matrix  $\mathbf{A}$ . Furthermore, the matrix  $\mathbf{A}$  can consist of linear operators and functionals in addition to numbers.

---

**Example 1.7:** For a vector  $\mathbf{q} = (q_1, q_2)$  and the square matrix

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

we have a quadratic form

$$\psi(\mathbf{q}) = (q_1 \ q_2) \begin{pmatrix} 2 & -5 \\ -5 & 2 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = 2q_1^2 - 10q_1q_2 + 2q_2^2.$$

The eigenvalues of the matrix  $\mathbf{A}$  is determined by

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

whose solutions are  $\lambda_1 = 7$  and  $\lambda_2 = -3$  (see the next section for further details). Their corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}.$$

We can see that  $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$ , which means that these two eigenvectors are orthogonal. Writing the quadratic form in terms of the intrinsic coordinates, we have

$$\psi(\mathbf{q}) = 7Q_1^2 - 3Q_2^2.$$

Furthermore, if we assume  $\psi(\mathbf{q}) = 1$  as a simple constraint, then the equation  $7Q_1^2 - 3Q_2^2 = 1$  corresponds to a hyperbola.

---

## 1.7 Eigenvalues and Eigenvectors

The eigenvalues  $\lambda$  of any  $n \times n$  square matrix  $\mathbf{A}$  is determined by

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \tag{1.92}$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0. \tag{1.93}$$

where  $\mathbf{I}$  is a unitary matrix with the same size as  $\mathbf{A}$ . Any non-trivial solution requires that

$$\det |\mathbf{A} - \lambda\mathbf{I}| = 0, \tag{1.94}$$

or

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} = 0, \quad (1.95)$$

which again can be written as a polynomial

$$\lambda^n + \alpha_{n-1}\lambda^{n-1} + \dots + \alpha_0 = (\lambda - \lambda_1)\dots(\lambda - \lambda_n) = 0, \quad (1.96)$$

where  $\lambda_i$  are the eigenvalues which could be complex numbers. In general, the determinant is zero, which leads to a polynomial of order  $n$  in  $\lambda$ . For each eigenvalue  $\lambda$ , there is a corresponding eigenvector  $\mathbf{u}$  whose direction can be uniquely determined. However, the length of the eigenvector is not unique because any non-zero multiple of  $\mathbf{u}$  will also satisfy equation (1.92), and thus can be considered as an eigenvector. For this reason, it is usually necessary to apply additional conditions by setting the length as unity, and subsequently the eigenvector becomes a unit eigenvector.

Generally speaking, a real  $n \times n$  matrix  $\mathbf{A}$  has  $n$  eigenvalues  $\lambda_i$  ( $i = 1, 2, \dots, n$ ), however, these eigenvalues are not necessarily distinct. If the real matrix is symmetric, that is to say  $\mathbf{A}^T = \mathbf{A}$ , then the matrix has  $n$  distinct eigenvectors, and all the eigenvalues are real numbers.

The eigenvalues  $\lambda_i$  are related to the trace and determinant of the matrix

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = \lambda_1 + \lambda_2 + \dots + \lambda_n = \sum_{i=1}^n \lambda_i, \quad (1.97)$$

and

$$\det(\mathbf{A}) = |\mathbf{A}| = \prod_{i=1}^n \lambda_i. \quad (1.98)$$

**Example 1.8:** *The eigenvalues of the square matrix*

$$\mathbf{A} = \begin{pmatrix} 4 & 9 \\ 2 & -3 \end{pmatrix},$$

can be obtained by solving

$$\begin{vmatrix} 4 - \lambda & 9 \\ 2 & -3 - \lambda \end{vmatrix} = 0.$$

We have

$$(4 - \lambda)(-3 - \lambda) - 18 = (\lambda - 6)(\lambda + 5) = 0.$$

Thus, the eigenvalues are  $\lambda = 6$  and  $\lambda = -5$ . Let  $\mathbf{v} = (v_1 \ v_2)^T$  be the eigenvector; we have for  $\lambda = 6$

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{pmatrix} -2 & 9 \\ 2 & -9 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0,$$

which means that

$$-2v_1 + 9v_2 = 0, \quad 2v_1 - 9v_2 = 0.$$

These two equations are virtually the same (not linearly independent), so the solution is

$$v_1 = \frac{9}{2}v_2.$$

Any vector parallel to  $\mathbf{v}$  is also an eigenvector. In order to get a unique eigenvector, we have to impose an extra requirement, that is, the length of the vector is unity. We now have

$$v_1^2 + v_2^2 = 1,$$

or

$$\left(\frac{9v_2}{2}\right)^2 + v_2^2 = 1,$$

which gives  $v_2 = \pm 2/\sqrt{85}$ , and  $v_1 = \pm 9/\sqrt{85}$ . As these two vectors are in opposite directions, we can choose any of the two directions. So the eigenvector for the eigenvalue  $\lambda = 6$  is

$$\mathbf{v} = \begin{pmatrix} 9/\sqrt{85} \\ 2/\sqrt{85} \end{pmatrix}.$$

Similarly, the corresponding eigenvector for the eigenvalue  $\lambda = -5$  is  $\mathbf{v} = (-\sqrt{2}/2 \ \sqrt{2}/2)^T$ .

Furthermore, the trace and determinant of  $\mathbf{A}$  are

$$\text{tr}(\mathbf{A}) = 4 + (-3) = 1, \quad \det(\mathbf{A}) = 4 \times (-3) - 2 \times 9 = -30.$$

The sum of the eigenvalues is

$$\sum_{i=1}^2 \lambda_i = 6 + (-5) = 1 = \text{tr}(\mathbf{A}),$$

while the product of the eigenvalues is

$$\prod_{i=1}^2 \lambda_i = 6 \times (-5) = -30 = \det(\mathbf{A}).$$

For any real square matrix  $\mathbf{A}$  with the eigenvalues  $\lambda_i = \text{eig}(\mathbf{A})$ , the eigenvalues of  $\alpha\mathbf{A}$  are  $\alpha\lambda_i$  where  $\alpha \neq 0 \in \Re$ . This property becomes handy when rescaling the matrices in some iteration formulae so that the rescaled scheme becomes more stable. This is also the major reason why the pivoting and removing/rescaling of exceptionally large elements works.

## 1.8 Definiteness of Matrices

A square symmetric matrix  $\mathbf{A}$  is said to be positive definite if all its eigenvalues are strictly positive ( $\lambda_i > 0$  where  $i = 1, 2, \dots, n$ ). By multiplying (1.92) by  $\mathbf{u}^T$ , we have

$$\mathbf{u}^T \mathbf{A} \mathbf{u} = \mathbf{u}^T \lambda \mathbf{u} = \lambda \mathbf{u}^T \mathbf{u}, \quad (1.99)$$

which leads to

$$\lambda = \frac{\mathbf{u}^T \mathbf{A} \mathbf{u}}{\mathbf{u}^T \mathbf{u}}. \quad (1.100)$$

This means that

$$\mathbf{u}^T \mathbf{A} \mathbf{u} > 0, \quad \text{if } \lambda > 0. \quad (1.101)$$

In fact, for any vector  $\mathbf{v}$ , the following relationship holds

$$\mathbf{v}^T \mathbf{A} \mathbf{v} > 0. \quad (1.102)$$

Since  $\mathbf{v}$  can be a unit vector, thus all the diagonal elements of  $\mathbf{A}$  should be strictly positive as well. If all the eigenvalues are non-negative or  $\lambda_i \geq 0$ , then the matrix is called positive semi-definite. In general, an indefinite matrix can have both positive and negative eigenvalues.

The inverse of a positive definite matrix is also positive definite. For a linear system  $\mathbf{A} \mathbf{u} = \mathbf{f}$  where  $\mathbf{f}$  is a known column vector, if  $\mathbf{A}$  is positive definite, then the system can be solved more efficiently by matrix decomposition methods.

**Example 1.9:** In general, a  $2 \times 2$  symmetric matrix  $\mathbf{A}$

$$\mathbf{A} = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix},$$

is positive definite if

$$\alpha u_1^2 + 2\beta u_1 u_2 + \gamma u_2^2 > 0,$$

for all  $\mathbf{u} = (u_1, u_2)^T \neq 0$ . The inverse of  $\mathbf{A}$  is

$$\mathbf{A}^{-1} = \frac{1}{\alpha\gamma - \beta^2} \begin{pmatrix} \gamma & -\beta \\ -\beta & \alpha \end{pmatrix},$$

which is also positive definite.

As the eigenvalues of

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

are  $\lambda = 3, -1$ , the matrix is indefinite. For another matrix

$$\mathbf{B} = \begin{pmatrix} 4 & 6 \\ 6 & 20 \end{pmatrix},$$

we can find its eigenvalues using a similar method as discussed earlier, and the eigenvalues are  $\lambda = 2, 22$ . So matrix  $\mathbf{B}$  is positive definite. The inverse of  $\mathbf{B}$

$$\mathbf{B}^{-1} = \frac{1}{44} \begin{pmatrix} 20 & -6 \\ -6 & 4 \end{pmatrix},$$

is also positive definite because  $\mathbf{B}^{-1}$  has two eigenvalues:  $\lambda = 1/2, 1/22$ .

---