Mathematical background

Jianxin Wu

LAMDA Group
National Key Lab for Novel Software Technology
Nanjing University, China
wujx2001@gmail.com

February 11, 2020

Contents

This chapter provides a brief review of the basic mathematical background that is required for understanding this book. Most of the contents in this chapter can be found in standard undergraduate math textbooks, hence details such as proofs will be omitted.

This book also requires some mathematics that are a little bit more advanced. We will provide the statements in this chapter, but detailed proofs are again omitted.

1 Linear algebra

We will not consider complex numbers in this book. Hence, what we will deal with are all real numbers.

Scalar. We use $\mathbb{R}$ to denote the set of real numbers. A real number $x \in \mathbb{R}$ is also called a scalar.

Vector. A sequence of real numbers form a vector. We use bold face letters to denote vectors, e.g., $\boldsymbol{x} \in \mathbb{R}^d$ is a vector formed by a sequence of d real numbers. We use

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$$

to indicate that $\boldsymbol{x}$ is formed by d numbers in a *column* shape, and the i-th number in the sequence is a scalar x_i, i.e., [1]

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} . \tag{1}$$

d is called the length (or dimensionality, or size) of the vector, and the vector is called a d-dimensional one. We use $\mathbf{1}_d$ and $\mathbf{0}_d$ to denote d-dimensional vectors whose elements are all 1 and all 0, respectively. When the vector size is obvious from its context, we simply write $\mathbf{1}$ or $\mathbf{0}$.

1.1 Inner product, norm, distance, and orthogonality

The inner product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $\boldsymbol{x}^T\boldsymbol{y}$ (or $\boldsymbol{x}\cdot\boldsymbol{y}$, or $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, or $\boldsymbol{x}'\boldsymbol{y}$, or $\boldsymbol{x}^t\boldsymbol{y}$; in this book, we will use the notation $\boldsymbol{x}^T\boldsymbol{y}$). It is also called the dot product. The dot product of two d-dimensional vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)^T$ is defined as

$$\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{d} x_i y_i \, . \tag{2}$$

[1] The ^{T} superscript means the transpose of a matrix, which will be defined soon.

Hence, the inner product is a scalar, and we obviously have

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}\,. \tag{3}$$

The above fact will sometimes help us in this book—e.g., making transformations:

$$(\boldsymbol{x}^T\boldsymbol{y})\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{x}^T\boldsymbol{y}) = \boldsymbol{z}\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{z}\boldsymbol{y}^T\boldsymbol{x} = (\boldsymbol{z}\boldsymbol{y}^T)\boldsymbol{x}\,, \tag{4}$$

and so on.

The norm of a vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|$, and defined by

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}\,. \tag{5}$$

Other types of vector norms are available. The specific form in Equation 5 is called the ℓ_2 norm. It is also called the *length* of $\boldsymbol{x}$ in some cases. Note that the norm $\|\boldsymbol{x}\|$ and the squared norm $\boldsymbol{x}^T\boldsymbol{x}$ are always non-negative for any $\boldsymbol{x} \in \mathbb{R}^d$.

A vector whose length is 1 is called a unit vector. We usually say a unit vector determines a *direction*. End points of unit vectors reside on the surface of the unit hypersphere in the d-dimensional space whose center is the zero vector $\mathbf{0}$ and radius is 1. A ray from the center to any unit vector uniquely determines a direction in that space, and vice versa. When $\boldsymbol{x} = c\boldsymbol{y}$ and $c > 0$, we say the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are in the same direction.

The distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $\|\boldsymbol{x} - \boldsymbol{y}\|$. A frequently used fact is about the squared distance:

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = (\boldsymbol{x} - \boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\boldsymbol{x}^T\boldsymbol{y}\,. \tag{6}$$

The above equality utilizes the facts that $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x}$ and $\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}$.

1.2 Angle and inequality

If $\boldsymbol{x}^T\boldsymbol{y} = 0$, we say the two vectors are orthogonal, or perpendicular, also denoted by $\boldsymbol{x} \perp \boldsymbol{y}$. From the geometry, we know the angle between these two vectors is $90°$ or $\frac{\pi}{2}$.

Let the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ be denoted by θ ($0 \le \theta \le 180°$); then

$$\boldsymbol{x}^T\boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta\,. \tag{7}$$

The above equation in fact defines the angle as

$$\theta = \arccos\left(\frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\right)\,. \tag{8}$$

Because $-1 \le \cos\theta \le 1$ for any θ, these equations also tell us

$$\boldsymbol{x}^T\boldsymbol{y} \le |\boldsymbol{x}^T\boldsymbol{y}| \le \|\boldsymbol{x}\|\|\boldsymbol{y}\|\,. \tag{9}$$

If we expand the vector form of this inequality and take the square of both sides, it appears as

$$\left(\sum_{i=1}^{d} x_i y_i\right)^2 \le \left(\sum_{i=1}^{d} x_i^2\right)\left(\sum_{i=1}^{d} y_i^2\right)\,, \tag{10}$$

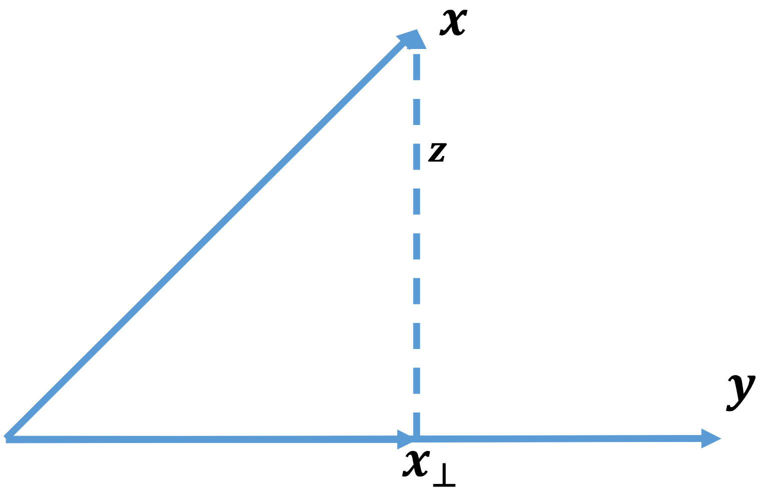


Figure 1: Illustration of vector projection.

which is the Cauchy–Schwarz inequality.[2] The equality holds if and only if there is a constant $c \in \mathbb{R}$ such that $x_i = cy_i$ for all $1 \le i \le d$. In the vector form, the equality condition is equivalent to $\boldsymbol{x} = c\boldsymbol{y}$ for some constant c.

This inequality (and the equality condition) can be extended to integrals:

$$\left(\int f(x)g(x)\,\mathrm{d}x\right)^2 \le \left(\int f^2(x)\,\mathrm{d}x\right)\left(\int g^2(x)\,\mathrm{d}x\right)\,, \tag{11}$$

assuming all integrals exist, in which $f^2(x)$ means $f(x)f(x)$.

1.3 Vector projection

Sometimes we need to compute the projection of one vector onto another. As illustrated in Figure 1, $\boldsymbol{x}$ is projected onto $\boldsymbol{y}$ (which must be non-zero). Hence, $\boldsymbol{x}$ is decomposed as

$$\boldsymbol{x} = \boldsymbol{x}_\perp + \boldsymbol{z}\,,$$

where $\boldsymbol{x}_\perp$ is the projected vector, and $\boldsymbol{z}$ can be considered as the residue (or error) of the projection. Note that $\boldsymbol{x}_\perp \perp \boldsymbol{z}$.

In order to determine $\boldsymbol{x}_\perp$, we take two steps: to *find its direction and norm separately*, and this trick will be useful in some other scenarios in this book, too.

For any non-zero vector $\boldsymbol{x}$, its norm is $\|\boldsymbol{x}\|$. Since $\boldsymbol{x} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\|\boldsymbol{x}\|$, the vector $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ is in the same direction as $\boldsymbol{x}$ and it is also a unit vector. Hence, $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ is the

[2] The two names in this inequality are Augustin-Louis Cauchy, the famous French mathematician who first published this inequality and Karl Hermann Amandus Schwarz, a German mathematician. The integral form generalization of this inequality was by Viktor Yakovlevich Bunyakovsky, a Ukrainian/Russian mathematician.

direction of $\boldsymbol{x}$. The combination of norm and direction uniquely determines any vector. The norm alone determines the zero vector.

The direction of $\boldsymbol{y}$ is $\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}$. It is obvious that the direction of $\boldsymbol{x}_\perp$ is $\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}$ if the angle θ between $\boldsymbol{x}$ and $\boldsymbol{y}$ is acute ($< 90°$), as illustrated in Figure 1. The norm of $\boldsymbol{x}_\perp$ is also simple:

$$\|\boldsymbol{x}_\perp\| = \|\boldsymbol{x}\| \cos\theta = \|\boldsymbol{x}\| \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{y}\|}\,. \tag{12}$$

Hence, the projection $\boldsymbol{x}_\perp$ is

$$\boldsymbol{x}_\perp = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{y}\|}\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}}\boldsymbol{y}\,. \tag{13}$$

Equation 13 is derived assuming θ is acute. However, it is easy to verify that this equation is correct too when the angle is right ($= 90°$), obtuse ($> 90°$), or straight ($= 180°$). The term $\frac{\boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}}$ (which is a scalar) is called the projected value, and $\frac{\boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}}\boldsymbol{y}$ is the projected vector, which is also denoted by $\text{proj}_{\boldsymbol{y}}\boldsymbol{x}$.

Vector projection is very useful in this book. For example, let $\boldsymbol{y} = (2, 1)$ and $\boldsymbol{x} = (1, 1)$. The direction of $\boldsymbol{y}$ specifies all the points that possess this property: its first dimension is twice its second dimension. Using Equation 13, we obtain $\text{proj}_{\boldsymbol{y}}\boldsymbol{x} = (1.2, 0.6)$, which also exhibits the same property. We may treat $\text{proj}_{\boldsymbol{y}}\boldsymbol{x}$ as the best approximation of $\boldsymbol{x}$ that satisfies the property specified in $\boldsymbol{y}$. The residue of this approximation $\boldsymbol{z} = \boldsymbol{x} - \text{proj}_{\boldsymbol{y}}\boldsymbol{x} = (-0.2, 0.4)$ does not satisfy this property and can be considered as noise or error in certain applications.

1.4 Basics of matrices

An $m \times n$ matrix contains mn numbers organized in m rows and n columns, and we use x_{ij} (or $x_{i,j}$) to denote the element at the i-th row and j-th column in a matrix X, that is,

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix}\,. \tag{14}$$

We also use $[X]_{ij}$ to refer to the element at the i-th row and j-th column in a matrix X.

There are a few special cases. When $m = n$, we call the matrix a square matrix. When $n = 1$, the matrix contains only one column, and we call it a column matrix, or a column vector, or simply a vector. When $m = 1$, we call it a row matrix, or a row vector. Note that when we say $\boldsymbol{x}$ is a vector, we mean a column vector if not otherwise specified. That is, when we write $\boldsymbol{x} = (1, 2, 3)^T$, we are referring to a column matrix $\left[\begin{smallmatrix}1\\2\\3\end{smallmatrix}\right]$.

There are also a few special cases within square matrices that are worth noting. In a square matrix X of size $n \times n$, the diagonal entries refer to those elements x_{ij} in X satisfying $i = j$. If $x_{ij} = 0$ whenever $i \neq j$ (i.e., when non-diagonal entries are all 0), we say X is a diagonal matrix. The unit matrix is a special diagonal matrix, whose diagonal entires are all 1. A unit matrix is usually denoted by I (when the size of it can be inferred from the context) or I_n (indicating the size is $n \times n$).

Following the Matlab convention, we use

$$X = \operatorname{diag}(x_{11}, x_{22}, \ldots, x_{nn})$$

to denote an $n \times n$ diagonal matrix whose diagonal entries are $x_{11}, x_{22}, \ldots, x_{nn}$, sequentially. Similarly, for an $n \times n$ square matrix X, $\operatorname{diag}(X)$ is a vector $(x_{11}, x_{22}, \ldots, x_{nn})^T$.

The transpose of a matrix X is denoted by X^T, and is defined by

$$[X^T]_{ji} = x_{ij} \,.$$

X^T has size $n \times m$ if X is $m \times n$. When X is square, X^T and X have the same size. If in addition $X^T = X$, then we say X is a symmetric matrix.

1.5 Matrix multiplication

Addition and subtraction can be applied to matrices with the same size. Let X and Y be two matrices with size $m \times n$, then

$$[X + Y]_{ij} = x_{ij} + y_{ij} \,, \tag{15}$$

$$[X - Y]_{ij} = x_{ij} - y_{ij} \,, \tag{16}$$

for any $1 \le i \le m$, $1 \le j \le n$. For any matrix X and a scalar c, the scalar multiplication cX is defined by

$$[cX]_{ij} = cx_{ij} \,.$$

Not any two matrices can be multiplied. The multiplication XY exists (i.e., is well defined) if and only if the number of columns in X equals the number of rows in Y—i.e., when there are positive integers m, n, and p such that the size of X is $m \times n$ and the size of Y is $n \times p$. The product XY is a matrix of size $m \times p$, and is defined by

$$[XY]_{ij} = \sum_{k=1}^{n} x_{ik} y_{kj} \,. \tag{17}$$

When XY is well defined, we always have

$$(XY)^T = Y^T X^T \,.$$

Note that YX does not necessarily exist when XY exists. Even if both XY and YX are well-defined, $XY \neq YX$ except in a few special cases. However, for any matrix X, both XX^T and $X^T X$ exist and both are symmetric matrices.

Let $\boldsymbol{x} = (x_1, x_2, \dots, x_m)^T$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_p)^T$ be two vectors. Treating vectors as special matrices, the dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}^T$ satisfy the multiplication constraint, such that $\boldsymbol{x}\boldsymbol{y}^T$ always exists for any $\boldsymbol{x}$ and $\boldsymbol{y}$. We call $\boldsymbol{x}\boldsymbol{y}^T$ the outer product between $\boldsymbol{x}$ and $\boldsymbol{y}$. This outer product is an $m \times p$ matrix, and $[\boldsymbol{x}\boldsymbol{y}^T]_{ij} = x_i y_j$. Note that in general $\boldsymbol{x}\boldsymbol{y}^T \neq \boldsymbol{y}\boldsymbol{x}^T$.

The block matrix representation is sometimes useful. Let $\boldsymbol{x}_{i:}$ denote the i-th row of X (of size $1 \times n$), and $\boldsymbol{x}_{:i}$ the i-th column (of size $m \times 1$); we can write X as either a column format

$$X = \left[\begin{array}{c} \boldsymbol{x}_{1:} \\ \hline \boldsymbol{x}_{2:} \\ \hline \vdots \\ \hline \boldsymbol{x}_{m:} \end{array}\right], \tag{18}$$

or a row format

$$X = [\boldsymbol{x}_{:1} | \boldsymbol{x}_{:2} | \dots | \boldsymbol{x}_{:n}] \,. \tag{19}$$

Using the block matrix representation, we have

$$XY = [\boldsymbol{x}_{:1} | \boldsymbol{x}_{:2} | \dots | \boldsymbol{x}_{:n}] \left[\begin{array}{c} \boldsymbol{y}_{1:} \\ \hline \boldsymbol{y}_{2:} \\ \hline \vdots \\ \hline \boldsymbol{y}_{n:} \end{array}\right] = \sum_{i=1}^{n} \boldsymbol{x}_{:i} \boldsymbol{y}_{i:} \,. \tag{20}$$

That is, the product XY is the summation of n outer products (between $\boldsymbol{x}_{:i}$ and $\boldsymbol{y}_{i:}^T$, which are both column vectors), X and Y^T have the same number of columns. If we compute the outer product of their corresponding columns, we get n $m \times p$ matrices. The summation of these matrices equals XY.

Similarly, we also have

$$XY = \left[\begin{array}{c} \boldsymbol{x}_{1:} \\ \hline \boldsymbol{x}_{2:} \\ \hline \vdots \\ \hline \boldsymbol{x}_{m:} \end{array}\right] [\boldsymbol{y}_{:1} | \boldsymbol{y}_{:2} | \dots | \boldsymbol{y}_{:p}] \,. \tag{21}$$

This (block) outer product tells us $[XY]_{ij} = \boldsymbol{x}_{i:} \boldsymbol{y}_{:j}$, which is exactly Equation 17.

For a square matrix X and a natural number k, the k-th power of X is well defined, as

$$X^k = \underbrace{XX\dots X}_{k \text{ times}} \,.$$

1.6 Determinant and inverse of a square matrix

There are many ways to define the determinant of a square matrix, and we adopt Laplace's formula to define it recursively. The determinant of X is usually denoted by $\det(X)$ or simply $|X|$, and is a scalar. Note that although the $|\cdot|$ symbol looks like the absolute value operator, its meaning is different. The determinant could be positive, zero, or negative, while absolute values are always non-negative.

Given an $n \times n$ square matrix X, by removing its i-th row and j-th column, we obtain an $(n-1) \times (n-1)$ matrix, and the determinant of this matrix is called the (i,j)-th minor of X, denoted by M_{ij}. Then, Laplace's formula states that

$$|X| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \tag{22}$$

for *any* $1 \le i \le n$. Similarly,

$$|X| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij}$$

for any $1 \le j \le n$. For a scalar (i.e., a 1×1 matrix), the determinant is itself. Hence, this recursive formula can be used to define the determinant of any square matrix.

It is easy to prove that

$$|X| = |X^T|$$

for any square matrix X, and

$$|XY| = |X||Y|$$

when the product is well-defined. For a scalar c and an $n \times n$ matrix X,

$$|cX| = c^n |X|\,.$$

For a square matrix X, if there exists another matrix Y such that $XY = YX = I$, then we say Y is the inverse of X, denoted by X^{-1}. When the inverse of X exists, we say X is invertible. X^{-1} is of the same size as X. If X^{-1} exists, then its transpose $(X^{-1})^T$ is abbreviated as X^{-T}.

The following statement is useful for determining whether X is invertible or not:

$$X \text{ is invertible} \Longleftrightarrow |X| \neq 0. \tag{23}$$

In other words, a square matrix is invertible if and only if its determinant is non-zero.

Assuming both X and Y are invertible, XY exists, and c is a non-zero scalar, we have the following properties.

- X^{-1} is also invertible and $\left(X^{-1}\right)^{-1} = X$;
- $(cX)^{-1} = \frac{1}{c} X^{-1}$;
- $(XY)^{-1} = Y^{-1}X^{-1}$; and,
- $X^{-T} = \left(X^{-1}\right)^T = \left(X^T\right)^{-1}$.

1.7 Eigenvalue, eigenvector, rank, and trace of a square matrix

For a square matrix A, if there exist a *non-zero* vector $\boldsymbol{x}$ and a scalar λ such that

$$A\boldsymbol{x} = \lambda \boldsymbol{x}\,,$$

we say λ is an eigenvalue of A and $\boldsymbol{x}$ is an eigenvector of A (which is associated with this eigenvalue λ). An $n \times n$ real square matrix has n eigenvalues, although some of them may be equal to each other. The eigenvalue and eigenvectors of a real square matrix, however, may contain complex numbers.

Eigenvalues have connections with the diagonal entries and the determinant of A. Denote the n eigenvalues by $\lambda_1, \lambda_2, \ldots, \lambda_n$; the following equations hold (even if the eigenvalues are complex numbers):

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} a_{ii}\,, \tag{24}$$

$$\prod_{i=1}^{n} \lambda_i = |A|\,. \tag{25}$$

The latter equation shows that a square matrix is invertible if and only if all of its eigenvalues are non-zero. The summation of all eigenvalues ($\sum_{i=1}^{n} \lambda_i$) has a special name: the trace. The trace of a square matrix X is denoted by $\operatorname{tr}(X)$. Now we know

$$\operatorname{tr}(X) = \sum_{i=1}^{n} x_{ii}\,. \tag{26}$$

If we assume all matrix multiplications are well-defined, we have

$$\operatorname{tr}(XY) = \operatorname{tr}(YX)\,. \tag{27}$$

Applying this rule, we can easily derive

$$\operatorname{tr}(XYZ) = \operatorname{tr}(ZXY) = \operatorname{tr}(YZX)$$

and many other similar results.

The rank of a square matrix X equals its number of non-zero eigenvalues, and is denoted by $\operatorname{rank}(X)$.

If X is also symmetric, then the properties of its eigenvalues and eigenvectors are a lot nicer. Given any $n \times n$ *real symmetric matrix* X, the following statements are true.

- All the eigenvalues of X are real numbers, hence can be sorted. We will denote the eigenvalues of an $n \times n$ real symmetric matrix as $\lambda_1, \lambda_2, \ldots, \lambda_n$, and assume $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$— i.e., they are sorted in descending order.

- All the eigenvectors of X only contain real values. We will denote the eigenvectors as $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n$, and $\boldsymbol{\xi}_i$ is associated with λ_i—i.e., the eigenvectors are also sorted according to their associated eigenvalues. The eigenvectors are normalized—i.e., $\|\boldsymbol{\xi}_i\| = 1$ for any $1 \le i \le n$.

- The eigenvectors satisfy (for $1 \leq i, j \leq n$)

$$\boldsymbol{\xi}_i^T \boldsymbol{\xi}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} . \tag{28}$$

  That is, the n eigenvectors form an orthogonal basis set of $\mathbb{R}^n$. Let E be an $n \times n$ matrix whose i-th column is $\boldsymbol{\xi}_i$, that is,

$$E = [\boldsymbol{\xi}_1 | \boldsymbol{\xi}_2 | \dots | \boldsymbol{\xi}_n] .$$

  Then, Equation 28 is equivalent to

$$EE^T = E^T E = I . \tag{29}$$

- $\operatorname{rank}(E) = n$, because E is an orthogonal matrix. It is also easy to see that $|E| = \pm 1$ and $E^{-1} = E^T$.

- If we define a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, then the eigendecomposition of X is

$$X = E \Lambda E^T . \tag{30}$$

- The eigendecomposition can also be written in an equivalent form as

$$X = \sum_{i=1}^{n} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T , \tag{31}$$

  which is called the spectral decomposition. The spectral decomposition says that the matrix X equals a weighted sum of n matrices, each being the outer product between one eigenvector and itself, and weighted by the corresponding eigenvalue.

We will also encounter the generalized eigenvalue problem. Let A and B be two square matrices (and we assume they are real symmetric in this book). Then, a vector $\boldsymbol{x}$ and a scalar λ that satisfy

$$A\boldsymbol{x} = \lambda B \boldsymbol{x}$$

are called the generalized eigenvector and generalized eigenvalue of $\underline{A \text{ and } B}$, respectively. The generalized eigenvectors, however, are usually not normalized, and are not orthogonal.

1.8 Singular value decomposition

Eigendecomposition is related to the Singular Value Decomposition (SVD). We will briefly introduce a few facts for real matrices.

Let X be an $m \times n$ matrix; the SVD of X is

$$X = U \Sigma V^T , \tag{32}$$

where U is an $m \times m$ matrix, Σ is an $m \times n$ matrix whose non-diagonal elements are all 0, and V is an $n \times n$ matrix.

If there are a scalar σ and two vectors $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^n$ (both are unit vectors) that satisfy the following two equalities simultaneously:

$$X\boldsymbol{v} = \sigma\boldsymbol{u} \quad \text{and} \quad X^T\boldsymbol{u} = \sigma\boldsymbol{v}\,, \tag{33}$$

we say σ is a singular value of X, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are its associated left- and right-singular vectors, respectively.

If $(\sigma, \boldsymbol{u}, \boldsymbol{v})$ satisfy the above equation, so does $(-\sigma, -\boldsymbol{u}, \boldsymbol{v})$. In order to remove this ambiguity, the singular value is always non-negative (i.e., $\sigma \geq 0$).

The SVD finds all singular values and singular vectors. The columns of U are called the left-singular vectors of X, and the columns of V are the right-singular vectors. The matrices U and V are orthogonal. The diagonal entries in Σ are the corresponding singular values.

Because $XX^T = (U\Sigma V^T)(V\Sigma^T U^T) = U\Sigma\Sigma^T U^T$ and $\Sigma\Sigma^T$ is diagonal, we get that the left-singular vectors of X are the eigenvectors of XX^T; similarly, the right-singular vectors of X are the eigenvectors of X^TX; and the non-zero singular values of X (diagonal non-zero entries in Σ) are the square roots of the non-zero eigenvalues of XX^T and X^TX. A by-product is that the non-zero eigenvalues of XX^T and X^TX are exactly the same.

This connection is helpful. When $m \gg n$ (e.g., $n = 10$ but $m = 100\,000$), the eigendecomposition of XX^T needs to perform eigendecomposition for a $100\,000 \times 100\,000$ matrix, which is infeasible or at least very inefficient. However, we can compute the SVD of X. The squared positive singular values and left-singular vectors are the positive eigenvalues and their associated eigenvectors of XX^T. The same trick also works well when $n \gg m$, in which the right-singular vectors are useful in finding the eigendecomposition of X^TX.

1.9 Positive (semi-)definite real symmetric matrices

We only consider real symmetric matrices and real vectors in this section, although the definition of positive definite and positive semi-definite matrices are wider than that.

An $n \times n$ matrix A is positive definite if for any non-zero real vector $\boldsymbol{x}$ (i.e., $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{x} \neq \boldsymbol{0}$),

$$\boldsymbol{x}^T A\boldsymbol{x} > 0\,. \tag{34}$$

We say A is positive semi-definite if $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ holds for any $\boldsymbol{x}$. The fact that matrix A is positive definite can be abbreviated as "A is PD" or in mathematical notation as $A \succ 0$. Similarly, A is positive semi-definite is equivalent to $A \succcurlyeq 0$ or "A is PSD."

The term

$$\boldsymbol{x}^T A\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij} \tag{35}$$

is a real quadratic form, which will be frequently used in this book.

There is a simple connection between eigenvalues and PD (PSD) matrices. A real symmetric matrix is PD/PSD if and only if all its eigenvalues are positive/non-negative.

One type of PSD matrix we will use frequently is of the form AA^T or A^TA, in which A is any real matrix. The proof is pretty simple: because

$$\boldsymbol{x}^T AA^T \boldsymbol{x} = \left(A^T\boldsymbol{x}\right)^T \left(A^T\boldsymbol{x}\right) = \|A^T\boldsymbol{x}\|^2 \geq 0\,,$$

AA^T is PSD; and similarly A^TA is also PSD.

Now, for a PSD real symmetric matrix, we sort its eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, in which the final ≥ 0 relationship is always true for PSD matrices, but not for all real symmetric matrices.

2 Probability

A random variable is usually denoted by an upper case letter, such as X. A random variable is a variable that can take value from a finite or infinite set. To keep things simple, we refrain from using the measure-theoretic definition of random variables and probabilities. We will use the terms *random variable* and *distribution* interchangeably.

2.1 Basics

If a random variable X can take a value from a finite or countably infinite set, it is called a discrete random variable. Suppose the outcome of a particular trial is either success or failure, and the chance of success is p ($0 \leq p \leq 1$). When multiple trials are tested, the chance of success in any one trial is not affected by any other trial (i.e., the trials are independent). Then, we denote the number of trials that are required till we see the first successful outcome as X. X is a random variable, which can take its value from the countably infinite set $\{1, 2, 3, \ldots\}$, hence is a discrete random variable. We say X follows a geometric distribution with parameter p.[3]

A random variable is different from a usual variable: it may take different values with different likelihoods (or probabilities). Hence, a random variable is a function rather a variable whose value can be fixed. Let the set $E = \{x_1, x_2, x_3, \ldots\}$ denote all values a discrete random variable X can possibly take. We call each x_i an event. The number of events should be either finite or countably infinite, and the events are mutually exclusive; that is, if an event x_i happens, then any other event x_j ($j \neq i$) cannot happen in the same trial. Hence, the probability that either one of two events x_i or x_j happens equals the sum of the probability of the two events:

$$\Pr(X = x_1 || X = x_2) = \Pr(X = x_1) + \Pr(X = x_2)\,,$$

[3] Another definition of the geometric distribution defines it as the number of failures before the first success, and possible values are $\{0, 1, 2, \ldots\}$.

in which $\Pr(\cdot)$ means the probability and $||$ is the logical or. The summation rule can be extended to a countable number of elements.

A discrete random variable is determined by a probability mass function (p.m.f.) $p(X)$. A p.m.f. is specified by the probability of each event: $\Pr(X = x_i) = c_i$ $(c_i \in \mathbb{R})$, and it is a valid p.m.f. if and only if

$$c_i \geq 0 \; (\forall \, x_i \in E) \quad \text{and} \quad \sum_{x_i \in E} c_i = 1 \, . \tag{36}$$

For the geometric distribution, we have $\Pr(X = 1) = p$, $\Pr(X = 2) = (1-p)p$, and in general, $c_i = (1-p)^{i-1}p$. Since $\sum_{i=1}^{\infty} c_i = 1$, this is a valid p.m.f. And, $\Pr(X \leq 2) = Pr(X = 1) + Pr(X = 2) = 2p - p^2$. The function

$$F(x) = \Pr(X \leq x) \tag{37}$$

is called the cumulative distribution function (c.d.f. or CDF).

If the set of possible values E is infinite and uncountable (for example and most likely, $\mathbb{R}$ or a subset of it), and in addition $\Pr(X = x) = 0$ for each possible $x \in E$, then we say X is a continuous random variable or continuous distribution.[4]

As in the discrete case, the c.d.f. of X is still

$$F(x) = \Pr(X \leq x) = \Pr(X < x) \, ,$$

where the additional equality follows from $\Pr(X = x) = 0$. The corresponding function of the discrete p.m.f. is called the probability density function (p.d.f.) $p(x)$, which should satisfy

$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1 \tag{38}$$

to make $p(x)$ a valid p.d.f. In this book, we assume a continuous c.d.f. is differentiable; then

$$p(x) = F'(x) \, , \tag{39}$$

where F' means the derivative of F.

The c.d.f. measures the accumulation of probability in both discrete and continuous domains. The p.d.f. $p(x)$ (which is the derivative of c.d.f.) measures the rate of accumulation of probability—in other words, how dense is X at x. Hence, the higher $p(x)$ is, the larger the probability $\Pr(x - \varepsilon \leq X \leq x + \varepsilon)$ (but not a larger $\Pr(x)$, which is always 0).

A few statements about c.d.f. and p.d.f.:

- $F(x)$ is non-decreasing, and

$$F(-\infty) \triangleq \lim_{x \to -\infty} F(x) = 0 \, ,$$

[4] In fact, if $\Pr(X = x) = 0$ for all $x \in E$, then E cannot be finite or countable.

$$F(\infty) \triangleq \lim_{x\to\infty} F(x) = 1\,,$$

in which $\triangleq$ means "defined as". This property is true for both discrete and continuous distributions.

- $\Pr(a \le X \le b) = \int_a^b p(x)\,\mathrm{d}x = F(b) - F(a)$.
- Although the p.m.f. is always between 0 and 1, the p.d.f. can be any non-negative value.
- If a continuous X only takes values in the range $E = [a, b]$, we can still say that $E = \mathbb{R}$, and let the p.d.f. $p(x) = 0$ for $x < a$ or $x > b$.

When there is more than one random variable, we will use a subscript to distinguish them—e.g., $p_Y(y)$ or $p_X(x)$. If Y is a continuous random variable and g is a fixed function (i.e., no randomness in the computation of g) and is monotonic, then $X = g(Y)$ is also a random variable, and its p.d.f. can be computed as

$$p_Y(y) = p_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_X(g(y))\,|g'(y)|\,, \tag{40}$$

in which $|\cdot|$ is the absolute value function.

2.2 Joint and conditional distributions, and Bayes' theorem

In many situations we need to consider two or more random variables simultaneously. For example, let A be the age and I be the annual income in year 2016 (in RMB, using 10 000 as a step) for a person in China. Then, the joint CDF $\Pr(A \le a, I \le i)$ is the percentage of people in China whose age is not larger than a years old and whose income is not higher than i RMB in the year 2016. If we denote the random vector $X = (A, I)^T$ and $\boldsymbol{x} = (30, 80\,000)^T$, then $F(\boldsymbol{x}) = \Pr(X \le \boldsymbol{x}) = \Pr(A \le 30, I \le 80\,000)$ defines the c.d.f. of a joint distribution. This definition also applies to any number of random variables. The joint distribution can be discrete (if all random variables in it are discrete), continuous (if all random variables in it are continuous), or hybrid (if both discrete and continuous random variables exist). We will not deal with hybrid distributions in this book.

For the discrete case, a multidimensional p.m.f. $p(\boldsymbol{x})$ requires $p(\boldsymbol{x}) \ge 0$ for any $\boldsymbol{x}$ and $\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$. For the continuous case, we require the p.d.f. $p(\boldsymbol{x})$ to satisfy $p(\boldsymbol{x}) \ge 0$ for any $\boldsymbol{x}$ and $\int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$.

It is obvious that for a discrete p.m.f.,

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y})$$

when $\boldsymbol{x}$ and $\boldsymbol{y}$ are two random vectors (and one or both can be random variables—i.e., 1-dimensional random vectors). In the continuous case,

$$p(\boldsymbol{x}) = \int_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,.$$

The distributions obtained by summing or integrating one or more random variables out are called marginal distributions. The summation is taken over all possible values of $\boldsymbol{y}$ (the variable that is summed or integrated out).

Note that in general

$$p(\boldsymbol{x},\boldsymbol{y}) \neq p(\boldsymbol{x})p(\boldsymbol{y})\,.$$

For example, let us guess $\Pr(A=3)=0.04$ and $\Pr(I=80\,000)=0.1$—i.e., in China the percentage of people aged 3 is 4%, and people with 80 000 yearly income is 10%; then, $\Pr(A=3)\Pr(I=80\,000)=0.004$. However, we would expect $\Pr(A=3, I=80\,000)$ to be almost 0: how many 3-year-old babies have 80 000 RMB yearly income?

In a random vector, if we know the value of one random variable for a particular example (or sample, or instance, or instantiation), it will affect our estimate of other random variable(s) in that sample. In the age–income hypothetic example, if we know $I=80\,000$, then we know $A=3$ is almost impossible for the same individual. Our estimate of the age will change to a new one when we know the income, and this new distribution is called the conditional distribution. We use $\boldsymbol{x}|Y=\boldsymbol{y}$ to denote the random vector (distribution) of $\boldsymbol{x}$ conditioned on $Y=\boldsymbol{y}$, and use $p(\boldsymbol{x}|Y=\boldsymbol{y})$ to denote the conditional p.m.f. or p.d.f. For conditional distributions, we have

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{y})}\,, \tag{41}$$

$$p(\boldsymbol{y}) = \int_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,. \tag{42}$$

In the discrete case, $\int$ in Equation 42 is changed to $\sum$, and is called the law of total probability. Putting these two together, we get Bayes' theorem:[5]

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}\,. \tag{43}$$

Let $\boldsymbol{y}$ be some random vectors we can observe (or measure), and $\boldsymbol{x}$ be the random variables that we cannot directly observe but want to estimate or predict. Then, knowing the values of $\boldsymbol{y}$ are "evidences" that will make us update our estimate ("belief") of $\boldsymbol{x}$. Bayes' theorem provides a mathematically precise way to perform such updates, and we will use it frequently in this book.

2.3 Expectation and variance/covariance matrices

The expectation (or mean, or average, or expected value) of a random vector X is denoted by $\mathbb{E}[X]$ (or EX, or $E(X)$, or $\mathcal{E}(X)$, etc.), and is computed as

$$\mathbb{E}[X] = \int_{\boldsymbol{x}} p(\boldsymbol{x})\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}\,, \tag{44}$$

[5] It is also called Bayes' rule or Bayes' law, named after Thomas Bayes, a famous British statistician and philosopher.

i.e., a weighted sum of $\boldsymbol{x}$, and the weights are the p.d.f. or p.m.f. (changing $\int$ to $\sum$ in the discrete case). Note that the expectation is a normal scalar or vector, which is not affected by randomness anymore (or at least not the randomness related to X). Two obvious properties of expectations are

- $\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]$; and,
- $\mathbb{E}[cX]=c\mathbb{E}[X]$ for a scalar c.

The expectation concept can be generalized. Let $g(\cdot)$ be a function; then $g(X)$ is a random vector, and its expectation is

$$\mathbb{E}[g(X)]=\int_{\boldsymbol{x}} p(\boldsymbol{x})g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,. \tag{45}$$

Similarly, $g(X)|Y$ is also a random vector, and its expectation (the conditional expectation) is

$$\mathbb{E}[g(X)|Y=\boldsymbol{y}]=\int_{\boldsymbol{x}} p(\boldsymbol{x}|Y=\boldsymbol{y})g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,. \tag{46}$$

And we can also write

$$h(\boldsymbol{y})=\mathbb{E}[g(X)|Y=\boldsymbol{y}]\,. \tag{47}$$

Note that the expectation $\mathbb{E}[g(X)|Y=\boldsymbol{y}]$ is not dependent on X because it is integrated (or summed) out. Hence, $h(\boldsymbol{y})$ is a normal function of $\boldsymbol{y}$, which is not affected by the randomness caused by X anymore.

Now, in Equation 45 we can specify

$$g(x)=(x-\mathbb{E}[X])^2\,,$$

in which $\mathbb{E}[X]$ is a completely determined scalar (if the p.m.f. or p.d.f. of X is known) and $g(x)$ is thus not affected by randomness. The expectation of this particular choice is called the variance (if X is a random variable) or covariance matrix (if X is a random vector) of X:

$$\mathrm{Var}(X)=\mathbb{E}[(X-\mathbb{E}[X])^2] \quad \text{or} \quad \mathrm{Cov}(X)=\mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T] \tag{48}$$

When X is a random variable, this expectation is called the variance and is denoted by $\mathrm{Var}(X)$. Variance is a scalar (which is always non-negative), and its square root is called the standard deviation of X, denoted by σ_X.

When X is a random vector, this expectation is called the covariance matrix and is denoted by $\mathrm{Cov}(X)$. The covariance matrix for a d-dimensional random vector is a $d\times d$ real symmetric matrix, and is always positive semi-definite.

For a random variable X, it is easy to prove the following useful formula:

$$\mathrm{Var}(X)=\mathbb{E}[X^2]-(\mathbb{E}[X])^2\,. \tag{49}$$

Hence, the variance is the difference between two terms: the expectation of the squared random variable and the square of the mean. Because variance is non-negative, we always have

$$\mathbb{E}[X^2]\ge(\mathbb{E}[X])^2\,.$$

A similar formula holds for random vectors:

$$\mathrm{Cov}(X) = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T \,. \tag{50}$$

In a complex expectation involving multiple random variables or random vectors, we may specify for which random variable (or vector) we want to compute the expectation by adding a subscript. For example, $\mathbb{E}_X[g(X,Y)]$ computes the expectation of $g(X,Y)$ with respect to X.

One final note about expectation is that an expectation may not exist—e.g., if the integration or summation is undefined. The standard Cauchy distribution provides an example.[6] The p.d.f. of the standard Cauchy distribution is defined as

$$p(x) = \frac{1}{\pi(1+x^2)} \,. \tag{51}$$

Since $\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} \,\mathrm{d}x = \frac{1}{\pi}\left(\arctan(\infty) - \arctan(-\infty)\right) = 1$ and $p(x) \ge 0$, this is a valid p.d.f. The expectation, however, does not exist because it is the sum of two infinite values, which is not well-defined in mathematical analysis.

2.4 Inequalities

If we are asked to estimate a probability $\Pr(a \le X \le b)$ but know nothing about X, the best we can say is: it is between 0 and 1 (including both ends). That is, if there is no information in, there is no information out—the estimation is valid for any distribution and is not useful at all.

If we know more about X, we can say more about probabilities involving X. Markov's inequality states that if X is a non-negative random variable (or, $\Pr(X < 0) = 0$) and $a > 0$ is a scalar, then

$$\Pr(X \ge a) \le \frac{\mathbb{E}[X]}{a} \,, \tag{52}$$

assuming the mean is finite.[7]

Chebyshev's inequality depends on both the mean and the variance. For a random variable X, if its mean is finite and its variance is non-zero, then for any scalar $k > 0$,

$$\Pr\left(|X - \mathbb{E}[X]| \ge k\sigma\right) \le \frac{1}{k^2} \,, \tag{53}$$

in which $\sigma = \sqrt{\mathrm{Var}(X)}$ is the standard deviation of X.[8]

There is also a one-tailed version of Chebyshev's inequality, which states that for $k > 0$,

$$\Pr(X - \mathbb{E}[X] \ge k\sigma) \le \frac{1}{1+k^2} \,. \tag{54}$$

[6] The Cauchy distribution is named, once again, after Augustin-Louis Cauchy.

[7] This inequality is named after Andrey (Andrei) Andreyevich Markov, a famous Russian mathematician.

[8] This inequality is named after Pafnuty Lvovich Chebyshev, another Russian mathematician.

2.5 Independence and correlation

Two random variables X and Y are independent if and only if the joint c.d.f. $F_{X,Y}$ and the marginal c.d.f. F_X and F_Y satisfy

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \tag{55}$$

for *any* x and y; or equivalently, if and only if the p.d.f. satisfies

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\,. \tag{56}$$

When X and Y are independent, knowing the distribution of X does not give us any information about Y, and vice versa; in addition, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. When X and Y are not independent, we say they are dependent.

Another concept related to independence (or dependence) is correlatedness (or uncorrelatedness). Two random variables are said to be uncorrelated if their covariance is zero and correlated if their covariance is nonzero. The covariance between two random variables X and Y is defined as

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\,, \tag{57}$$

which measures the level of linear relationship between them.

The range of $\operatorname{Cov}(X,Y)$ is not bounded. A proper normalization could convert it to a closed interval. Pearson's correlation coefficient is denoted by $\rho_{X,Y}$ or $\operatorname{corr}(X,Y)$,[9] and is defined as

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma_X\sigma_Y}\,. \tag{58}$$

The range of Pearson's correlation coefficient is $[-1,+1]$. When the correlation coefficient is $+1$ or -1, X and Y are related by a perfect linear relationship $X = cY + b$; when the correlation coefficient is 0, they are uncorrelated.

When X and Y are random vectors (m- and n-dimensional, respectively), $\operatorname{Cov}(X,Y)$ is an $m \times n$ covariance matrix, and defined as

$$\begin{aligned}
\operatorname{Cov}(X,Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right] && (59)\\
&= \mathbb{E}\left[XY^T\right] - \mathbb{E}[X]\mathbb{E}[Y]^T\,. && (60)
\end{aligned}$$

Note that when $X = Y$, we get the covariance matrix of X (cf. Equation 50).

Independence is a much stronger condition than uncorrelatedness:

$$\begin{aligned}
&X \text{ and } Y \text{ are independent} \Longrightarrow X \text{ and } Y \text{ are uncorrelated.} && (61)\\
&X \text{ and } Y \text{ are uncorrelated} \not\Longrightarrow X \text{ and } Y \text{ are independent.} && (62)
\end{aligned}$$

[9] It is named after Karl Pearson, a famous British mathematician and biostatistician.

2.6 The normal distribution

Among all distributions, the normal distribution is probably the most widely used. A random variable X follows a normal distribution if its p.d.f. is in the form of

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{63}$$

for some $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. We can denote it as $X \sim N(\mu, \sigma^2)$ or $p(x) = N(x; \mu, \sigma^2)$. A normal distribution is also called a Gaussian distribution.[10] Note that the parameters that determine a normal distribution are (μ, σ^2), not (μ, σ).

A d-dimensional random vector is jointly normal (or has a multivariate normal distribution) if its p.d.f. is in the form of

$$p(\boldsymbol{x}) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{64}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive semi-definite symmetric matrix Σ, and $|\cdot|$ is the determinant of a matrix. We can write this distribution as $X \sim N(\boldsymbol{\mu}, \Sigma)$ or $p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$.

Examples of the normal p.d.f. are shown in Figure 2. Figure 2a is a normal distribution with $\mu = 0$ and $\sigma^2 = 1$, and Figure 2b is a 2-dimensional normal distribution with $\boldsymbol{\mu} = \boldsymbol{0}$ and $\Sigma = I_2$.

The expectation of single- and multi-variate normal distributions are μ and $\boldsymbol{\mu}$, respectively. Their variance and covariance matrices are σ^2 and Σ, respectively. Hence, μ and σ^2 (not σ) are the counterparts of $\boldsymbol{\mu}$ and Σ, respectively.

We might remember $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ very well, but are less familiar with the multivariate version $p(\boldsymbol{x})$. However, the 1D distribution can help us remember the more complex multivariate p.d.f. If we rewrite the univariate normal density into an equivalent form

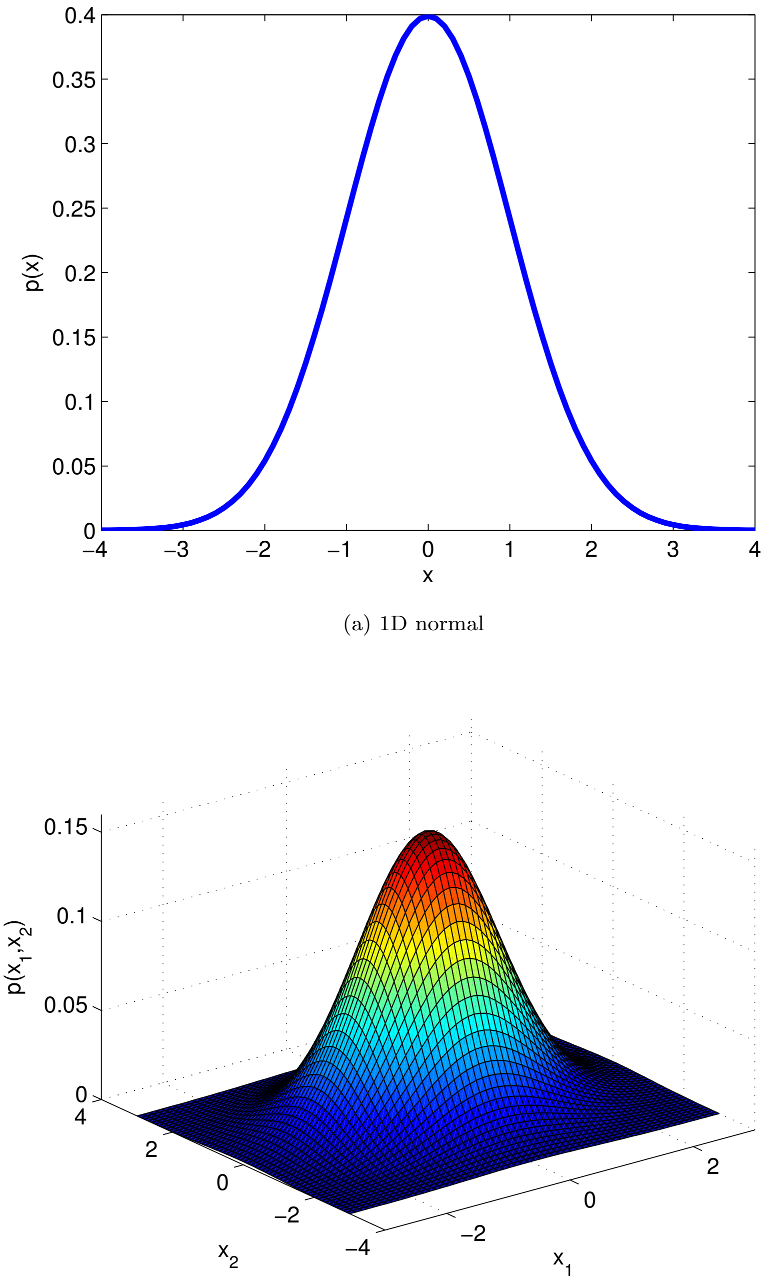
$$p(x) = (2\pi)^{-1/2} (\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T (\sigma^2)^{-1} (x-\mu)\right) \tag{65}$$

and change the dimensionality from 1 to d, the variance from σ^2 to the covariance matrix Σ or its determinant $|\Sigma|$, x to $\boldsymbol{x}$, and the mean from μ to $\boldsymbol{\mu}$, we get exactly the multivariate p.d.f.

$$p(\boldsymbol{x}) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{66}$$

The Gaussian distribution has many nice properties, some of which can be found in Chapter 13, the chapter devoted to the properties of normal distributions. One particularly useful property is: if X and Y are jointly Gaussian and X and Y are uncorrelated, then they are independent.

[10] It is named after Johann Carl Friedrich Gauss, a very influential German mathematician.



(a) 1D normal

(b) 2D normal

Figure 2: Probability density function of example normal distributions.

3 Optimization and matrix calculus

Optimization will be frequently encountered in this book. However, details of optimization principles and techniques are beyond the scope of this book. We will touch only a little bit on this huge topic in this chapter.

Informally speaking, given a cost (or objective) function $f(\boldsymbol{x}) : \mathcal{D} \mapsto \mathbb{R}$, the purpose of mathematical optimization is to find an $\boldsymbol{x}^\star$ in the domain $\mathcal{D}$, such that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{D}$. This type of optimization problem is called a minimization problem, and is usually denoted as

$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x}) \,. \tag{67}$$

A solution $\boldsymbol{x}^\star$ that makes $f(\boldsymbol{x})$ reach its minimum value is called a minimizer of f, and is denoted as

$$\boldsymbol{x}^\star = \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x}) \,. \tag{68}$$

Note that the minimizers of a minimization objective can be a (possibly infinite) set of values rather than a single point. For example, the minimizers for $\min_{x \in \mathbb{R}} \sin(x)$ is a set containing infinitely many numbers: $-\frac{\pi}{2} + 2n\pi$ for any integer n. We are, however, satisfied by any one of the minimizers in practice in many applications.

In contrast, an optimization problem can also be maximizing a function $f(\boldsymbol{x})$, denoted as $\max_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$. The maximizers are similarly denoted as $\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$. However, since maximizing $f(\boldsymbol{x})$ is equivalent to minimizing $-f(\boldsymbol{x})$, we often only talk about minimization problems.

3.1 Local minimum, necessary condition, and matrix calculus

An $\boldsymbol{x}^\star \in \mathcal{D}$ that satisfies $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{D}$ is called a global minimum. However, global minima are difficult to find in many (complex) optimization problems. In such cases, we are usually satisfied by a local minimum.

A local minimum, in layman's language, is some $\boldsymbol{x}$ that leads to the smallest objective value in its local neighborhood. In mathematics, $\boldsymbol{x}^\star$ is a local minimum if it belongs to the domain $\mathcal{D}$, and there exists some radius $r > 0$ such that for all $\boldsymbol{x} \in \mathcal{D}$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq r$, we always have $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$.

There is one commonly used criterion for determining whether a particular point $\boldsymbol{x}$ is a candidate for being a minimizer of $f(\boldsymbol{x})$. If f is differentiable, then

$$\frac{\partial f}{\partial \boldsymbol{x}} = \boldsymbol{0} \tag{69}$$

is a *necessary* condition for $\boldsymbol{x}$ to be a local minimum (or a local maximum). In other words, for $\boldsymbol{x}$ to be either a minimum or a maximum point, the gradient at that point should be an all-zero vector. Note that this is only a necessary condition, but may not be sufficient. And, we do not know an $\boldsymbol{x}$ satisfying this gradient test is a maximizer, a minimizer, or a saddle point (neither a maximum nor a minimum). Points with a all-zero gradient are also called stationary points or critical points.

The gradient $\frac{\partial f}{\partial \boldsymbol{x}}$ is defined in all undergraduate mathematics texts if $\boldsymbol{x} \in \mathbb{R}$, i.e., a scalar variable. In this case, the gradient is the derivative $\frac{\mathrm{d}f}{\mathrm{d}x}$. The gradient $\frac{\partial f}{\partial \boldsymbol{x}}$ for multivariate functions, however, is rarely included in these textbooks. These gradients are defined via matrix calculus as partial derivatives. For vectors $\boldsymbol{x}$, $\boldsymbol{y}$, scalars x, y, and matrix X, the matrix form is defined as

$$\left[\frac{\partial \boldsymbol{x}}{\partial y}\right]_i = \frac{\partial x_i}{\partial y}\,, \tag{70}$$

$$\left[\frac{\partial x}{\partial \boldsymbol{y}}\right]_i = \frac{\partial x}{\partial y_i}\,, \tag{71}$$

$$\left[\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right]_{ij} = \frac{\partial x_i}{\partial y_j} \quad \text{(which is a matrix)}\,, \tag{72}$$

$$\left[\frac{\partial y}{\partial X}\right]_{ij} = \frac{\partial y}{\partial x_{ij}}\,. \tag{73}$$

Using these definitions, it is easy to calculate some gradients (partial derivatives)—e.g.,

$$\frac{\partial \boldsymbol{x}^T\boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{y}\,, \tag{74}$$

$$\frac{\partial \boldsymbol{a}^T X \boldsymbol{b}}{\partial X} = \boldsymbol{a}\boldsymbol{b}^T\,. \tag{75}$$

However, for more complex gradients—for example, those involving matrix inverse, eigenvalues, and matrix determinant, the solutions are not obvious. We recommend *The Matrix Cookbook*, which lists many useful results—e.g.,

$$\frac{\partial \det(X)}{\partial X} = \det(X) X^{-T}\,. \tag{76}$$

3.2 Convex and concave optimization

Some functions have nicer properties than others in the optimization realm. For example, when $f(\boldsymbol{x})$ is a *convex* function whose domain is $\mathbb{R}^d$, any local minimum is also a global minimum. More generally, a convex minimization problem is to minimize a convex objective on a convex set. In convex minimization, any local minimum must also be a global minimum.

In this book, we only consider subsets of $\mathbb{R}^d$. If $S \subseteq \mathbb{R}^d$, then S is a convex set if for any $\boldsymbol{x} \in S$, $\boldsymbol{y} \in S$ and $0 \le \lambda \le 1$,

$$\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in S$$

always holds. In other words, if we pick any two points from a set S and the line segment connecting them falls entirely inside S, then S is convex. For example,

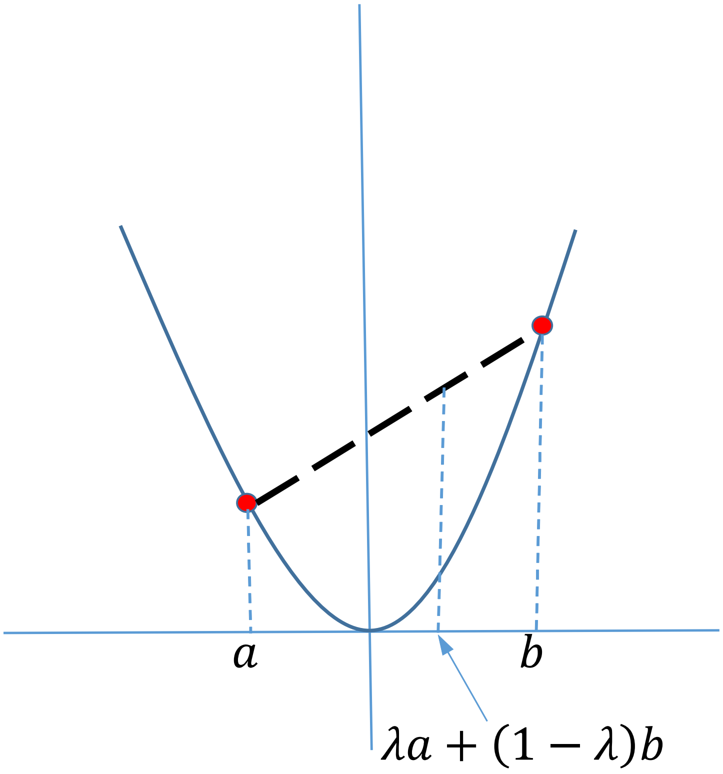


Figure 3: Illustration of a simple convex function.

in the 2-D space, all points inside a circle form a convex set, but the set of points outside that circle is not convex.

A function f (whose domain is S) is convex if for any $\boldsymbol{x}$ and $\boldsymbol{y}$ in S and any λ ($0 \le \lambda \le 1$), we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \le \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})\,.$$

$f(x) = x^2$ is a convex function. If we pick any two points $a < b$ on its curve, the line segment that connects them is above the $f(x) = x^2$ curve in the range (a, b), as illustrated in Figure 3.

If f is a convex function, we say that $-f$ is a concave function. Any local maximum of a concave function (on a convex domain) is also a global maximum.

Jensen's inequality shows that the constraints in the convex function definition can be extended to an arbitrary number of points. Let $f(\boldsymbol{x})$ be a convex function defined on a convex set S, and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are points in S. Then, for weights $w_1, w_2, \ldots, w_n$ satisfying $w_i \ge 0$ (for all $1 \le i \le n$) and $\sum_{i=1}^n w_i = 1$,

Jensen's inequality states that

$$f\left(\sum_{i=1}^{n} w_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)\,. \tag{77}$$

If $f(x)$ is a concave function defined on a convex set S, we have

$$f\left(\sum_{i=1}^{n} w_i \boldsymbol{x}_i\right) \geq \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)\,. \tag{78}$$

If we assume $w_i > 0$ for all i, the equality holds if and only if $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \cdots = \boldsymbol{x}_n$ or f is a linear function. If $w_i = 0$ for some i, we can remove these w_i and their corresponding $\boldsymbol{x}_i$ and apply the equality condition again.

A twice differentiable function is convex (concave) if its second derivative is non-negative (non-positive). For example, $\ln(x)$ is concave on $(0, \infty)$ because $\ln''(x) = -\frac{1}{x^2} < 0$. Similarly, $f(x) = x^4$ is convex because its second derivative is $12x^2 \geq 0$. The same convexity test applies to a convex subset of $\mathbb{R}$—e.g., an interval $[a, b]$.

For a scalar-valued function involving many variables, if it is continuous and twice differentiable, its second-order partial derivatives form a square matrix, called the Hessian matrix, or simply the Hessian. Such a function is convex if the Hessian is positive semi-definite. For example, $f(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$ is convex if A is positive semi-definite.

A function f is *strictly* convex if for any $\boldsymbol{x} \neq \boldsymbol{y}$ in a convex domain S and any λ $(0 < \lambda < 1)$, we have $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$.[11]

A twice differentiable function is strictly convex if its second derivative is positive (or its Hessian is positive definite in the multivariate case). For example, $f(x) = x^2$ is strictly convex, but any linear function is not. Hence, the equality condition for Jensen's inequality applied to a strictly convex function is $\boldsymbol{x}_i = \boldsymbol{c}$ if $w_i > 0$, in which $\boldsymbol{c}$ is a fixed vector.

For more treatments on convex functions and convex optimization, *Convex Optimization* is an excellent textbook and reference.

3.3 Constrained optimization and the Lagrange multipliers

Sometimes beyond the objective $f(\boldsymbol{x})$, we also require the variables $\boldsymbol{x}$ to satisfy some constraints. For example, we may require that $\boldsymbol{x}$ has unit length (which will appear quite frequently later in this book). For $\boldsymbol{x} = (x_1, x_2)^T$ and the domain $\mathcal{D} = \mathbb{R}^2$, a concrete example is

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) = \boldsymbol{v}^T \boldsymbol{x} && (79) \\ \text{s.t.} \quad & \boldsymbol{x}^T \boldsymbol{x} = 1\,, && (80) \end{aligned}$$

[11] Please pay attention to the three changes in this definition (compared to the convex function definition): $\boldsymbol{x} \neq \boldsymbol{y}$, $(0, 1)$ instead of $[0, 1]$, and $<$ instead of $\leq$.

in which $\boldsymbol{v} = \left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]$ is a constant vector and "s.t." means "subject to," which specifies a constraint on $\boldsymbol{x}$. There can be more than one constraint, and the constraint can also be an inequality.

Let us focus on equality constraints for now. A minimization problem with only equality constraints is

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) & (81)\\ \text{s.t.} \quad & g_1(\boldsymbol{x}) = 0, & (82)\\ & \ldots & \\ & g_m(\boldsymbol{x}) = 0\,. & (83) \end{aligned}$$

The method of Lagrange multipliers is a good tool to deal with this kind of problem.[12] This method defines a Lagrange function (or Lagrangian) as

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{x})\,, \tag{84}$$

in which $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^T$ are the m Lagrange multipliers, with the i-th Lagrange multiplier λ_i associated with the i-th constraint $g_i(\boldsymbol{x}) = 0$; and we use $\boldsymbol{g}(\boldsymbol{x})$ to denote $(g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}))^T$, the values of all m constraints. Then, L is an unconstrained optimization objective, and

$$\frac{\partial L}{\partial \boldsymbol{x}} = 0\,, \tag{85}$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = 0 \tag{86}$$

are necessary conditions for $(\boldsymbol{x}, \boldsymbol{\lambda})$ to be a stationary point of $L(\boldsymbol{x}, \boldsymbol{\lambda})$. Note that the domain of the Lagrange multipliers is $\mathbb{R}^m$—i.e., without any restriction. Hence, we can also change the minus sign $(-)$ to the plus sign $(+)$ in the Lagrangian.

The method of Lagrange multipliers states that if $\boldsymbol{x}_0$ is a stationary point of the original constrained optimization problem, there always exists a $\boldsymbol{\lambda}_0$ such that $(\boldsymbol{x}_0, \boldsymbol{\lambda}_0)$ is also a stationary point of the unconstrained objective $L(\boldsymbol{x}, \boldsymbol{\lambda})$. In other words, we can use Equations 85 and 86 to find all stationary points of the original problem.

If we move back to our example at the beginning of this section, its Lagrangian is

$$L(\boldsymbol{x}, \lambda) = \boldsymbol{v}^T \boldsymbol{x} - \lambda(\boldsymbol{x}^T \boldsymbol{x} - 1)\,. \tag{87}$$

Setting $\frac{\partial L}{\partial \boldsymbol{x}} = 0$ leads to $\boldsymbol{v} = 2\lambda \boldsymbol{x}$; and setting $\frac{\partial L}{\partial \boldsymbol{\lambda}} = 0$ gives us the original constraint $\boldsymbol{x}^T \boldsymbol{x} = 1$.

Because $\boldsymbol{v} = 2\lambda \boldsymbol{x}$, we have $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T \boldsymbol{v} = 4\lambda^2 \boldsymbol{x}^T \boldsymbol{x} = 4\lambda^2$. Hence, $|\lambda| = \frac{1}{2}\|\boldsymbol{v}\|$, and the stationary point is $\boldsymbol{x} = \frac{1}{2\lambda}\boldsymbol{v}$.

Since $\boldsymbol{v} = (1, 2)^T$ in our example, we have $\lambda^2 = \frac{1}{4}\|\boldsymbol{v}\|^2 = \frac{5}{4}$; or, $\lambda = \pm\frac{\sqrt{5}}{2}$. Thus, $f(\boldsymbol{x}) = \boldsymbol{v}^T \boldsymbol{x} = 2\lambda \boldsymbol{x}^T \boldsymbol{x} = 2\lambda$. Hence, $\min f(\boldsymbol{x}) = -\sqrt{5}$ and $\max f(\boldsymbol{x}) = \sqrt{5}$. The minimizer is $-\frac{1}{\sqrt{5}}(1, 2)^T$, and the maximizer is $\frac{1}{\sqrt{5}}(1, 2)^T$.

[12] This method is named after Joseph-Louis Lagrange, an Italian mathematician and astronomer.

These solutions are easily verified. Applying the Cauchy–Schwarz inequality, we know $|f(\boldsymbol{x})| = |\boldsymbol{v}^T\boldsymbol{x}| \leq \|\boldsymbol{v}\|\|\boldsymbol{x}\| = \sqrt{5}$. That is, $-\sqrt{5} \leq f(\boldsymbol{x}) \leq \sqrt{5}$, and the equality is obtained when $\boldsymbol{v} = c\boldsymbol{x}$ for some constant c. Because $\|\boldsymbol{v}\| = \sqrt{5}$ and $\|\boldsymbol{x}\| = 1$, we know $c = \sqrt{5}$ and get the maximum and minimum points as above.

The handling of inequality constraints is more complex than equality constraints, and involves duality, saddle points, and duality gaps. The method of Lagrange multipliers can be extended to handle these cases, but are beyond the scope of this book. Interested readers are referred to the *Convex Optimization* book for more details.

4 Complexity of algorithms

In the next chapter, we will discuss how modern algorithms and systems require a lot of computing and storage resources, in terms of the number of instructions to be executed by CPU and GPU, or the amount of data that need to be stored in main memory or hard disk. Of course, we prefer algorithms whose resource consumption (i.e., running time or storage complexity) is low.

In the theoretical analysis of an algorithm's complexity, we are often interested in how fast the complexity grows when the input size gets larger. The unit of such complexity, however, is usually variable. For example, when the complexity analysis is based on a specific algorithm's pseudocode, and we are interested in how many arithmetic operations are involved when the input size is 100, the running time complexity might evaluate to 50,000 arithmetic operations. If instead we are interested in the number of CPU instructions that are executed, the same algorithm may have a complexity of 200,000 in terms of CPU instructions.

The big-O notation ($\mathcal{O}$) is often used to analyze the theoretical complexity of algorithms, which measures how the running time or storage requirement grows when the size of the input increases. Note that the input size may be measured by more than one number—e.g., by both the number of training examples n and the length of the feature vector d.

When the input size is a single number n and the complexity is $f(n)$, we say this algorithm's complexity is $\mathcal{O}(g(n))$ if and only if there exist a positive constant M and an input size n_0, such that when $n \geq n_0$, we always have

$$f(n) \leq Mg(n)\,. \tag{88}$$

We assume both f and g are positive in the above equation. With a slight abuse of notation, we can write the complexity as

$$f(n) = \mathcal{O}(g(n))\,. \tag{89}$$

Informally speaking, Equation 89 states that $f(n)$ grows at most as fast as $g(n)$ after the problem size n is large enough.

An interesting observation can be made from Equation 88. If this equation holds, then we should have $f(n) \leq cMg(n)$ when $c > 1$, too. That is, when

$c > 1$, $f(n) = \mathcal{O}(g(n))$ implies $f(n) = \mathcal{O}(cg(n))$. In other words, a positive constant scalar will not change the complexity result in the big O notation. A direct consequence of this observation is that we do not need to be very careful in deciding the unit for our complexity result.

However, in a specific application, different constant scalars might have very different impacts. Both $f_1(n) = 2n^2$ and $f_2(n) = 20n^2$ are $\mathcal{O}(n^2)$ in the big O notation, but their running speed may differ by a factor of 10, and this speed variation makes a big difference in real world systems.

The big O notation can be generalized to scenarios when there are more variables involved in determining the input size. For example, the first pattern recognition system we introduce (cf. Chapter 3) has a complexity $f(n, d) = \mathcal{O}(nd)$ when there are n training examples and each training example is d-dimensional. This notation means that there exist numbers n_0 and d_0, and a positive constant M, such that when $n \geq n_0$ and $d \geq d_0$, we always have

$$f(n, d) \leq Mnd\,. \tag{90}$$

The generalization to more than two variables is trivial.

Exercises

1. Let $\boldsymbol{x} = (\sqrt{3}, 1)^T$ and $\boldsymbol{y} = (1, \sqrt{3})^T$ be two vectors, and $\boldsymbol{x}_\perp$ be the projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$.

   (a) What is the value of $\boldsymbol{x}_\perp$?

   (b) Prove that $\boldsymbol{y} \perp (\boldsymbol{x} - \boldsymbol{x}_\perp)$.

   (c) Draw a graph to illustrate the relationship between these vectors.

   (d) Prove that for any $\lambda \in \mathbb{R}$, $\|\boldsymbol{x} - \boldsymbol{x}_\perp\| \le \|\boldsymbol{x} - \lambda\boldsymbol{y}\|$. (Hint: the geometry between these vectors suggests that $\|\boldsymbol{x} - \boldsymbol{x}_\perp\|^2 + \|\boldsymbol{x}_\perp - \lambda\boldsymbol{y}\|^2 = \|\boldsymbol{x} - \lambda\boldsymbol{y}\|^2$.)

2. Let X be a 5×5 real symmetric matrix, whose eigenvalues are 1, 1, 3, 4, and x.

   (a) Define a necessary and sufficient condition on x such that X is a positive definite matrix.

   (b) If $\det(X) = 72$, what is the value of x?

3. Let $\boldsymbol{x}$ be a d-dimensional random vector, and $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$.

   (a) Let $p(\boldsymbol{x})$ be the probability density function for $\boldsymbol{x}$. What is the equation that defines $p(\boldsymbol{x})$?

   (b) Write down the equation that defines $\ln p(\boldsymbol{x})$.

   (c) If you have access to *The Matrix Cookbook*, which equation will you use to help you derive $\frac{\partial \ln p(\boldsymbol{x})}{\partial \boldsymbol{\mu}}$? What is your result?

   (d) Similarly, if we treat Σ^{-1} as variables (rather than Σ), which equation (or equations) will you use to help you derive $\frac{\partial \ln p(\boldsymbol{x})}{\partial \Sigma^{-1}}$, and what is the result?

4. (Schwarz inequality) Let X and Y be two random variables (discrete or continuous) and $\mathbb{E}[XY]$ exists. Prove that

$$(\mathbb{E}[XY])^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

5. Prove the following equality and inequalities.

   (a) Starting from the definition of covariance matrix for a random vector X, prove

$$\operatorname{Cov}(X) = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T.$$

   (b) Let X and Y be two random variables. Prove that for any constant $u \in \mathbb{R}$ and $v \in \mathbb{R}$,

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(X + u, Y + v).$$

   (c) Let X and Y be two random variables (discrete or continuous). Prove that the correlation coefficient $\rho_{X,Y}$ satisfies

$$-1 \le \rho_{X,Y} \le 1.$$

6. Answer the following questions related to the exponential distribution.

   (a) Calculate the expectation and variance of the exponential distribution with p.d.f. $p(x) = \begin{cases} \beta e^{-\beta x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$ (in which $\beta > 0$).

   (b) What is the c.d.f. of this distribution?

   (c) (Memoryless property) Let X denote the continuous exponential random variable. Prove that for any $a > 0$ and $b > 0$,

$$\Pr(X \geq a + b | X \geq a) = \Pr(X \geq b)\,.$$

   (d) If we assume X, the lifetime of a light bulb, follows the exponential distribution with $\beta = 10^{-3}$. What is its expected lifetime? If a particular light bulb has worked 2000 hours, what is the expectation of its remaining lifetime?

7. Suppose X is a random variable following the exponential distribution, whose probability density function is $p(x) = 3e^{-3x}$ for $x \geq 0$ and $p(x) = 0$ for $x < 0$.

   (a) What is the value of $\mathbb{E}[X]$ and $\operatorname{Var}(X)$? Just give the results, no derivation is needed.

   (b) Can we apply Markov's inequality to this distribution? If the answer is yes, what is the estimate for $\Pr(X \geq 1)$?

   (c) Can we apply Chebyshev's inequality? If the answer is yes, what is the estimate for $\Pr(X \geq 1)$?

   (d) The one-sided (or one-tailed) Chebyshev inequality states that: if $\mathbb{E}[X]$ and $\operatorname{Var}(X)$ both exist, for any positive number $a > 0$, we have $\Pr(X \geq \mathbb{E}[X] + a) \leq \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)+a^2}$ and $\Pr(X \leq \mathbb{E}[X] - a) \leq \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)+a^2}$. Apply this inequality to estimate $\Pr(X \geq 1)$.

   (e) What is the exact value for $\Pr(X \geq 1)$?

   (f) Compare the four values: estimate based on Markov's inequality, estimate based on Chebyshev's inequality, estimate based on one-sided Chebyshev inequality, and the true value; what conclusion do you get?

8. Let A be a $d \times d$ real symmetric matrix, whose eigenvalues are sorted and denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. The eigenvector associated with λ_i is $\boldsymbol{\xi}_i$. All the eigenvectors form an orthogonal matrix E, whose i-th column is $\boldsymbol{\xi}_i$. If we denote $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, we have $A = E\Lambda E^T$.

   (a) For any non-zero vector $\boldsymbol{x} \neq \boldsymbol{0}$, the term

$$\frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

   is called the *Rayleigh quotient*, denoted by $R(\boldsymbol{x}, A)$. Prove that for any $c \neq 0$,

$$R(\boldsymbol{x}, A) = R(c\boldsymbol{x}, A)\,.$$

(b) Show that

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \max_{\boldsymbol{x}^T\boldsymbol{x}=1} \boldsymbol{x}^T A\boldsymbol{x}\,.$$

(c) Show that any unit norm vector $\boldsymbol{x}$ (i.e., $\|\boldsymbol{x}\| = 1$) can be expressed as a linear combination of the eigenvectors, as $\boldsymbol{x} = E\boldsymbol{w}$, or equivalently,

$$\boldsymbol{x} = \sum_{i=1}^{d} w_i \boldsymbol{\xi}_i$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_d)^T$, with $\|\boldsymbol{w}\| = 1$.

(d) Prove that

$$\max_{\boldsymbol{x}^T\boldsymbol{x}=1} \boldsymbol{x}^T A\boldsymbol{x} = \lambda_1\,,$$

i.e., the maximum value of the Rayleigh quotient $R(\boldsymbol{x}, A)$ is λ_1, the largest eigenvalue of A. What is the optimal $\boldsymbol{x}$ that achieves this maximum? (Hint: Express $\boldsymbol{x}$ as a linear combination of $\boldsymbol{\xi}_i$.)

(e) Prove that

$$\min_{\boldsymbol{x}^T\boldsymbol{x}=1} \boldsymbol{x}^T A\boldsymbol{x} = \lambda_d\,,$$

i.e., the minimum value of the Rayleigh quotient $R(\boldsymbol{x}, A)$ is λ_d, the smallest eigenvalue of A. What is the optimal $\boldsymbol{x}$ that achieves this minimum? (Hint: Express $\boldsymbol{x}$ as a linear combination of $\boldsymbol{\xi}_i$.)

9. Answer the following questions on the Cauchy distribution.

(a) Show that the Cauchy distribution is a valid continuous distribution.

(b) Show that the expectation of the Cauchy distribution does not exist.

10. Answer the following questions related to convex and concave functions.

(a) Show that $f(x) = e^{ax}$ is a convex function for any $a \in \mathbb{R}$.

(b) Show that $g(x) = \ln(x)$ is a concave function on $\{x|x > 0\}$.

(c) Show that $h(x) = x\ln(x)$ is a convex function on $\{x|x \geq 0\}$ (we define $0 \ln 0 = 0$).

(d) Given a discrete distribution with its p.m.f. $(p_1, p_2, \ldots, p_n)$ $(p_i \geq 0)$, its *entropy* is defined as

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i\,,$$

in which we assume $0 \ln 0 = 0$. Use the method of Lagrange multipliers to find which values of p_i will maximize the entropy.

11. Let X and Y be two random variables.

(a) Prove that if X and Y are independent, then they are uncorrelated.

(b) Let X be uniformly distributed on $[-1, 1]$, and $Y = X^2$. Show that X and Y are uncorrelated but not independent.

(c) Let X and Y be two discrete random variables whose values can be either 1 or 2. The joint probability is $p_{ij} = \Pr(X = i, Y = j)$ $(i, j \in \{1, 2\})$. Prove that if X and Y are uncorrelated, then they are independent, too.