

Chapter 2

Multivariable Calculus

2.1 Review of Euclidean n -Space R^n

2.2 Geometric Approach: Vectors in 2-Space and 3-Space

A *vector* is a quantity that is determined by both its magnitude and its direction: thus, it is a directed line segment (Figure 2.1).

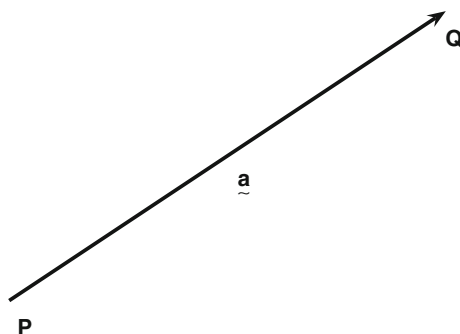


Fig. 2.1 A vector

The length of a vector \mathbf{a} is also called its *norm* (Euclidean norm), denoted by $\|\mathbf{a}\|$.

This chapter is based on the lectures of Professor D.V. Pai, IIT Bombay, and IIT Gandhinagar, Gujarat, India.

2.2.1 Components and the Norm of a Vector

For a vector \mathbf{a} with the initial point $P : (u_1, u_2, u_3)$ and the terminal point $Q : (v_1, v_2, v_3)$, its *components* are the three numbers:

$$a_1 = v_1 - u_1, \quad a_2 = v_2 - u_2, \quad a_3 = v_3 - u_3,$$

and we write

$$\mathbf{a} = (a_1, a_2, a_3).$$

The norm of \mathbf{a} is the number

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2}.$$

2.2.2 Position Vector

Given a Cartesian coordinate system, each point $P : (x_1, x_2, x_3)$ is determined by its *position vector* $\mathbf{r} = (x_1, x_2, x_3)$ with the initial point origin and the terminal point P . The origin is determined by the null vector or *zero-vector* $\mathbf{O} = (0, 0, 0)$ with length 0 and no direction.

2.2.3 Vectors as Ordered Triplets of Real Numbers

There is a 1-1 correspondence between vectors and ordered triplets of R . Given vectors $\mathbf{a} = (a_1, a_2, a_3)$, $\mathbf{b} = (b_1, b_2, b_3)$, we define (Figure 2.2)

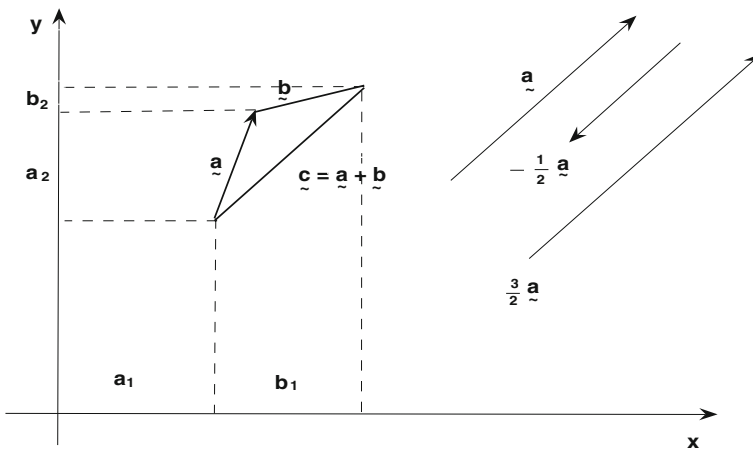


Fig. 2.2 Scalar multiplication and addition of vectors

$\mathbf{a} = \mathbf{b} \Leftrightarrow a_1 = b_1, a_2 = b_2, a_3 = b_3$ (*equality*)
 $\mathbf{a} + \mathbf{b} = (a_1 + b_1, a_2 + b_2, a_3 + b_3)$ (*vector addition*)
 $\alpha \mathbf{a} = (\alpha a_1, \alpha a_2, \alpha a_3)$ for all $\alpha \in R$; (*scalar multiplication*).

2.3 Analytic Approach

Given $n \in N$, let $R^n := \{\mathbf{a} = (a_1, a_2, \dots, a_n) : a_i \in R, i = 1, 2, \dots, n\}$ denote the *Cartesian n -space*. The elements of R^n are called *vectors* (or more precisely *n -vectors*).

Definition 2.1. Given two vectors $\mathbf{a} = (a_1, \dots, a_n)$, $\mathbf{b} = (b_1, \dots, b_n)$, we define:
(equality) $\mathbf{a} = \mathbf{b}$ iff $a_i = b_i, i = 1, 2, \dots, n$.
(addition) $\mathbf{a} + \mathbf{b} = (a_1 + b_1, \dots, a_n + b_n)$.
(scalar multiplication) $c\mathbf{a} = (ca_1, \dots, ca_n)$ for all $c \in R$.

2.3.1 Properties of Addition and Scalar Multiplication

For all $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in R^n and α, β in R , we have

- (i) $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$ (*commutativity*),
- (ii) $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$ (*associativity*),
- (iii) $\mathbf{a} + \mathbf{O} = \mathbf{O} + \mathbf{a} = \mathbf{a}$ (*zero element*),
- (iv) $\mathbf{a} + (-\mathbf{a}) = \mathbf{O}$ (*inverse element*),
- (v) $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$ (*distributivity*),
- (vi) $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$ (*distributivity*),
- (vii) $\alpha(\beta\mathbf{a}) = (\alpha\beta)\mathbf{a}$,
- (viii) $1\mathbf{a} = \mathbf{a}$.

The above properties of addition and scalar multiplication are used as axioms for defining a general *vector space*.

2.4 The Dot Product or the Inner Product

Definition 2.2. For \mathbf{a}, \mathbf{b} in R^n , their *dot* (or *inner*) product, written $\mathbf{a} \cdot \mathbf{b}$ (or (\mathbf{a}, \mathbf{b})) is the number

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i. \quad (2.4.1)$$

2.4.1 The Properties of Dot Product

For all $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in R^n and α in R , we have:

- (i) $\mathbf{a} \cdot \mathbf{a} \geq 0$,
- (ii) $\mathbf{a} \cdot \mathbf{a} = 0$ iff $\mathbf{a} = \mathbf{O}$ ((i)&(ii) \Rightarrow positivity),
- (iii) $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$ (commutativity),
- (iv) $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$
- (v) $\alpha(\mathbf{a} \cdot \mathbf{b}) = (\alpha\mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot (\alpha\mathbf{b})$ ((iv)&(v) \Rightarrow bilinearity).

Theorem 2.1. (The Cauchy–Schwarz inequality) For all \mathbf{a}, \mathbf{b} in R^n , we have

$$(\mathbf{a} \cdot \mathbf{b})^2 \leq (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}). \quad (2.4.2)$$

Equality holds in (2.4.2) if and only if \mathbf{a} is a scalar multiple of \mathbf{b} , or one of them is \mathbf{O} .

Proof. If either \mathbf{a} or \mathbf{b} equals \mathbf{O} , then (2.4.2) holds trivially. So let $\mathbf{a} \neq \mathbf{O}$ and $\mathbf{b} \neq \mathbf{O}$. Let $\mathbf{x} = \alpha\mathbf{a} - \beta\mathbf{b}$ where $\alpha = \mathbf{b} \cdot \mathbf{b}$ and $\beta = \mathbf{a} \cdot \mathbf{b}$, then

$$\begin{aligned} 0 \leq \mathbf{x} \cdot \mathbf{x} &= (\alpha\mathbf{a} - \beta\mathbf{b}) \cdot (\alpha\mathbf{a} - \beta\mathbf{b}) \\ &= \alpha^2(\mathbf{a} \cdot \mathbf{a}) - 2\alpha\beta(\mathbf{a} \cdot \mathbf{b}) + \beta^2(\mathbf{b} \cdot \mathbf{b}) \\ &= (\mathbf{b} \cdot \mathbf{b})^2(\mathbf{a} \cdot \mathbf{a}) - 2(\mathbf{b} \cdot \mathbf{b})(\mathbf{a} \cdot \mathbf{b})^2 + (\mathbf{a} \cdot \mathbf{b})^2(\mathbf{b} \cdot \mathbf{b}) \\ &= (\mathbf{b} \cdot \mathbf{b})^2(\mathbf{a} \cdot \mathbf{a}) - (\mathbf{a} \cdot \mathbf{b})^2(\mathbf{b} \cdot \mathbf{b}). \end{aligned} \quad (2.4.3)$$

Since $\mathbf{b} \cdot \mathbf{b} \neq 0$, we may divide by $(\mathbf{b} \cdot \mathbf{b})$ to obtain

$$(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2 \geq 0,$$

which is (2.4.2). Next, if the equality $(\mathbf{a} \cdot \mathbf{b})^2 = (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b})$ holds in (2.4.2), then by (2.4.3), $\mathbf{x} \cdot \mathbf{x} = 0$ and so $\mathbf{x} = \mathbf{O}$, that is, $(\mathbf{b} \cdot \mathbf{b})\mathbf{a} = (\mathbf{a} \cdot \mathbf{b})\mathbf{b}$. Thus, \mathbf{a} equals the scalar multiple $\frac{(\mathbf{a} \cdot \mathbf{b})}{(\mathbf{b} \cdot \mathbf{b})}\mathbf{b}$ of \mathbf{b} in case $\mathbf{b} \neq \mathbf{O}$. On the other hand, if either one of \mathbf{a}, \mathbf{b} equals \mathbf{O} or $\mathbf{a} = k\mathbf{b}$ for some $k \in R$, then it is easily seen that

$$(\mathbf{a} \cdot \mathbf{b})^2 = (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}).$$

Definition 2.3. For a vector $\mathbf{a} \in R^n$, the length or *norm* of \mathbf{a} is the number

$$\|\mathbf{a}\| = (\mathbf{a} \cdot \mathbf{a})^{\frac{1}{2}}. \quad (2.4.4)$$

This enables us to rewrite the Cauchy–Schwarz inequality (2.4.2) as

$$|(\mathbf{a} \cdot \mathbf{b})| \leq \|\mathbf{a}\| \|\mathbf{b}\|. \quad (2.4.5)$$

2.4.2 Properties of the Norm

For all \mathbf{a}, \mathbf{b} in R^n and $\alpha \in R$, we have

- (i) $\|\mathbf{a}\| \geq 0$,
- (ii) $\|\mathbf{a}\| = 0$ iff $\mathbf{a} = \mathbf{O}$ ((i)&(ii) \Rightarrow positivity),
- (iii) $\|\alpha\mathbf{a}\| = |\alpha| \|\mathbf{a}\|$ (homogeneity),
- (iv) $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$ (triangle inequality).

Remark 2.1. (i) The equality holds in the triangle inequality

$$\|\mathbf{a} + \mathbf{b}\| = \|\mathbf{a}\| + \|\mathbf{b}\| \quad (2.4.6)$$

iff \mathbf{a} is a positive scalar multiple of \mathbf{b} , or one of them is \mathbf{O} .

(ii) The *Pythagorean identity*

$$\|\mathbf{a} + \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$$

holds iff $\mathbf{a} \cdot \mathbf{b} = 0$.

Definition 2.4. Two vectors \mathbf{a}, \mathbf{b} in R^n are said to be *orthogonal* if $\mathbf{a} \cdot \mathbf{b} = 0$.

2.4.3 Geometric Interpretation of the Dot Product

The Cauchy–Schwarz inequality (2.4.2) shows that for any two non-null vectors \mathbf{a}, \mathbf{b} in R^n , we have

$$-1 \leq \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \leq 1.$$

Thus, there is exactly one real θ in the interval $0 \leq \theta \leq \pi$ such that

$$\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \cos \theta.$$

This defines the *angle* θ between the vectors \mathbf{a} and \mathbf{b} . With this definition of θ , we obtain

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta. \quad (2.4.7)$$

Recall that for the vectors in 3-space, (2.4.7) is, in fact, taken as the definition of the dot product for the geometric approach.

Given two vectors \mathbf{x}, \mathbf{y} in R^n , the *distance* between them is defined by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \left[\sum_{i=1}^n |x_i - y_i|^2 \right]^{\frac{1}{2}} \quad (\text{Euclidean distance}).$$

For developing the basic notions of multivariable calculus, just as in the univariate case, the notion of *neighborhood* of a point is crucial to us.

Definition 2.5. Let $\mathbf{x}^0 = (x_1^0, \dots, x_n^0) \in R^n$ and $r > 0$ be given. A convenient neighborhood of \mathbf{x}^0 is the set of all vectors $\mathbf{x} \in R^n$ such that

$$d(\mathbf{x}, \mathbf{x}^0) = \|\mathbf{x} - \mathbf{x}^0\| < r.$$

This is called the *open ball* with center \mathbf{x}^0 and radius r , denoted by $B(\mathbf{x}^0, r)$. Thus,

$$B(\mathbf{x}^0, r) = \{\mathbf{x} \in R^n : \|\mathbf{x} - \mathbf{x}^0\| < r\}.$$

As in the case of univariate calculus, we may prefer to begin with the notion of convergence of a sequence.

Definition 2.6. Given a sequence $\{\mathbf{a}^k\}_{k \in N}$ of vectors in R^n , we say that this sequence converges to a vector $\mathbf{a} \in R^n$, provided the sequence of real numbers $\{\|\mathbf{a}^k - \mathbf{a}\| : k \in N\}$ converges to 0. Put differently, $\mathbf{a}^k \rightarrow \mathbf{a}$ provided for every $\epsilon > 0$, there corresponds a number $K \in N$, such that

$$\|\mathbf{a}^k - \mathbf{a}\| < \epsilon \text{ for all } k \geq K.$$

This is the same as saying that for every $\epsilon > 0$, \mathbf{a}^k belongs to the open ball $B(\mathbf{a}, \epsilon)$ for k large enough.

Remark 2.2. If $\mathbf{a}^{(k)} = (a_1^{(k)}, \dots, a_n^{(k)})$, $k \in N$, and $\mathbf{a} = (a_1, \dots, a_n)$, $k \in N$, then clearly from the above definition, we have

$$\lim_{k \rightarrow \infty} \mathbf{a}^{(k)} = \mathbf{a} \iff \lim_{k \rightarrow \infty} a_i^{(k)} = a_i, \quad i = 1, 2, \dots, n.$$

2.5 Multivariable Functions

Definition 2.7. Let $D \subset R^n$. By a function \mathbf{F} on D to R^m , we mean a correspondence that assigns a unique vector

$$\mathbf{y} = \mathbf{F}(\mathbf{x})$$

in R^m to each element $\mathbf{x} = (x_1, \dots, x_n)$ in D . We write $\mathbf{F} : D \subset R^n \rightarrow R^m$ to signify that the set D is the *domain* of \mathbf{F} and R^m is the target space. The *range* of \mathbf{F} denoted $\mathcal{R}_{\mathbf{F}}$ is the set $\{\mathbf{F}(\mathbf{x}) : \mathbf{x} \in D\}$ of all images of elements of D . Here, \mathbf{x} is called an *input vector* and \mathbf{y} is called an *output vector*. If $m = 1$, the function \mathbf{F} is *real-valued* or *scalar-valued* which we simply denote by F . If $m > 1$, the function is called *vector-valued*. If $n > 1$, we call such a function a *function of several variables* or a *multivariate function*.

Remark 2.3. Note that for $m > 1$, each vector-valued function $\mathbf{F} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be written in terms of its components

$$\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \dots, F_m(\mathbf{x})), \quad \mathbf{x} \in D. \quad (2.5.1)$$

Here, if $\mathbf{y} = (y_1, \dots, y_m) = \mathbf{F}(\mathbf{x})$, then we define

$$F_i(\mathbf{x}) = F_i(x_1, \dots, x_n) = y_i, \quad \mathbf{x} \in D, \quad i = 1, \dots, m, \quad (2.5.2)$$

called the i^{th} component of \mathbf{F} . Conversely, given m scalar-valued functions $F_i : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$, we can get a vector-valued function $\mathbf{F} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ defined by (2.5.1).

2.5.1 Scalar and Vector Functions and Fields

If to each point P of a set $D \subset \mathbb{R}^3$ is assigned a scalar $f(P)$, then a *scalar field* is said to be defined in D and the function $f : D \rightarrow \mathbb{R}$ is called a *scalar function* (or a *scalar field* itself). Likewise, if to each point P in D is assigned a vector $\mathbf{F}(P) \in \mathbb{R}^3$ then a *vector field* is said to be defined in D and the vector-valued function $\mathbf{F} : D \rightarrow \mathbb{R}^3$ is called a *vector function* (or a *vector field* itself).

If we introduce Cartesian coordinates x, y, z , then instead of $f(P)$ we can write $f(x, y, z)$ and

$$\mathbf{F}(x, y, z) = (F_1(x, y, z), F_2(x, y, z), F_3(x, y, z))$$

where F_1, F_2, F_3 are the components of \mathbf{F} .

Remark 2.4. (i) A scalar field or a vector field arising from geometric or physical considerations must depend only on the points P where it is defined and not on the particular choice of Cartesian coordinates.

(ii) More generally, if D is a subset of \mathbb{R}^n , then a *scalar field* in D is a function $f : D \rightarrow \mathbb{R}$ and a *vector field* in D is a function $\mathbf{F} : D \rightarrow \mathbb{R}^n$. In the latter case,

$$\mathbf{F}(x_1, \dots, x_n) = (F_1(x_1, \dots, x_n), \dots, F_n(x_1, \dots, x_n))$$

where F_1, \dots, F_n are the components of \mathbf{F} . If $n = 2$, $f(\text{resp. } \mathbf{F})$ is called a *scalar* (resp. *vector*) *field in the plane*. If $n = 3$, $f(\text{resp. } \mathbf{F})$ is a *scalar* (resp. *vector*) *field in space*.

Example 2.1. (*Euclidean distance*) Let $D = \mathbb{R}^3$ and $f(P) = \|\vec{PP}_0\|$ the distance of point P from a fixed point P_0 in space. $f(P)$ defines a scalar field in space. If we introduce a Cartesian coordinate system in which $P_0 : (x_0, y_0, z_0)$, then

$$\begin{aligned} f(P) &= f(x, y, z) = \|(x - x_0, y - y_0, z - z_0)\| \\ &= \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}. \end{aligned}$$

Note that the value of $f(P)$ does not depend on the particular choice of Cartesian coordinate system.

Example 2.2. (*Thermal field*) In a region D of R^3 , one may be required to specify steady state temperature distribution function $u : D \rightarrow R$, $(x, y, z) \in D \rightarrow u(x, y, z)$. In this case, D becomes a scalar field.

Example 2.3. (*Gravitational force field*) Place the origin of a coordinate system at the center of the earth (assumed spherical). By Newton's law of gravity, the force of attraction of the earth (assumed to be of mass M) on a mass m situated at point P is given by

$$\mathbf{F} = -\frac{c}{r^3}\mathbf{r}, \quad c = GMm, \quad G = \text{the gravitational constant.}$$

Here \mathbf{r} denotes the position vector of point P . If we introduce Cartesian coordinates and $P : (x, y, z)$, then

$$\mathbf{F}(x, y, z) = \left(-\frac{c}{r^3}x, -\frac{c}{r^3}y, -\frac{c}{r^3}z\right) = -\frac{c}{r^3}(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}),$$

where $r = \|\mathbf{r}\|$.

Example 2.4. (*Electrostatic force field*) According to Coulomb's law, the force acting on a charge e at a position \mathbf{r} due to a charge Q at the origin is given by

$$\mathbf{F} = \frac{\epsilon Qe}{r^3}\mathbf{r},$$

ϵ being a constant. For $Qe > 0$ (like charges) the force is repulsive and for $Qe < 0$ (unlike charges) the force is attractive (Figure 2.3).

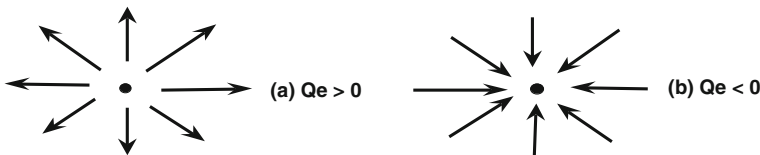


Fig. 2.3 Electrostatic force field

Example 2.5. One may be required to specify the reaction rate of a solution consisting of say five reacting chemicals C_1, C_2, \dots, C_5 . This requires a scalar function $\phi : D \subset R^5 \rightarrow R$ where $\phi(x_1, x_2, \dots, x_5)$ gives the rate when the chemicals are in the indicated proportion. Again, in this case, D becomes a scalar field.

Example 2.6. In medical diagnostics, for carrying out a *stress test*, it may be required to specify the *cardiac vector* (the vector giving the magnitude and direction of electric current flow in the heart) as it depends on time. This requires a vector-valued function $\mathbf{r} : R \rightarrow R^3$ (which is *not* a vector field).

2.5.2 Visualization of Scalar-Valued Functions

In view of Remark 2.3, we first consider a scalar-valued function $f : R^n \rightarrow R$ of n real variables. The *natural domain* \mathcal{D}_f of such a function is the set of all vectors $\mathbf{x} \in R^n$ such that $f(\mathbf{x}) \in R$.

Example 2.7. Let $f(x, y, z) = \sqrt{x^2 + y^2 + z^2}$. Here, $\mathcal{D}_f = R^3$ and $\mathcal{R}_f = \{x \in R : x \geq 0\}$.

Example 2.8. Let $g(x, y, z) = \sqrt{25 - x^2 - y^2 - z^2}$. Here, $\mathcal{D}_g = \{(x, y, z) : x^2 + y^2 + z^2 \leq 25\}$ and $\mathcal{R}_g = [0, 5]$.

Example 2.9. Let $h(x, y, z) = \sin(\frac{1}{xyz})$. Here, $\mathcal{D}_h = \{(x, y, z) \in R^3 : xyz \neq 0\}$ and $\mathcal{R}_h = [-1, 1]$.

Definition 2.8. (i) If $f : D \subset R^n \rightarrow R$ is a given function, then its *graph* $G(f)$ is the set

$$\{(x_1, \dots, x_n, f(x_1, \dots, x_n)) \in R^{n+1} : \mathbf{x} = (x_1, \dots, x_n) \in D\}$$

in R^{n+1} . For $n = 2$, the graph of the function $f(x, y)$ of two variables is the surface consisting of all the points (x, y, z) such that $(x, y) \in D$ and $z = f(x, y)$.

(ii) Given a function $f : R^n \rightarrow R$, its *level set at height* α is the set

$$\{\mathbf{x} = (x_1, \dots, x_n) : f(\mathbf{x}) = \alpha\}.$$

For $n = 2$, it is called a *level curve* and for $n = 3$, it is called a *level surface*.

Example 2.10. (i) For the function $f(x, y) = x + y + 1$, its graph is the plane $\{(x, y, z = x + y + 1) : (x, y) \in R^2\}$, and its level curves are the straight lines $x + y = c$.

(ii) For the function $f(x, y) = x^2 + y^2$, its graph is the paraboloid of revolution

$$\{(x, y, z = x^2 + y^2) : (x, y) \in R^2\}$$

and its level curves are circles $x^2 + y^2 = c^2$ in the xy -plane.

(iii) For the function $f(x, y) = \sqrt{4 - x^2 - y^2}$, its graph is the upper hemisphere of radius 2, and its level curves are circles $x^2 + y^2 = 4 - c^2$, $|c| \leq 2$.

2.6 Limits and Continuity

We first consider scalar-valued function of multivariables. For simplicity, we confine ourselves to functions of two variables. Let $(x_0, y_0) \in \mathbb{R}^2$ and $r > 0$. Recall that the *open ball* $B_r(x_0, y_0)$ of center (x_0, y_0) and radius r (for the Euclidean distance) is the set

$$\{(x, y) \in \mathbb{R}^2 : \|(x, y) - (x_0, y_0)\| = \sqrt{(x - x_0)^2 + (y - y_0)^2} < r\},$$

which is, in fact, the open disk of center (x_0, y_0) and radius r . Let us also denote by $\tilde{B}_r(x_0, y_0)$ the set

$$\{(x, y) \in \mathbb{R}^2 : |x - x_0| < r \text{ and } |y - y_0| < r\},$$

which is, in fact, the open square centered at (x_0, y_0) of side $2r$. It is easy to see that $\tilde{B}_r(x_0, y_0)$ is the open ball of center (x_0, y_0) and radius r for the metric

$$d_\infty((x_1, y_1), (x_2, y_2)) = \max\{|x_1 - x_2|, |y_1 - y_2|\}, (x_1, y_1), (x_2, y_2) \in \mathbb{R}^2.$$

Observe that

$$B_r(x_0, y_0) \subset \tilde{B}_r(x_0, y_0) \subset B_{\sqrt{2}r}(x_0, y_0).$$

This inclusion of open balls makes it possible to choose either of the two sets $B_r(x_0, y_0)$, $\tilde{B}_r(x_0, y_0)$ as a neighborhood of the point (x_0, y_0) .

2.6.1 Limits

The notion of the limit $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y)$ is defined analogously as in the univariate case. Let us recall that a point $(x_0, y_0) \in \mathbb{R}^2$ is called a *limit point* of the set $D \subset \mathbb{R}^2$ if for every $r > 0$, the neighborhood $B_r(x_0, y_0)$ of (x_0, y_0) contains a point of D other than (x_0, y_0) . Throughout in this subsection, we will make one of the following assumptions:

- (i) The natural domain \mathcal{D}_f of f contains a neighborhood $B_r(x_0, y_0)$ of (x_0, y_0) except possibly (x_0, y_0) itself;
- (ii) $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ is a given function and (x_0, y_0) is a limit point of D .

Definition 2.9. Assume either (i) or (ii) above holds. Given a number $L \in \mathbb{R}$, one says

$$f(x, y) \rightarrow L \text{ as } (x, y) \rightarrow (x_0, y_0),$$

written $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = L$ if for every $\epsilon > 0$, there corresponds a $\delta > 0$, such that

$$(x, y) \in D \cap B_\delta(x_0, y_0), (x, y) \neq (x_0, y_0) \Rightarrow |f(x, y) - L| < \epsilon.$$

In case, we are assuming (i), the above condition can be simply replaced by

$$(x, y) \in B_\delta(x_0, y_0), (x, y) \neq (x_0, y_0) \Rightarrow |f(x, y) - L| < \epsilon.$$

Remarks 2.5. (i) In the above definition, $B_\delta(x_0, y_0)$ can be replaced by $\tilde{B}_\delta(x_0, y_0)$.

(ii) Intuitively, the definition simply says that $f(x, y)$ comes arbitrarily close to L whenever (x, y) is sufficiently close to (x_0, y_0) .

(iii) Generalization of this definition to a function $f(x_1, \dots, x_n)$ of n variables is clear. Given a vector $\mathbf{x}^0 = (x_1^0, \dots, x_n^0) \in R^n$ and $r > 0$, we take the open ball $B_r(\mathbf{x}^0) = \{\mathbf{x} \in R^n : \|\mathbf{x} - \mathbf{x}^0\| < r\}$ as a neighborhood of the vector \mathbf{x}^0 . Under assumptions analogous to (i) and (ii), we proceed exactly as before to define the notion of the limit $\lim_{\mathbf{x} \rightarrow \mathbf{x}^0} f(\mathbf{x}) = L$.

Example 2.11. Let $f : R^3 \rightarrow R$ be defined by

$$f(x, y, z) = \begin{cases} \frac{xyz}{\sqrt{x^2 + y^2 + z^2}} \sin\left(\frac{1}{xyz}\right), & \text{if } x \neq 0, y \neq 0 \text{ and } z \neq 0 \\ 0, & \text{if } x = 0 \text{ or } y = 0 \text{ or } z = 0. \end{cases}$$

The natural domain of the function is $\mathcal{D}_f = \{(x, y, z) : x \neq 0, y \neq 0 \text{ and } z \neq 0\}$.

We have

$$|f(x, y, z) - 0| \leq \frac{|x| |y| |z|}{\sqrt{x^2 + y^2 + z^2}} \leq (x^2 + y^2 + z^2) \leq \epsilon,$$

whenever $0 < \sqrt{x^2 + y^2 + z^2} < \delta \leq \epsilon^{\frac{1}{2}}$. This shows that

$$\lim_{(x, y, z) \rightarrow (0, 0, 0)} f(x, y, z) = 0.$$

Theorem 2.2. Assume conditions (i) or (ii) as in the definition of limit. Let $y = \phi(x)$, $x \in [a, b]$ be a curve such that $x_0 \in (a, b)$ and $\lim_{x \rightarrow x_0} \phi(x) = y_0$. If

$$\psi(x) = f(x, \phi(x)), x \in [a, b], \text{ and } \lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y) = L,$$

then $\lim_{x \rightarrow x_0} \psi(x) = L$.

Proof: Exercise.

The above theorem gives the following test for nonexistence of a limit: If there is some curve as in the last theorem, along which the limit does not exist or that the limit is different along two different curves as (x, y) approaches (x_0, y_0) , then $\lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y)$ does not exist.

Example 2.12. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{xy^2}{x^2+y^4}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0). \end{cases}$$

Then, $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$ does not exist. Indeed, along the curve $x = my^2$, $y \neq 0$, the function has the constant value $f(my^2, y) = \frac{m}{1+m^2}$. Therefore,

$$\lim_{(x,y) \rightarrow (0,0) \text{ along } x=my^2} f(x, y) = \lim_{y \rightarrow 0} f(my^2, y) = \frac{m}{1+m^2}.$$

The limit is different for curves with different values of m . Hence, $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$ does not exist.

It is sometimes convenient to use polar coordinates for examining the limit of a function of two variables as illustrated by the next example.

Example 2.13. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{x^4y - 2x^3y^2 + 3x^2y^3 + y^5}{(x^2+y^2)^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0). \end{cases}$$

We have

$$|f(r \cos \theta, r \sin \theta) - 0| \leq r(1 + 2 + 3 + 1) = 7r = 7\sqrt{x^2 + y^2}.$$

Therefore, $|f(x, y) - 0| < \epsilon$, whenever $0 < \sqrt{x^2 + y^2} < \delta < \epsilon/7$. This shows that $\lim_{(x,y) \rightarrow (0,0)} f(x, y) = 0$.

2.6.2 Continuity

The notion of continuity of an univariate function extends easily to multivariate functions. As before, we start with a real-valued function of two variables.

Definition 2.10. (*continuity*) Let $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function and $(x_0, y_0) \in D$. We say that f is *continuous* at (x_0, y_0) if for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$(x, y) \in B_\delta(x_0, y_0) \Rightarrow |f(x, y) - f(x_0, y_0)| < \epsilon.$$

Further, we say that f is *continuous on* D if it is continuous at each point of D .

Remarks 2.6. (i) If (x_0, y_0) is a limit point of D , then f is continuous at $(x_0, y_0) \Leftrightarrow \lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y)$ exists and equals $f(x_0, y_0)$.

- (ii) If (x_0, y_0) is an isolated point of D , that is, for some $r > 0$, $B_r(x_0, y_0) \cap D = \{(x_0, y_0)\}$, then f is trivially continuous at (x_0, y_0) .
- (iii) Let $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function and $(x_0, y_0) \in D$. Then, f is continuous at $(x_0, y_0) \Leftrightarrow f$ is sequentially continuous at (x_0, y_0) . Recall that f is sequentially continuous at (x_0, y_0) if whenever a sequence $\{(x_n, y_n)\}$ in D is such that $(x_n, y_n) \rightarrow (x_0, y_0)$, we have $f(x_n, y_n) \rightarrow f(x_0, y_0)$.

Theorem 2.3. (*Composition of continuous functions*) (1): $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ is continuous at a point $(x_0, y_0) \in D$. (2): $E \subset \mathbb{R}$ is such that $\mathcal{R}_f \subset E$. (3): $g : E \rightarrow \mathbb{R}$ is continuous at $t_0 = f(x_0, y_0)$. Then, the composed function $g \circ f : D \rightarrow \mathbb{R}$ is continuous at (x_0, y_0) .

Proof: Prove this result as an exercise using the definition of norm in \mathbb{R}^m .

We remark that the notions of limit and continuity extend easily to vector-valued functions $\mathbf{F} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$. In fact, let $\mathbf{x}^0 \in \mathbb{R}^n$ be a limit point of D and let $\mathbf{L} \in \mathbb{R}^m$. Then, we say

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}^0} \mathbf{F}(\mathbf{x}) = \mathbf{L} \text{ provided } \lim_{\mathbf{x} \rightarrow \mathbf{x}^0} \|\mathbf{F}(\mathbf{x}) - \mathbf{L}\| = 0.$$

Remarks 2.7. If the functions $F_i : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$ are the components of the function $\mathbf{F} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$:

$$\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \dots, F_m(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^n,$$

then clearly

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}^0} \mathbf{F}(\mathbf{x}) = \mathbf{L} \Leftrightarrow \lim_{\mathbf{x} \rightarrow \mathbf{x}^0} F_i(\mathbf{x}) = L_i, \quad i = 1, \dots, m \text{ where } \mathbf{L} = (L_1, \dots, L_m).$$

For the sake of completeness, we mention that the following sandwich theorem also holds.

Theorem 2.4. (*Sandwich Theorem*) Let f, g, h be functions defined on $D \subset \mathbb{R}^n$ to \mathbb{R} . Let \mathbf{x}^0 be a limit point of D . If

$$g(\mathbf{x}) \leq f(\mathbf{x}) \leq h(\mathbf{x}), \quad \mathbf{x} \in D,$$

and

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}^0} g(\mathbf{x}) = L = \lim_{\mathbf{x} \rightarrow \mathbf{x}^0} h(\mathbf{x}),$$

then $\lim_{\mathbf{x} \rightarrow \mathbf{x}^0} f(\mathbf{x}) = L$.

Proof: Prove this as an exercise. (**Hint:** Let $H(\mathbf{x}) = h(\mathbf{x}) - g(\mathbf{x})$, $F(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x})$. Then, $0 \leq F(\mathbf{x}) \leq H(\mathbf{x})$ and $\lim_{\mathbf{x} \rightarrow \mathbf{x}^0} H(\mathbf{x}) = 0$.)

It is now clear that it looks natural to define a function $\mathbf{F} : D \subset R^n \rightarrow R^m$ to be continuous at a point $\mathbf{x}^0 \in D$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\mathbf{x} \in B_\delta(\mathbf{x}^0) \cap D \Rightarrow \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^0)\| < \epsilon.$$

Remark 2.8. (i) If $D \subset R^n$ and $\mathbf{x}^0 \in D$ is a limit point of D , then $\mathbf{F} : D \subset R^n \rightarrow R^m$ is continuous at $\mathbf{x}^0 \Leftrightarrow \lim_{\mathbf{x} \rightarrow \mathbf{x}^0} \mathbf{F}(\mathbf{x})$ exists and equals $\mathbf{F}(\mathbf{x}^0)$.

(ii) If \mathbf{x}^0 is an isolated point of D , then \mathbf{F} is trivially continuous at \mathbf{x}^0 .

(iii) $\mathbf{F} : D \subset R^n \rightarrow R^m$ is continuous at a point $\mathbf{x}^0 \in D$ if and only if each of its component function F_i is continuous at \mathbf{x}^0 , $i = 1, \dots, m$.

2.7 Partial Derivatives and Differentiability

2.7.1 Partial Derivatives

Let $f : D \subset R^2 \rightarrow R$ be a function, and let (x_0, y_0) be an **interior point** of D . By that we mean that there exists $\delta > 0$ such that $B_\delta(x_0, y_0) \subset D$. Let us recall that the partial derivative of f with respect to x at (x_0, y_0) denoted by $f_x(x_0, y_0)$ or $\frac{\partial f}{\partial x}(x_0, y_0)$ is this limit

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h},$$

if it exists. Geometrically, it is the slope of the tangent to the curve $z = f(x, y_0)$ obtained by intersecting the graph $(x, y, z = f(x, y_0))$, $(x, y) \in D$ of the function $f(x, y)$ with the plane $y = y_0$ at the point $(x_0, y_0, f(x_0, y_0))$. The other partial derivative $f_y(x_0, y_0)$ or $\frac{\partial f}{\partial y}(x_0, y_0)$ is defined analogously. More generally, if $f : D \subset R^n \rightarrow R$ and $\mathbf{x}^0 \in R^n$, we define the partial derivative of f with respect to x_i at \mathbf{x}^0 denoted by

$$D_i f(\mathbf{x}^0) \text{ or } \frac{\partial f}{\partial x_i}(\mathbf{x}^0)$$

as the limit

$$\lim_{h_i \rightarrow 0} \frac{f(x_1^0, \dots, x_i^0 + h_i, \dots, x_n^0) - f(x_1^0, \dots, x_n^0)}{h_i}$$

if it exists.

Example 2.14. (i) Let $f(x_1, x_2, x_3, x_4) = x_2 \sin(x_1 x_2) + e^{x_2} \cos(x_3) + x_4^2$. Then, $D_2 f = \sin(x_1 x_2) + x_1 x_2 \cos(x_1 x_2) + e^{x_2} \cos(x_3)$.

(ii) Let $f(x, y, z) = \sqrt{x^2 + y^2 + z^2}$. Then, for $(x_0, y_0, z_0) \neq (0, 0, 0)$, $f_x(x_0, y_0, z_0) = \frac{x_0}{\sqrt{x_0^2 + y_0^2 + z_0^2}}$ etc. However, it is easily seen that $f_x(0, 0, 0)$, $f_y(0, 0, 0)$, $f_z(0, 0, 0)$ do not exist.

2.7.2 Differentiability

In the case of an univariate function $f : (a, b) \rightarrow \mathbb{R}$, one says that f is differentiable at a point $x_0 \in (a, b)$ provided the derivative $f'(x_0)$ exists and that it is differentiable in (a, b) if it is differentiable at each point of (a, b) . Standard facts about univariate real-valued differentiable functions are that such functions are continuous and that the chain rule for differentiation applies to them. Going from univariate to multivariate functions, one may be tempted to believe that the existence of partial derivatives constitutes differentiability of such functions. The following example refutes such a belief.

Example 2.15. Let

$$f(x, y) = \begin{cases} 0, & \text{if } x = 0 \text{ or } y = 0 \\ 1, & \text{if } x \neq 0 \text{ and } y \neq 0. \end{cases}$$

Clearly, $f_x(0, 0) = f_y(0, 0) = 0$. However, $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$ does not exist, and hence, f is not continuous at $(0, 0)$.

This example shows that existence of partial derivatives alone is not adequate for its differentiability at a point. In the univariate case, the second approach for differentiability of a function $f : (a, b) \rightarrow \mathbb{R}$ at a point $x_0 \in (a, b)$ is the following. f is said to be differentiable at x_0 , provided we can write

$$f(x_0 + h) = f(x_0) + hf'(x_0) + h\eta(h), \quad |h| < \delta,$$

for a suitable $\delta > 0$ and a suitable function $\eta(h)$ defined in this range such that $\lim_{h \rightarrow 0} \eta(h) = 0$. Geometrically, this says f is differentiable at x_0 , provided the tangent line approximation $L(x) = f(x_0) + (x - x_0)f'(x_0)$ is a good approximation to f in a neighborhood of the point x_0 .

The next theorem is in the same spirit for a function of two variables.

Theorem 2.5. Let $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ be a given function, and let (x_0, y_0) be an interior point of D . The following statements are equivalent: (i) There exist numbers $\alpha, \beta \in \mathbb{R}$ such that

$$\lim_{(h,k) \rightarrow (0,0)} \frac{|f(x_0 + h, y_0 + k) - f(x_0, y_0) - \alpha h - \beta k|}{\sqrt{h^2 + k^2}} = 0. \quad (2.7.1)$$

(ii) There exist numbers $\alpha, \beta \in \mathbb{R}$ and functions $\epsilon_1(h, k), \epsilon_2(h, k)$ defined in an open ball $B_r(0, 0)$, for a suitable $r > 0$, such that

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = \alpha h + \beta k + \epsilon_1(h, k)h + \epsilon_2(h, k)k \quad (2.7.2)$$

and

$$\lim_{(h,k) \rightarrow (0,0)} \epsilon_1(h, k) = 0 = \lim_{(h,k) \rightarrow (0,0)} \epsilon_2(h, k). \quad (2.7.3)$$

Proof: (i) \Rightarrow (ii) : Suppose (2.7.1) holds. Define

$$\eta(h, k) := \begin{cases} \frac{f(x_0+h, y_0+k) - f(x_0, y_0) - \alpha h - \beta k}{\sqrt{h^2 + k^2}}, & (h, k) \neq (0, 0) \\ 0, & \text{if } (h, k) = (0, 0). \end{cases}$$

By (i), $\lim_{(h,k) \rightarrow (0,0)} \eta(h, k) = 0$, and

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = \alpha h + \beta k + \sqrt{h^2 + k^2} \eta(h, k).$$

We write

$$\sqrt{h^2 + k^2} \eta(h, k) = \frac{h^2 + k^2}{\sqrt{h^2 + k^2}} \eta(h, k),$$

and let

$$\epsilon_1(h, k) = \frac{h}{\sqrt{h^2 + k^2}} \eta(h, k), \quad \epsilon_2(h, k) = \frac{k}{\sqrt{h^2 + k^2}} \eta(h, k).$$

Then clearly,

$$\lim_{(h,k) \rightarrow (0,0)} \epsilon_1(h, k) = 0 = \lim_{(h,k) \rightarrow (0,0)} \epsilon_2(h, k)$$

and (2.7.3) holds. (ii) \Rightarrow (i): Note that (ii) implies

$$\frac{|f(x_0 + h, y_0 + k) - f(x_0, y_0) - \alpha h - \beta k|}{\sqrt{h^2 + k^2}} = \epsilon_1 \frac{h}{\sqrt{h^2 + k^2}} + \epsilon_2 \frac{k}{\sqrt{h^2 + k^2}},$$

which tends to 0 as $(h, k) \rightarrow (0, 0)$.

Definition 2.11. Let (x_0, y_0) be an interior point of $D \subset \mathbb{R}^2$ and $f : D \rightarrow \mathbb{R}$ be a function. The function f is said to be *differentiable at* (x_0, y_0) if f satisfies condition (ii) of the last theorem. If D is an open set, then f is said to be *differentiable in* D if it is differentiable at each point of D .

Theorem 2.6. Let $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function and let (x_0, y_0) be an interior point of D . If f is differentiable at (x_0, y_0) , then (a): f is continuous at (x_0, y_0) ; (b): both the partial derivatives of f exist at (x_0, y_0) . In fact $\alpha = f_x(x_0, y_0)$ and $\beta = f_y(x_0, y_0)$ where α, β are as in (ii) of the last theorem.

Proof: This is an easy exercise.

Theorem 2.7. (Increment Theorem) Let $D \subset \mathbb{R}^2$ and (x_0, y_0) be an interior point of D . If $f : D \rightarrow \mathbb{R}$ is such that the partial derivatives f_x, f_y exist at all points in an open ball $B(x_0, y_0)$ around (x_0, y_0) and these are continuous at (x_0, y_0) , then f is differentiable at (x_0, y_0) .

Proof: Proof using MVT is left as an exercise.

Remarks 2.9. The conditions in the Increment Theorem are only sufficient but not necessary for differentiability of f at (x_0, y_0) . As an example, consider

$$f(x, y) = \begin{cases} (x^2 + y^2) \sin\left(\frac{1}{x^2 + y^2}\right), & (x, y) \neq (0, 0) \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

Here $f_x(0, 0) = f_y(0, 0) = 0$ and f is differentiable at $(0, 0)$. In fact, we have

$$\epsilon_1(h, k) = h \sin\left(\frac{1}{h^2 + k^2}\right), \quad \epsilon_2(h, k) = k \sin\left(\frac{1}{h^2 + k^2}\right),$$

and both $\epsilon_1(h, k), \epsilon_2(h, k) \rightarrow 0$ as $(h, k) \rightarrow (0, 0)$. However,

$$f_x(x, y) = 2x \sin\left(\frac{1}{x^2 + y^2}\right) - \frac{2x}{x^2 + y^2} \cos\left(\frac{1}{x^2 + y^2}\right), \quad (x, y) \neq (0, 0),$$

and it is easily seen that along $y = 0$, $f_x(x, 0) \rightarrow -\infty$ as $x \rightarrow 0^+$ and $f_x(x, 0) \rightarrow \infty$ as $x \rightarrow 0^-$.

2.7.3 Chain Rule

We can easily extend the notion of differentiability to a function $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$. More precisely, let \mathbf{x}^0 be an interior point of D . By the discussion in the preceding subsection, we can define f to be differentiable at \mathbf{x}^0 if the partial derivatives $D_i f(\mathbf{x}^0)$, $i = 1, \dots, m$ exist and we have

$$\begin{aligned} f(x_1^0 + h_1, \dots, x_n^0 + h_n) - f(x_1^0, \dots, x_n^0) &= h_1 D_1 f(\mathbf{x}^0) + \dots \\ &+ h_n D_n f(\mathbf{x}^0) + \epsilon_1(\mathbf{h})h_1 + \dots + \epsilon_n(\mathbf{h})h_n \end{aligned} \quad (2.7.4)$$

where $\lim_{\mathbf{h} \rightarrow \mathbf{0}} \epsilon_i(\mathbf{h}) = 0$, $i = 1, \dots, m$. Writing

$$\mathbf{D}f(\mathbf{x}^0) = (D_1 f, \dots, D_n f)_{\mathbf{x}^0} \quad (2.7.5)$$

as the *total derivative* of f at \mathbf{x}^0 , we can write (2.7.4) in compact form as

$$f(\mathbf{x}^0 + \mathbf{h}) - f(\mathbf{x}^0) = \mathbf{D}f(\mathbf{x}^0) \cdot \mathbf{h} + \epsilon(\mathbf{h}) \cdot \mathbf{h} \quad (2.7.6)$$

where $\epsilon(\mathbf{h}) = (\epsilon_1(\mathbf{h}), \dots, \epsilon_n(\mathbf{h})) \rightarrow \mathbf{0}$ as $\mathbf{h} \rightarrow \mathbf{0}$. Frequently, in calculus, one writes $w = f(x_1, \dots, x_n)$ and (2.7.4) is written in the form

$$\Delta w = \Delta x_1 D_1 f(\mathbf{x}^0) + \dots + \Delta x_n D_n f(\mathbf{x}^0) + \epsilon_1 \Delta x_1 + \dots + \epsilon_n \Delta x_n, \quad (2.7.7)$$

by taking $h_i = \Delta x_i$, $i = 1, \dots, n$. We now consider the first version of the chain rule for the case under consideration.

Theorem 2.8. (1): $D \subset R^n$ and \mathbf{x}^0 is an interior point of D . (2): $f : D \subset R^n \rightarrow R$ is differentiable at \mathbf{x}^0 . (3): $x_1 = x_1(t), \dots, x_n = x_n(t)$ are functions defined from (a, b) to R , which are differentiable at $t_0 \in (a, b)$. (4): $(x_1(t_0), \dots, x_n(t_0)) = \mathbf{x}^0$ and $(x_1(t), \dots, x_n(t)) \in D$ for $t \in (a, b)$. Then, $W = f(x_1(t), \dots, x_n(t))$ is differentiable at t_0 and

$$W'(t_0) = \sum_{i=1}^n D'_i(\mathbf{x}^0) x'_i(t_0). \quad (2.7.8)$$

Proof: As t changes from t_0 to $t_0 + \Delta t$, the function x_i changes to $x_i + \Delta x_i$ where $\Delta x_i = x_i(t_0 + \Delta t) - x_i(t_0)$, $i = 1, \dots, n$. Differentiability of f entails

$$f(x_1^0 + \Delta x_1, \dots, x_n^0 + \Delta x_n) - f(x_1^0, \dots, x_n^0) = \sum_{i=1}^n D_i f(\mathbf{x}^0) \Delta x_i + \epsilon(\Delta \mathbf{x}) \cdot \Delta \mathbf{x}.$$

Divide both sides by Δt and let $\Delta t \rightarrow 0$, to obtain (2.7.8).

2.7.4 Gradient and Directional Derivatives

Let $D \subset R^n$ and \mathbf{x}^0 be an interior point of D . Let $f : D \subset R^n \rightarrow R$ be given. If the partial derivatives $D_i f(\mathbf{x}^0)$ all exist, then the vector $(D_1 f(\mathbf{x}^0), \dots, D_n f(\mathbf{x}^0))$ is called the *gradient* of f at \mathbf{x}^0 . It is denoted by $\nabla f(\mathbf{x}^0)$ or $\text{grad } f(\mathbf{x}^0)$. Now, fix up any direction $\mathbf{u} = (u_1, \dots, u_n)$ in R^n . The requirement that $\|\mathbf{u}\| = 1$ is not essential. The line through \mathbf{x}^0 in the direction \mathbf{u} has the equation: $\mathbf{x} = \mathbf{x}^0 + t\mathbf{u}$, $t \in R$. This gives rise to the parametric equations

$$x_i = x_i(t) = x_i^0 + tu_i, \quad i = 1, \dots, n.$$

Definition 2.12. The directional derivative of f at \mathbf{x}^0 in the direction \mathbf{u} is the limit

$$\lim_{t \rightarrow 0} \frac{f(\mathbf{x}^0 + t\mathbf{u}) - f(\mathbf{x}^0)}{t}$$

if it exists. It is denoted variously by $D_{\mathbf{u}} f(\mathbf{x}^0)$, $\frac{\partial f}{\partial \mathbf{u}}(\mathbf{x}^0)$, or $f'(\mathbf{x}^0, \mathbf{u})$.

The notion of directional derivative extends the notion of partial derivative: If $\mathbf{e}_i = (0, \dots, 1, 0, \dots, 0)$, then $D_{\mathbf{e}_i} f(\mathbf{x}^0)$, $i = 1, \dots, n$. It is clear from the definition that if $D_{\mathbf{u}} f(\mathbf{x}^0)$ exists, then $D_{-\mathbf{u}} f(\mathbf{x}^0)$ also exists, and $D_{-\mathbf{u}} f(\mathbf{x}^0) = -D_{\mathbf{u}} f(\mathbf{x}^0)$. The

next theorem is a crucial link between differentiability and the existence of directional derivatives of a function.

Theorem 2.9. (1): $D \subset R^n$ and \mathbf{x}^0 is an interior point of D . (2): $f : D \subset R^n \rightarrow R$ is differentiable at \mathbf{x}^0 . Then, in every direction $\mathbf{u} \in R^n$, the directional derivative $D_{\mathbf{u}}f(\mathbf{x}^0)$ exists and is equal to

$$\nabla f(\mathbf{x}^0) \cdot \mathbf{u} = D_1 f(\mathbf{x}^0)u_1 + \dots + D_n f(\mathbf{x}^0)u_n.$$

Proof: Indeed, the differentiability of f at \mathbf{x}^0 entails

$$f(\mathbf{x}^0 + t\mathbf{u}) - f(\mathbf{x}^0) = (tu_1)D_1 f(\mathbf{x}^0) + \dots + (tu_n)D_n f(\mathbf{x}^0) + t\mathbf{u} \cdot \epsilon(t\mathbf{u})$$

where $\epsilon(t\mathbf{u}) \rightarrow \mathbf{0}$ as $t \rightarrow 0$. Dividing both sides by t and letting $t \rightarrow 0$, we conclude that $D_{\mathbf{u}}f(\mathbf{x}^0)$ exists and is equal to $\nabla f(\mathbf{x}^0) \cdot \mathbf{u}$.

Remarks 2.10. (i): Note that existence of all partial derivatives need not imply existence of directional derivative $D_{\mathbf{u}}f(\mathbf{x}^0)$ in every direction. By way of an example, let

$$f(x, y) = \begin{cases} x + y, & \text{if } x = 0 \text{ or } y = 0 \\ 1, & \text{otherwise.} \end{cases}$$

Clearly, $f_x(0, 0) = f_y(0, 0) = 1$. Nevertheless, if we take $\mathbf{u} = (u_1, u_2)$, $u_1 \neq 0$, and $u_2 \neq 0$, then

$$\frac{f(\mathbf{0} + t\mathbf{u}) - f(\mathbf{0})}{t} = \frac{f(t\mathbf{u})}{t} = \frac{1}{t},$$

and this does not tend to a limit as $t \rightarrow 0$. (ii): The converse of the preceding theorem is false. For example, let

$$f(x, y) = \begin{cases} \frac{xy^2}{x^2+y^4}, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0. \end{cases}$$

It is easily seen that f is not continuous at $(0, 0)$; hence, it is, a fortiori, not differentiable at $(0, 0)$. Let $\mathbf{u} = (u_1, u_2)$. Then,

$$\frac{f(0 + tu_1, 0 + tu_2) - f(0, 0)}{t} = \frac{f(tu_1, tu_2)}{t} = \frac{u_1 u_2^2}{u_1^2 + t^2 u_2^4}.$$

Therefore,

$$D_{\mathbf{u}}f(0, 0) = \begin{cases} \frac{u_2^2}{u_1}, & \text{if } u_1 \neq 0 \\ 0, & \text{if } u_1 = 0. \end{cases}$$

The next example shows that the directional derivative $D_{\mathbf{u}}f(\mathbf{x}^0)$ may exist in every direction and f may be continuous, and yet f may be non-differentiable at \mathbf{x}^0 .

Example 2.16. Let

$$f(x, y) = \begin{cases} \frac{y}{|y|}\sqrt{x^2 + y^2}, & \text{if } y \neq 0 \\ 0, & \text{if } y = 0. \end{cases}$$

Clearly, $|f(x, y)| = \sqrt{x^2 + y^2}$, which implies f is continuous at $(0, 0)$. Fix up $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ such that $u_1^2 + u_2^2 = 1$. Then,

$$\frac{f(tu_1, tu_2) - f(0, 0)}{t} = \frac{\frac{tu_2}{|tu_2|}\sqrt{t^2u_1^2 + t^2u_2^2}}{t} = \frac{u_2}{|u_2|}, \quad u_2 \neq 0.$$

Therefore,

$$D_{\mathbf{u}}f(0, 0) = \begin{cases} \frac{u_2}{|u_2|}, & \text{if } u_2 \neq 0 \\ 0, & \text{if } u_2 = 0. \end{cases}$$

Also, it is easily seen that $f_x(0, 0) = 0$, $f_y(0, 0) = 1$, and $\text{grad } f \cdot \mathbf{u} \neq D_{\mathbf{u}}f(0, 0) = \frac{u_2}{|u_2|}$ if $u_2 \neq 0$ and $|u_2| \neq 1$. This shows that f is not differentiable at $(0, 0)$.

2.7.5 Tangent Plane and Normal Line

We consider a differentiable function $f : D \subset \mathbb{R}^3 \rightarrow \mathbb{R}$ defined in an open set D . If $C : \mathbf{r} = \mathbf{r}(t) = (x(t), y(t), z(t))$, $a \leq t \leq b$ is a smooth curve (i.e., we are assuming $\mathbf{r}(t)$ is continuous on $[a, b]$) on the level surface $S : f(x, y, z) = c$ of f , then $f(x(t), y(t), z(t)) = c$. Differentiating both sides of this equation with respect to t , by chain rule, we get

$$\frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt} = 0.$$

Equivalently, we can write $\nabla f \cdot \dot{\mathbf{r}} = 0$, which says that at every point along the curve C , ∇f is orthogonal to the tangent vector $\dot{\mathbf{r}}$ to C . Thus, if we fix up a point P_0 on S and consider all possible curves on S passing through P_0 , all the tangent lines to these curves are orthogonal to the vector $\nabla f(P_0)$. This motivates us to define

Definition 2.13. The *tangent plane* at the point $P_0 : (x_0, y_0, z_0)$ to the level surface $f(x, y, z) = c$ is the plane through P_0 which is orthogonal to the vector $\nabla f(P_0)$. The *normal line* to the surface at P_0 is the line through P_0 parallel to $\nabla f(P_0)$.

The vector equations of the tangent plane and normal line are, respectively,

$$(\mathbf{r} - \mathbf{r}_0) \cdot \nabla \mathbf{f}(P_0) = 0, \quad \mathbf{r} = \mathbf{r}_0 + t \nabla \mathbf{f}(P_0), \quad \mathbf{r}_0 = (x_0, y_0, z_0).$$

The corresponding scalar equations are, respectively,

$$\begin{aligned} f_x(P_0)(x - x_0) + f_y(P_0)(y - y_0) + f_z(P_0)(z - z_0) &= 0, \\ x &= x_0 + t f_x(P_0), \quad y = y_0 + t f_y(P_0), \quad z = z_0 + t f_z(P_0). \end{aligned}$$

If we are given a differentiable function $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ where D is an open set, then its graph $\{(x, y, z = f(x, y)) : (x, y) \in D\}$ is the level surface $F(x, y, z) = 0$ of the function

$$F(x, y, z) = f(x, y) - z,$$

whose partial derivatives are $F_x = f_x$, $F_y = f_y$, $F_z = -1$, respectively. In this case, the equations of the tangent plane and the normal line become, respectively,

$$\begin{aligned} f_x(P_0)(x - x_0) + f_y(P_0)(y - y_0) - (z - z_0), \\ x = x_0 + t f_x(P_0), \quad y = y_0 + t f_y(P_0), \quad z = z_0 - t. \end{aligned}$$

2.7.6 Differentiability of Vector-Valued Functions, Chain Rule

We are now ready for the general definition of differentiability of functions from \mathbb{R}^n to \mathbb{R}^m . Let $\mathbf{F} : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a given function and let \mathbf{x}^0 be an interior point of Ω . As before, we write

$$\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \dots, F_m(\mathbf{x})), \quad \mathbf{x} \in \Omega \quad (2.7.9)$$

where each F_i maps Ω to \mathbb{R} , so that F_i 's are the components of \mathbf{F} . Let us continue to denote by $D_j F_i$, the j^{th} partial derivative $\frac{\partial F_i}{\partial x_j}$ of F_i , $i = 1, \dots, m$, $j = 1, \dots, n$.

Definition 2.14. The *derivative matrix* of \mathbf{F} at \mathbf{x}^0 is the matrix

$$D\mathbf{F}(\mathbf{x}^0) = [D_j F_i]_{1 \leq i \leq m, 1 \leq j \leq n} \quad (2.7.10)$$

where the partial derivatives $D_j F_i$ evaluated at \mathbf{x}^0 are assumed to exist. We say that the vector-valued function \mathbf{F} is differentiable at \mathbf{x}^0 , provided these partial derivatives exist at \mathbf{x}^0 and if

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x}^0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}^0) - D\mathbf{F}(\mathbf{x}^0)\mathbf{h}\|}{\|\mathbf{h}\|} = 0. \quad (2.7.11)$$

In the last equation, we regard $\mathbf{h} = [h_1, \dots, h_n]^T$ as a n -column vector, so that it can be multiplied by the $m \times n$ matrix $\mathbf{DF}(\mathbf{x}^0)$. In case \mathbf{F} is differentiable at \mathbf{x}^0 , the derivative matrix of \mathbf{F} at \mathbf{x}^0 is sometimes called the *total derivative* of \mathbf{F} at \mathbf{x}^0 .

Remark 2.11. Applying Remark 2.7 to equation (2.7.11), it is immediately clear that $\mathbf{F} : \Omega \subset R^n \rightarrow R^m$ is differentiable at an interior point \mathbf{x}^0 of Ω if and only if each component function $F_i : \Omega \subset R^n \rightarrow R$, $i = 1, \dots, m$ is differentiable at \mathbf{x}^0 .

2.7.7 Particular Cases

(i) Let $m = 1$. Here, $F : \Omega \subset R^n \rightarrow R$ and

$$\mathbf{DF}(\mathbf{x}^0) = [D_1 F, \dots, D_n F]_{\mathbf{x}^0},$$

as already seen earlier.

(ii) Let $n = 1$ and $m > 1$. Here, $\mathbf{F} : (a, b) \subset R \rightarrow R^m$, written

$$\mathbf{F}(t) = (F_1(t), \dots, F_m(t)), \quad t \in (a, b),$$

which is a vector-valued function encountered frequently for parametrizing a curve in R^m . Clearly,

$$\mathbf{DF}(t) = [F'_1(t), \dots, F'_m(t)]^T$$

which is written frequently as $\mathbf{F}'(t)$ or $\dot{\mathbf{F}}(t)$ and it represents tangent vector to the curve $C : \mathbf{F} = \mathbf{F}(t)$, $t \in (a, b)$ in R^m .

For the sake of completeness, we state below without proof the following analogue of the *Increment Theorem* for the general case.

Theorem 2.10. Let $\mathbf{F} : \Omega \subset R^n \rightarrow R^m$ and \mathbf{x}^0 be an interior point of Ω . Writing \mathbf{F} in terms of its components

$$\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \dots, F_m(\mathbf{x})), \quad \mathbf{x} \in \Omega,$$

if the partial derivatives

$$D_j F_i, \quad i = 1, \dots, m; \quad j = 1, \dots, n$$

all exist in an open ball $B_\delta(\mathbf{x}^0)$ around \mathbf{x}^0 and are continuous at \mathbf{x}^0 , then \mathbf{F} is differentiable at \mathbf{x}^0 .

Remark 2.12. It follows from Remark 2.9 and the example therein that the conditions in the above theorem are only sufficient but not necessary for the differentiability of \mathbf{F} at \mathbf{x}^0 .

Next, we state, without proof, the general form of the chain rule.

Theorem 2.11. (Chain Rule) (1): $\mathbf{G} : \Omega \subset R^p \rightarrow R^n$ is differentiable at an interior point \mathbf{a} of Ω . (2): $\mathbf{F} : \Omega_1 \subset R^n \rightarrow R^m$ is differentiable at $\mathbf{b} = \mathbf{G}(\mathbf{a})$, which is an interior point of Ω_1 . Then, the composite function $\mathbf{H} : \Omega \subset R^p \rightarrow R^m$ defined by

$$\mathbf{H} = \mathbf{F} \circ \mathbf{G} : \mathbf{H}(\mathbf{x}) = \mathbf{F}(\mathbf{G}(\mathbf{x})), \quad \mathbf{x} \in R^p$$

is differentiable at \mathbf{a} and we have

$$\mathbf{DH}(\mathbf{a}) = \mathbf{DF}(\mathbf{b}) \cdot \mathbf{DG}(\mathbf{a}). \quad (2.7.12)$$

Note that the matrix on the left-hand side is of order $m \times p$, which is a product of an $m \times n$ matrix with an $n \times p$ matrix. The following special case is encountered frequently in multivariable calculus: $F : \Omega_1 \subset R^n \rightarrow R$, and $\mathbf{G} : \Omega \subset R^p \rightarrow R^n$. Here, $m = 1$, and $H : \Omega \rightarrow R$. We have

$$\mathbf{DH}(\mathbf{a}) = [D_1 H, \dots, D_p H]_{\mathbf{a}}, \quad \mathbf{DF}(\mathbf{b}) = [D_1 F, \dots, D_n F]_{\mathbf{b}},$$

and $\mathbf{DG}(\mathbf{a}) = [D_j G_i]_{i=1, \dots, n; j=1, \dots, p}(\mathbf{a})$. Hence, from (2.7.12),

$$[D_1 H, \dots, D_p H]_{\mathbf{a}} = [D_1 F, \dots, D_n F]_{\mathbf{b}} [D_j G_i]_{\mathbf{a}}.$$

This gives the familiar formulae using chain rule:

$$D_j H(\mathbf{a}) = \sum_{i=1}^n D_i F(\mathbf{b}) D_j G_i(\mathbf{a}), \quad j = 1, \dots, p. \quad (2.7.13)$$

For example, let $f : R^3 \rightarrow R$, $\mathbf{g} : R^3 \rightarrow R^3$ be differentiable. Write

$$g(u, v, w) = (x(u, v, w), y(u, v, w), z(u, v, w))$$

and define $h : R^3 \rightarrow R$ by setting

$$h(u, v, w) = f(x(u, v, w), y(u, v, w), z(u, v, w)).$$

Then,

$$\begin{bmatrix} \frac{\partial h}{\partial u} & \frac{\partial h}{\partial v} & \frac{\partial h}{\partial w} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{bmatrix}.$$

which gives the familiar chain rule for three intermediate and three independent variables.

2.7.8 Differentiation Rules

Here and in the sequel, it would be convenient for us to regard R^n as the space of column n -vectors $\mathbf{x} = [x_1, \dots, x_n]^T$ with real entries. Let us recall that if the function $f : R^n \rightarrow R$ is differentiable, then the function $\nabla : R^n \rightarrow R^n$ called the *gradient* of f is defined by

$$\nabla f(\mathbf{x}) = \mathbf{D}f(\mathbf{x})^T = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right]^T.$$

Given $f : R^n \rightarrow R$, if ∇f is differentiable, then f is said to be *twice differentiable*, and we write the derivative of ∇f as

$$\mathbf{D}^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

The matrix $\mathbf{D}^2 f(\mathbf{x})$ is called the *Hessian matrix* of f at \mathbf{x} . Let us also recall that a function $\mathbf{F} : \Omega \rightarrow R^m$, where Ω is an open subset of R^n is said to be continuously differentiable in Ω if it is differentiable, and $\mathbf{D}\mathbf{F} : \Omega \rightarrow R^{m \times n}$ is continuous. Here, $R^{m \times n}$ denotes the space of all $m \times n$ matrices with real entries. (This amounts to saying that the components of \mathbf{F} have continuous first partial derivatives.) In this case, we write $\mathbf{F} \in \mathcal{C}^{(1)}(\Omega)$.

Theorem 2.12. *Let Ω be an open subset of R^n , $g : \Omega \rightarrow R$ be differentiable in Ω , and let $\mathbf{f} : (a, b) \rightarrow \Omega$ be differentiable in (a, b) . Then, $h = g \circ \mathbf{f} : (a, b) \rightarrow R$ defined by $h(t) = g(\mathbf{f}(t))$, $t \in (a, b)$ is differentiable in (a, b) and*

$$h'(t) = Dg(\mathbf{f}(t))\mathbf{D}\mathbf{f}(t) = \nabla g(\mathbf{f}(t))^T [f_1'(t), \dots, f_n'(t)]^T, \quad t \in [a, b].$$

Theorem 2.13. *Let $\mathbf{f} : R^n \rightarrow R^m$, $\mathbf{g} : R^m \rightarrow R^m$ be two differentiable functions. Let $h : R^n \rightarrow R$ be defined by*

$$h(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \mathbf{g}(\mathbf{x}) = \langle \mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x}) \rangle. \quad (2.7.14)$$

Then, h is differentiable and

$$Dh(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T D\mathbf{g}(\mathbf{x}) + \mathbf{g}(\mathbf{x})^T D\mathbf{f}(\mathbf{x}). \quad (2.7.15)$$

Remarks 2.13. (Some useful formulae)

Let $\mathbf{A} \in R^{m \times n}$ and $\mathbf{y} \in R^m$ be given. Then, we have:

(i):

$$D(\mathbf{y}^T \mathbf{A} \mathbf{x}) = \mathbf{y}^T \mathbf{A}, \quad \mathbf{x} \in R^n. \quad (2.7.16)$$

(ii):

$$D(\mathbf{x}^T \mathbf{A} \mathbf{x}) = \mathbf{x}^T (\mathbf{A} + \mathbf{A}^T), \quad \mathbf{x} \in R^n. \quad (2.7.17)$$

2.7.9 Taylor's Expansion

Let us recall the following order symbols. Let $g : B \rightarrow R$ be a function defined in an open ball B around $\mathbf{0} \in R^n$ such that $g(\mathbf{x}) \neq 0$ for $\mathbf{x} \neq \mathbf{0}$, and let $f : C \rightarrow R^m$ be defined in a set $C \subseteq R^n$ that contains $\mathbf{0} \in R^n$. Then, (a) : the symbol

$$\mathbf{f}(\mathbf{x}) = O(g(\mathbf{x}))$$

means that $\frac{\|\mathbf{f}(\mathbf{x})\|}{|g(\mathbf{x})|}$ is bounded near $\mathbf{0}$. More precisely, there are numbers $K > 0$ and $\delta > 0$ such that

$$x \in C \text{ and } \|x\| < \delta \Rightarrow \frac{\|\mathbf{f}(\mathbf{x})\|}{|g(\mathbf{x})|} \leq K.$$

(b) : The symbol $f(\mathbf{x}) = o(g(\mathbf{x}))$ means that

$$\lim_{\mathbf{x} \rightarrow \mathbf{0}, \mathbf{x} \in C} \frac{\|\mathbf{f}(\mathbf{x})\|}{|g(\mathbf{x})|} = 0.$$

Theorem 2.14. (Taylor's Theorem) Let $f : R \rightarrow R$ be in $C^m[a, b]$, and $0 \leq h \leq b - a$. Then,

$$f(a + h) = f(a) + \frac{h}{1!} f'(a) + \dots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + R_m \quad (2.7.18)$$

where $R_m = \frac{h^m}{m!} f^{(m)}(a + \theta h)$, for a suitable $\theta \in (0, 1)$.

Remarks 2.14. Note that since $f \in C^m[a, b]$, $f^{(m)}(a + \theta h) = f^{(m)} + o(1)$. Thus, if $f \in C^{(m)}$, then we can write Taylor's formula as

$$f(a + h) = f(a) + \frac{h}{1!} f'(a) + \dots + \frac{h^m}{m!} f^{(m)}(a) + o(h^m). \quad (2.7.19)$$

In addition, if we assume that $f \in C^{(m+1)}[a, b]$, then $R_{m+1} = \frac{h^{m+1}}{(m+1)!} f^{(m+1)}(a + \theta h)$, and since $f^{(m+1)}$ is bounded, we can conclude that $R_{m+1} = O(h^{m+1})$. Thus, in this case, we can write Taylor's formula as

$$f(a + h) = f(a) + \frac{h}{1!} f'(a) + \dots + \frac{h^m}{m!} f^{(m)}(a) + O(h^{m+1}). \quad (2.7.20)$$

Theorem 2.15. Let $f : \Omega \rightarrow R$, where Ω is an open subset of R^n , be a function in class $C^2(\Omega)$. Let $\mathbf{x}_0 \in \Omega$ and \mathbf{x} be in an open ball around \mathbf{x}_0 contained in Ω . Let $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$. Then,

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \frac{1}{1!} \mathbf{D}f(\mathbf{x}_0)\mathbf{h} + \frac{1}{2!} \mathbf{h}^T \mathbf{D}^2 f(\mathbf{x}_0)\mathbf{h} + o(\|\mathbf{h}\|^2). \quad (2.7.21)$$

Moreover, if $f \in \mathcal{C}^3(\Omega)$, then we have:

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \frac{1}{1!} \mathbf{D}f(\mathbf{x}_0)\mathbf{h} + \frac{1}{2!} \mathbf{h}^T \mathbf{D}^2 f(\mathbf{x}_0)\mathbf{h} + O(\|\mathbf{h}\|^2). \quad (2.7.22)$$

Proof: This is left as an exercise by considering the function $\mathbf{y}(t) = \mathbf{x}_0 + t(\frac{\mathbf{x} - \mathbf{x}_0}{\|\mathbf{x} - \mathbf{x}_0\|})$ for $0 \leq t \leq \|\mathbf{x} - \mathbf{x}_0\|$. Let $h : R \rightarrow R$ be defined by $h(t) = f(\mathbf{y}(t))$. Use chain rule and the univariate Taylor theorem, Theorem 2.15 to complete the proof.

2.8 Introduction to Optimization

Let us begin by recalling the so-called *Max-Min Theorem* which asserts that a continuous function $f : D \subset R^n \rightarrow R$ defined on a compact subset D of R^n is *bounded* and that it attains its (global) maximum and its (global) minimum at some points of D . Since the topology of R^n that we are using is *metrizable*, by Heine–Borel theorem, saying that D is compact is equivalent to saying that D is closed and bounded. More importantly, it is equivalent to saying that every sequence $\mathbf{x}^{(n)}$ in D has a convergent subsequence $\mathbf{x}^{(n_k)}$ converging in D .

We intend to give here a slightly more general result than the above stated result. For this purpose, let us observe that since $\max_D f = -\min_D(-f)$, the problem of maximizing f is equivalent to the problem of minimizing $-f$. Thus, without loss of generality, we may confine ourself to the minimization problem. We need the following definitions.

Definition 2.15. Let X be a normed linear space. A function $f : X \rightarrow R \cup \{\infty\}$ is said to be (i): *inf-compact*, if for each $\alpha \in R$, the sublevel set of f at height α :

$$\text{lev}_\alpha f = \{x \in X : f(x) \leq \alpha\}$$

is compact. It is said to be (ii): *lower semi-continuous(lsc)*, if $\text{lev}_\alpha f$ is closed for each $\alpha \in R$. Furthermore, f is said to be (iii): *coercive*, if $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$.

Remarks 2.15. It is clear from the definitions that for a function $f : X \rightarrow R \cup \{\infty\}$ defined on a normed space X , f is coercive if and only if f is *inf-bounded*, that is to say that the sublevel set $\text{lev}_\alpha f$ of f at height α is bounded for each $\alpha \in R$. As a result, we see that if $X = R^n$ with the usual topology, then for a function $f : R^n \rightarrow R \cup \{\infty\}$ which is lsc, f is coercive if and only if f is inf-compact.

One main reason for bringing in extended real-valued functions in optimization is that these provide a flexible modelization of minimization problems with constraints. Most minimization problems can be formulated as

$$\min\{f_0(x) : x \in \Omega\} \quad (2.8.1)$$

where $f_0 : X \rightarrow R$ is a real-valued function, and $\Omega \subseteq X$ is the so-called *constraint set* or *feasible set*, X being some vector space, usually R^n . A natural way of dealing with such a problem is to apply penalization to it. For example, introduce a distance d on X and for any positive real number λ , let us consider the minimization problem

$$\min\{f_0(x) + \lambda \operatorname{dist}(x, \Omega) : x \in X\} \quad (2.8.2)$$

where

$$\operatorname{dist}(x, \Omega) = \inf\{d(x, y) : y \in \Omega\} \quad (2.8.3)$$

is the distance function from x to Ω . Let us note that the penalization is equal to zero if $x \in \Omega$ (that is if the constraint is satisfied), and when $x \notin \Omega$ it takes larger and larger values with λ (when the constraint is violated). Notice that the approximated problem (2.8.2) can be written as

$$\min\{f_\lambda(x) : x \in X\} \quad (2.8.4)$$

where

$$f_\lambda(x) = f_0(x) + \lambda \operatorname{dist}(x, \Omega) \quad (2.8.5)$$

is a real-valued function. Thus, the approximated problems (2.8.4) are unconstrained problems. As $\lambda \rightarrow +\infty$, the (generalized) sequence of functions (2.8.5) increases to the function $f : X \rightarrow R \cup \{\infty\}$, which is equal to

$$f(x) = \begin{cases} f_0(x), & \text{if } x \in \Omega, \\ +\infty, & \text{otherwise.} \end{cases} \quad (2.8.6)$$

Thus, we are led to minimization of an extended real-valued function f :

$$\min\{f(x) : x \in X\}$$

where f is given by (2.8.6). Let us note that if we introduce the *indicator function* δ_Ω of the set Ω :

$$\delta_\Omega(x) = \begin{cases} 0, & \text{if } x \in \Omega, \\ +\infty, & \text{otherwise,} \end{cases} \quad (2.8.7)$$

then we have the convenient expression $f = f_0 + \delta_\Omega$. We now come to the following generalization of the *Min-Max Theorem* called the extended Weierstrass theorem.

Theorem 2.16. *Let X be a normed linear space and let $f : X \rightarrow R \cup \{+\infty\}$ be an extended real-valued function which is lower semi-continuous and inf-compact. Then, $\inf_X f > -\infty$, and there exists some $\hat{x} \in X$ which minimizes f on X :*

$$f(\hat{x}) \leq f(x) \text{ for all } x \in X.$$

Proof: Given a function $f : X \rightarrow R \cup \{+\infty\}$, by the definition of $\inf_X f$, the infimum of f on X , we can construct a *minimizing sequence*, that is, a sequence $\{x_n\}$ such that $f(x_n) \rightarrow \inf_X f$ as $n \rightarrow +\infty$. Indeed, if $\inf_X f > -\infty$, pick $\{x_n\}$ such that

$$\inf_X f \leq f(x_n) \leq \inf_X f + \frac{1}{n};$$

and if $\inf_X f = -\infty$, pick $\{x_n\}$ such that $f(x_n) \leq -n$. Without any restriction, we may assume that f is proper (that is finite somewhere), and hence, $\inf_X f < +\infty$. Thus, for $n \in N$,

$$f(x_n) \leq \max\{\inf_X f + 1/n, -n\} \leq \max\{\inf_X f + 1, -1\} := \alpha_0.$$

Since, $\alpha_0 > \inf_X f$, $\text{lev}_{\alpha_0}(f) \neq \emptyset$, and $x_n \in \text{lev}_{\alpha_0}(f)$, $n \in N$. By inf-compactness of f , this sublevel set in which the sequence $\{x_n\}$ is contained is compact. Hence, we can extract a subsequence $\{x_{n_k}\}$ converging to some element $\hat{x} \in X$. By lower semi-continuity of f , we have

$$f(\hat{x}) \leq \lim_k f(x_{n_k}) = \inf_X f.$$

This proves that $\inf_X f > -\infty$, since $f : X \rightarrow R \cup \{+\infty\}$, and

$$f(\hat{x}) \leq f(x) \text{ for all } x \in X,$$

which completes the proof.

Remarks 2.16. The set of all the minimizers of f on X is usually denoted by $\arg \min_X(f)$. The above theorem gives the conditions under which $\arg \min_X(f) \neq \emptyset$. As a corollary of the preceding theorem, we have the following:

Corollary 2.1. (Weierstrass theorem) *Let X be a normed linear space and K be a compact subset of X . If $f : X \rightarrow R \cup \{+\infty\}$ is lower semi-continuous, then $\arg \min_X(f) \neq \emptyset$.*

2.8.1 Unconstrained and Constrained Extremizers: Conditions for Local Extremizers

We consider here the minimization problem (D, f) :

$$\min\{f(\mathbf{x}) : \mathbf{x} \in D\}.$$

Here $f : R^n \rightarrow R$ is a given function called the *objective function* or *cost function*, the vector \mathbf{x} is called the vector of *decision variables* x_1, \dots, x_n , and the set D is a

subset of R^n , called the *constraint set* or *feasible set*. The above problem is a general *constrained* minimization problem. In case $D = R^n$, one refers to the problem as *unconstrained* minimization problem. Constrained and unconstrained maximization problems are defined analogously.

Definition 2.16. Let $f : D \subseteq R^n \rightarrow R$ be a given function and let \mathbf{x}^0 be an interior point of D . Then, (i): \mathbf{x}^0 is called a *local minimizer* of f , if there is some $\delta > 0$ such that

$$f(\mathbf{x}) \geq f(\mathbf{x}^0), \text{ for all } \mathbf{x} \in B_\delta(\mathbf{x}^0).$$

In this case, $f(\mathbf{x}^0)$ is called a *local minimum value* of f . (ii): A *local maximizer* \mathbf{x}^0 of f and *local maximum value* $f(\mathbf{x}^0)$ of f are defined analogously. (iii): A point $\mathbf{x}^0 \in D$ is called a *global minimizer* of f if

$$f(\mathbf{x}^0) \leq f(\mathbf{x}), \forall \mathbf{x} \in D.$$

A *global maximizer* of f is defined analogously. A local (resp.global) minimizer or maximizer of f is called a *local (resp.global) extremizer* of f .

Let us consider the constrained minimization problem (D, f) where D, f are as given before. A minimizer \mathbf{x}^0 of problem (D, f) can be either an interior point or a boundary point of D . The following definition is useful in this respect.

Definition 2.17. A vector $\mathbf{u} \in R^n, \mathbf{u} \neq \mathbf{0}$ is called a *feasible direction* (f.d.) at $\mathbf{x}^0 \in D$, if there exists $\delta > 0$ such that $\mathbf{x}^0 + \lambda \mathbf{u} \in D$ for all $\lambda \in [0, \delta]$.

Let us recall Definition 2.16 of the directional derivative $f'(\mathbf{x}^0; \mathbf{u})$, which is also sometimes written as $\frac{df}{d\mathbf{u}}$, given by

$$f'(\mathbf{x}^0; \mathbf{u}) = \mathbf{f} \nabla f(\mathbf{x}^0)^T \mathbf{u} = \langle \nabla f(\mathbf{x}^0), \mathbf{u} \rangle = \mathbf{u}^T \nabla f(\mathbf{x}^0),$$

in case f is differentiable at \mathbf{x}^0 . Note that if \mathbf{u} is a unit vector, then

$$|f'(\mathbf{x}^0; \mathbf{u})| = | \langle \nabla f(\mathbf{x}^0), \mathbf{u} \rangle | \leq |\nabla f(\mathbf{x}^0)|.$$

Theorem 2.17. (First-Order Necessary Condition) Let D be a subset of R^n and $f : D \rightarrow R$ be a $C^{(1)}$ function. If $\mathbf{x}^0 \in D$ is a local minimizer of f , then for every f.d. \mathbf{u} at \mathbf{x}^0 , we have

$$\nabla f(\mathbf{x}^0)^T \mathbf{u} \geq 0.$$

Proof: Note that \mathbf{u} is a f.d. at $\mathbf{x}^0 \Rightarrow \exists \delta > 0$ such that $\mathbf{x}^0 + \lambda \mathbf{u} \in D, \forall \lambda \in [0, \delta]$. Let us define

$$\phi(\lambda) = f(\mathbf{x}^0 + \lambda \mathbf{u}), \lambda \in [0, \delta].$$

By Taylor's theorem, we have

$$\begin{aligned}
f(\mathbf{x}^0 + \lambda \mathbf{u}) - f(\mathbf{x}^0) &= \phi(\lambda) - \phi(0) \\
&= \lambda \phi'(\lambda) + o(\lambda) \\
&= \lambda \nabla f(\mathbf{x}^0)^T \mathbf{u} + o(\lambda).
\end{aligned}$$

Since \mathbf{x}^0 is a local minimizer of f , $\phi(\lambda) - \phi(0) \geq 0$ for sufficiently small values of $\lambda > 0$. This implies that $\nabla f(\mathbf{x}^0)^T \mathbf{u} \geq 0$.

Remarks 2.16. (i): The theorem holds under the weaker assumption: f is differentiable at \mathbf{x}^0 .

(ii): An alternative way to express the first-order necessary condition is $f'(\mathbf{x}^0; \mathbf{u}) \geq 0$ for every f.d. \mathbf{u} at \mathbf{x}^0 .

Corollary 2.2. (Interior point case) Let $D \subseteq \mathbb{R}^n$, \mathbf{x}^0 be an interior point of D , and $f : D \rightarrow \mathbb{R}$ be differentiable at \mathbf{x}^0 . If \mathbf{x}^0 is a local minimizer of f , then we have

$$\nabla f(\mathbf{x}^0) = \mathbf{0}.$$

Proof: Indeed, if \mathbf{x}^0 is an interior point of D , then the set of f.d.s at \mathbf{x}^0 is the whole of \mathbb{R}^n . Thus, for any $\mathbf{u} \in \mathbb{R}^n$, both $\nabla f(\mathbf{x}^0)^T \mathbf{u} \geq 0$ as well as $\nabla f(\mathbf{x}^0)^T (-\mathbf{u}) \geq 0$ hold. Thus, $\nabla f(\mathbf{x}^0) = \mathbf{0}$.

Theorem 2.18. Let $D \subseteq \mathbb{R}^n$ and $f : D \rightarrow \mathbb{R}$ be a \mathcal{C}^2 function. If $\mathbf{x}^0 \in D$ is a local minimizer of f and \mathbf{u} is a f.d. at \mathbf{x}^0 such that $\nabla f(\mathbf{x}^0)^T \mathbf{u} = 0$, then

$$\mathbf{u}^T \mathbf{H}_f(\mathbf{x}^0) \mathbf{u} \geq 0.$$

Here, $\mathbf{H}_f(\mathbf{x}^0)$ denotes the Hessian of f at \mathbf{x}^0 .

Proof: This is left as an exercise to the reader. One makes use of Theorem 2.11.

Remark 2.18. Theorem holds under the weaker assumption: f is twice differentiable at \mathbf{x}^0 .

Corollary 2.3. Let \mathbf{x}^0 be an interior point of $D \subseteq \mathbb{R}^n$, $f : D \rightarrow \mathbb{R}$ be a given function. If f is twice differentiable at \mathbf{x}^0 and \mathbf{x}^0 is a local minimizer of f , then $\nabla f(\mathbf{x}^0) = \mathbf{0}$, and $\mathbf{H}_f(\mathbf{x}^0)$ is positive semi-definite: $\mathbf{u}^T \mathbf{H}_f(\mathbf{x}^0) \mathbf{u} \geq 0$, $\forall \mathbf{u} \in \mathbb{R}^n$.

Proof: Note that \mathbf{x}^0 is an interior point of D implies all directions $\mathbf{u} \in \mathbb{R}^n$ are feasible. Hence, the result follows from the preceding theorem and the last corollary.

Exercises I

2.1. Find the natural domains of the following functions of two variables.

$$(i) \frac{xy}{x^2 - y^2}, \quad (ii) \ln(x^2 + y^2 + z^2), \quad (iii) \sqrt{25 - x^2 - y^2 - z^2}, \quad (iv) \frac{1}{xyz}.$$

2.2. (a): Use level curves to sketch graphs of the following functions.

$$(i) f(x, y) = x^2 + y^2, \quad (ii) f(x, y) = x^2 + y^2 - 4x - 6y + 13.$$

(b): Sketch some level surfaces of the following functions.

$$(i) f(x, y, z) = 4x^2 + y^2 + 9z^2, \quad (ii) f(x, y, z) = x^2 + y^2.$$

2.3. Using definition, examine the following functions for continuity at $(0, 0)$. The expressions below give the value at $(x, y) \neq (0, 0)$. At $(0, 0)$, the value should be taken as zero.

$$(i) \frac{x^3 y}{x^6 + y^2}, \quad (ii) \frac{x^2 y}{x^2 + y^2}, \quad (iii) xy \frac{x^2 - y^2}{x^2 + y^2},$$

$$(iv) ||x| - |y|| - |x| - |y|, \quad (v) \frac{\sin^2(x + y)}{|x| + |y|}.$$

2.4. Using definition, examine the following functions for continuity at $(0, 0, 0)$. The expressions below give the value at $(x, y, z) \neq (0, 0, 0)$. At $(0, 0, 0)$, the value should be taken as zero.

$$(i) \frac{xyz}{\sqrt{x^2 + y^2 + z^2}} \sin\left(\frac{1}{x^2 + y^2 + z^2}\right), \quad (ii) \frac{2xy}{x^2 - 3z^2}, \quad (iii) \frac{xy + yz + zx}{\sqrt{x^2 + y^2 + z^2}}.$$

2.5. Suppose $f, g : \mathbb{R} \rightarrow \mathbb{R}$ are continuous functions. Show that each of the following functions of $(x, y) \in \mathbb{R}^2$ is continuous.

$$(i) f(x) \pm g(y), \quad (ii) f(x)g(y), \quad (iii) \max\{f(x), g(y)\}, \quad (iv) \min\{f(x), g(y)\}.$$

2.6. Using the results from the above problem, show that $f(x, y) = x + y$ and $g(x, y) = xy$ are continuous in \mathbb{R}^2 . Deduce that every polynomial function in two variables is continuous in \mathbb{R}^2 .

2.7. Examine each of the following functions for continuity.

$$(i) f(x, y) = \begin{cases} \frac{y}{|y|} \sqrt{x^2 + y^2}, & \text{if } y \neq 0 \\ 0, & \text{if } y = 0. \end{cases}$$

$$(ii) f(x, y) = \begin{cases} x \sin \frac{1}{x} + y \sin \frac{1}{y}, & \text{if } x \neq 0, y \neq 0 \\ x \sin \frac{1}{x}, & \text{if } x \neq 0, y = 0 \\ y \sin \frac{1}{y}, & \text{if } x = 0, y \neq 0 \\ 0, & \text{if } x = 0, y = 0. \end{cases}$$

2.8. Let $f(x, y) = \frac{x^2 y^2}{x^2 y^2 + (x-y)^2}$, for $(x, y) \neq (0, 0)$. Show that the iterated limits $\lim_{x \rightarrow 0}[\lim_{y \rightarrow 0} f(x, y)]$ and $\lim_{y \rightarrow 0}[\lim_{x \rightarrow 0} f(x, y)]$ exist and both are equal to 0, but $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$ does not exist.

2.9. Express the definition of $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$ in terms of polar coordinates and analyze it for the following functions:

- (i) $f(x, y) = \frac{x^3 - xy^2}{x^2 + y^2}$. (ii) $f(x, y) = \tan^{-1} \left(\frac{|x| + |y|}{x^2 + y^2} \right)$.
 (iii) $f(x, y) = \frac{y^2}{x^2 + y^2}$. (iv) $f(x, y) = \frac{x^4 y - 3x^2 y^3 + y^5}{(x^2 + y^2)^2}$.

Exercises II

2.10. Examine the following functions for the existence of partial derivatives at $(0, 0)$. The expressions below give the value at $(x, y) \neq (0, 0)$. At $(0, 0)$, the value should be taken as zero.

- (i) $\frac{x^3 y}{x^6 + y^2}$, (ii) $xy \frac{x^3}{x^2 + y^2}$, (iii) $\frac{x^2 y}{x^4 + y^2}$
 (iv) $xy \frac{x^2 - y^2}{x^2 + y^2}$, (v) $||x| - y| - |x| - |y|$, (vi) $\frac{\sin^2(x+y)}{|x| + |y|}$.

2.11. Let $D \subset \mathbb{R}^2$ be an open disk centered at (x_0, y_0) and $f : D \rightarrow \mathbb{R}$ be such that both f_x and f_y exist and are bounded in D . Prove that f is continuous at (x_0, y_0) . Conclude that f is continuous everywhere in D .

2.12. Let $f(0, 0) = 0$ and $f(x, y) = (x^2 + y^2) \sin \frac{1}{x^2 + y^2}$ for $(x, y) \neq (0, 0)$. Show that f is continuous at $(0, 0)$, and the partial derivatives of f exist but are not bounded in any disk (however small) around $(0, 0)$.

2.13. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(0, 0) = 0$ and $f(x, y) = \frac{xy}{x^2 + y^2}$ if $(x, y) \neq (0, 0)$. Show that f_x and f_y exist at every $(x_0, y_0) \in \mathbb{R}^2$, but f is not continuous at $(0, 0)$.

2.14. Let $f(0, 0) = 0$ and

$$f(x, y) = \begin{cases} x \sin(\frac{1}{x}) + y \sin(\frac{1}{y}), & \text{if } x \neq 0, y \neq 0 \\ x \sin(\frac{1}{x}), & \text{if } x \neq 0, y = 0 \\ y \sin(\frac{1}{y}), & \text{if } y \neq 0, x = 0. \end{cases}$$

Show that none of the partial derivatives of f exist at $(0, 0)$ although f is continuous at $(0, 0)$.

2.15. Suppose the implicit equation $F(x, y, z) = 0$ determines z as a function of x and y , that is, there exists a function w of two variables such that $F(x, y, w(x, y)) = 0$. Assume that F_x, F_y, F_z exist and are continuous, $F_z \neq 0$ and w_y exists. Show that $w_y = -F_y/F_z$.

2.16. Suppose $F : R^3 \rightarrow R$ has the property that there exists $n \in N$ such that $F(tx, ty, tz) = t^n F(x, y, z)$ for all $t \in R$ and $(x, y, z) \in R^3$. [Such a function is said to be *homogeneous* of degree n .] If the first-order partial derivatives of f exist and are continuous, then show that $x \frac{\partial F}{\partial x} + y \frac{\partial F}{\partial y} + z \frac{\partial F}{\partial z} = nF$. [This result is sometimes called *Euler's Theorem*.]

2.17. Examine the following functions for the existence of directional derivatives and differentiability at $(0, 0)$. The expressions below give the value at $(x, y) \neq (0, 0)$. At $(0, 0)$, the value should be taken as zero.

(i) $xy \frac{x^2 - y^2}{x^2 + y^2}$, (ii) $\frac{x^3}{x^2 + y^2}$, (iii) $(x^2 + y^2) \sin \frac{1}{x^2 + y^2}$.

2.18. Let $f(x, y) = 0$ if $y = 0$ and $f(x, y) = \frac{y}{|y|} \sqrt{x^2 + y^2}$ if $y \neq 0$. Show that f is continuous at $(0, 0)$, $D_{\mathbf{u}} f(0, 0)$ exists for every vector \mathbf{u} , yet f is not differentiable at $(0, 0)$.

2.19. Assume that f_x and f_y exist and are continuous in some disk centered at $(1, 2)$. If the directional derivative of f at $(1, 2)$ toward $(2, 3)$ is $2\sqrt{2}$ and toward $(1, 0)$ is -3 , then find $f_x(1, 2)$, $f_y(1, 2)$ and the directional derivative of f at $(1, 2)$ toward $(4, 6)$.

Exercises III

2.20. Given $z = x^2 + 2xy$, $x = u \cos v$ and $y = u \sin v$, find $\frac{\partial z}{\partial u}$ and $\frac{\partial z}{\partial v}$.

2.21. Given $\sin(x + y) + \sin(y + z) = 1$, find $\frac{\partial^2 z}{\partial x \partial y}$, provided $\cos(y + z) \neq 0$.

2.22. If $f(0, 0) = 0$ and $f(x, y) = xy \frac{x^2 - y^2}{x^2 + y^2}$ for $(x, y) \neq (0, 0)$, show that both f_{xy} and f_{yx} exist at $(0, 0)$, but they are not equal. Are f_{xy} and f_{yx} continuous at $(0, 0)$?

2.23. Show that the following functions have local minima at the indicated points.

- (i) $f(x, y) = x^4 + y^4 + 4x - 32y - 7$, $(x_0, y_0) = (-1, 2)$
 (ii) $f(x, y) = x^3 + 3x^2 - 2xy + 5y^2 - 4y^3$, $(x_0, y_0) = (0, 0)$.

2.24. Analyze the following functions for local maxima, local minima, and saddle points:

- (i) $f(x, y) = (x^2 - y^2)e^{-\frac{1}{2}(x^2 + y^2)}$, (ii) $f(x, y) = x^3 - 3xy^2$,
 (iii) $f(x, y) = 6x^2 - 2x^3 + 3y^2 + 6xy$, (iv) $f(x, y) = x^3 + y^3 - 3xy + 15$,
 (v) $f(x, y) = (x^2 + y^2) \cos(x + 2y)$.

2.25. Find the absolute minimum and the absolute maximum of $f(x, y) = 2x^2 - 4x + y^2 - 4y + 1$ on the closed triangular plate bounded by the lines $x = 0$, $y = 2$ and $y = 2x$.

2.26. Find the absolute maximum and the absolute minimum of $f(x, y) = (x^2 - 4x) \cos y$ over the region $1 \leq x \leq 3$, $-\frac{\pi}{4} \leq y \leq \frac{\pi}{4}$.

2.27. The temperature at a point (x, y, z) in 3-space is given by $T(x, y, z) = 400xyz$. Find the highest temperature on the unit sphere $x^2 + y^2 + z^2 = 1$.

2.28. Find the point nearest to the origin on the surface defined by the equation $z = xy + 1$

2.29. Find the absolute maximum and minimum of $f(x, y) = \frac{1}{2}x^2 + \frac{1}{2}y^2$ in the elliptic region D defined by $\frac{1}{2}x^2 + y^2 \leq 1$.

2.30. A space probe in the shape of the ellipsoid $4x^2 + y^2 + 4z^2 = 16$ enters the earth's atmosphere and its surface begins to heat. After one hour, the temperature at the point (x, y, z) on the surface of the probe is given by $T(x, y, z) = 8x^2 + 4yz - 16z + 600$. Find the hottest and the coolest points on the surface of the probe.

2.31. Maximize the quantity xyz subject to the constraints $x + y + z = 40$ and $x + y = z$.

2.32. Minimize the quantity $x^2 + y^2 + z^2$ subject to the constraints $x + 2y + 3z = 6$ and $x + 3y + 4z = 9$.

2.33. Consider the minimization problem: minimize $x_1^2 + \frac{1}{2}x_2^2 + 3x_2 + \frac{9}{2}$ subject to $x_1, x_2 \geq 0$. Answer the following questions: (In the following, we abbreviate first-order necessary condition by f.o.n.c.)

- (a) Is the f.o.n.c. for a local minimizer satisfied at $\mathbf{x}_0 = [1, 3]^T$?
- (b) Is the f.o.n.c. for a local minimizer satisfied at $\mathbf{x}_0 = [0, 3]^T$?
- (c) Is the f.o.n.c. for a local minimizer satisfied at $\mathbf{x}_0 = [1, 0]^T$?
- (d) Is the f.o.n.c. for a local minimizer satisfied at $\mathbf{x}_0 = [0, 0]^T$?

2.34. Consider the quadratic function $f : R^2 \rightarrow R$ given below:

$$f(\mathbf{x}) = \mathbf{x}^T \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix} \mathbf{x} + \mathbf{x}^T [3, 5]^T + 6$$

- (a) Find the gradient and Hessian of f at the point $[1, 1]^T$.
- (b) Find the directional derivative of f at $[1, 1]^T$ with respect to a unit vector in the direction in which the function decreases most rapidly.
- (c) Find a point that satisfies the f.o.n.c.(interior case) for f . Does this point satisfy the second-order necessary condition(s.o.n.c.)(for a minimizer)?

2.35. Consider the problem: minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in \Omega$, where $f : R^2 \rightarrow R$ is given by $f(\mathbf{x}) = 5x_2$ with $\mathbf{x} = [x_1, x_2]^T$, and $\Omega = \{\mathbf{x} = [x_1, x_2]^T : x_1^2 + x_2 \geq 1\}$. Answer each of the following questions giving justification.

- (a) Does the point $\mathbf{x}_0 = [0, 1]^T$ satisfy the f.o.n.c.?
- (b) Does the point $\mathbf{x}_0 = [0, 1]^T$ satisfy the s.o.n.c.?
- (c) Is the point $\mathbf{x}_0 = [0, 1]^T$ a local minimizer?

2.9 Classical Approximation Problems: A Relook

2.9.1 Input–Output Process

In, many practical situations of interest, such as in *economic forecasting*, *weather prediction*, etc, one is required to construct a model for what is generally called an *input-output process*. For example, one may be interested in the price of a stock 5 years from now. The *rating industry* description of a stock typically lists such indicators as, the increase in the price over the last 1 year or the increase in the price over the last 2 years/ 3 years, the life of the stock, P/E ratio, α , β risk factors etc. The investor is expected to believe that the price of the stock depends on these parameters. Of course, no one knows precisely a formula to compute this price as a closed form function of these parameters. (Else, one would strike it rich quickly!) Typically, this is an example of an input-output process.

$$\text{input data } \mathbf{x} \in R^n \longrightarrow \text{actual process} \longrightarrow \text{output } f(\mathbf{x})$$

where $f(\mathbf{x})$ is the so-called *target function*. In practice, the computed model is as follows:

$$\text{finite data} \longrightarrow \text{computed model} \longrightarrow P_f(\mathbf{x}).$$

Here, the finite data may consist of the values of the function or the values of some of its derivatives sampled at some points or possibly it may consist of the Fourier coefficients or coefficients in some other series associated with f , etc.

Broadly speaking, there are two kinds of errors in using the model $P_f(\mathbf{x})$ as a predictor for the target function $f(\mathbf{x})$.

(a) **Noise:** This comes from the fact that the observations on which the model is based are usually subjected to errors-human errors, machine errors, interference from nature, etc. Also, this could arise from faulty assumptions about what one is modeling. Statistics is mainly concerned with the *reliability* of the model by eliminating or controlling the noise.

(b) **Intrinsic Error:** This comes from the fact that one is computing a *model* rather than the *actual function*. Approximation Theory is concerned for the most part with “intrinsic error.”

2.9.2 Approximation by Algebraic Polynomials

Under the setting of an input-output process in Section 2.9.1, the problems of *Approximation Theory* can be broadly classified into four categories.

Let I denote the compact interval $[a, b]$ of R , and let $X = \mathcal{C}(I)$ denote the space of functions

$$\{f : I \rightarrow R : f \text{ continuous} \}$$

normed by

$$\|f\| = \max\{|f(t)| : t \in I\}, f \in \mathcal{C}(I).$$

Let us assume that our target function f be in the class $\mathcal{C}(I)$. The *model* for such a function is typically a real algebraic polynomial

$$P_n(t) = \sum_{k=0}^n a_k x^k, a_k \in R, k = 0, \dots, n.$$

The integer n is called the degree of the polynomial and we denote by Π_n the class of all real algebraic polynomials of degree $\leq n$. The first problem that we encounter in this context is the *density problem*: to decide whether it is feasible to approximate the target function arbitrarily closely by choosing more and more complex models. More precisely, let

$$E_n(f) := \inf_{P \in \Pi_n} \|f - P\|$$

denote the *degree of approximation* of f from Π_n . The density problem is to decide whether $E_n(f) \rightarrow 0$ as $n \rightarrow \infty$. It is the classical Weierstrass approximation theorem that answers this question affirmatively.

The next problem is the so-called *complexity problem* which is concerned with estimating the rate at which $E_n(f) \rightarrow 0$. Frequently in applications, the target function is usually unknown, but one may assume that $f \in \mathcal{F}$, where \mathcal{F} is a certain class of functions. For example, \mathcal{F} may consist of functions f such that f' exists, is continuous, and satisfies $\|f'\| \leq 1$. Typically, then, one is interested in estimating $\sup_{f \in \mathcal{F}} E_n(f)$.

The next problem is addressed in the *theory of best approximation*. This deals with the existence, uniqueness, characterization, and computability of elements $P \in \Pi_n$ such that

$$\|f - P\| = E_n(f).$$

Lastly, the *theory of good approximation* deals with the approximation capabilities of different procedures for computing the approximants based on the values of the function, its derivatives, its Fourier coefficients, etc. It is to be emphasized here that these approximants are sometimes more interesting than the best approximant, P , because of the *relative ease of its computability* or because of some desirable properties such as *shape preservation*. For a more detailed exposition of ideas in this direction, the reader may consult the book by Mhaskar and Pai [3].

2.10 Introduction to Optimal Recovery of Functions

Let us begin by considering what is usually called the problem of *best simultaneous approximation* in the literature in Approximation Theory. Here, the main concern is with *simultaneous approximation* which is best approximation in some sense of

sets, rather than that of just single elements. One of the motivations for treating this problem is the following practical situation involving *optimal estimation*. We assume that in the mathematical modeling of a physical process an entity E is represented by an unknown element x_E of some normed linear space X . Mostly, X is a function space such as $\mathcal{C}([a, b], R)$, $H^m[a, b]$, etc. By means of suitable experiments, observations are obtained for E which give rise to *limited information* concerning x_E . For instance, the information could be the values of x_E and/some of its derivatives sampled on a discrete set of points, or it could be the Fourier coefficients of x_E etc. In addition, the information could be error contaminated due to experimental inaccuracies. We assume that the information so gathered is incomplete to specify x_E completely; it only identifies a certain subset $F \subset X$ called the *data set*. Our estimation problem then is to find the best estimate of x_E given only that $x_E \in F$ (Figure 2.4).

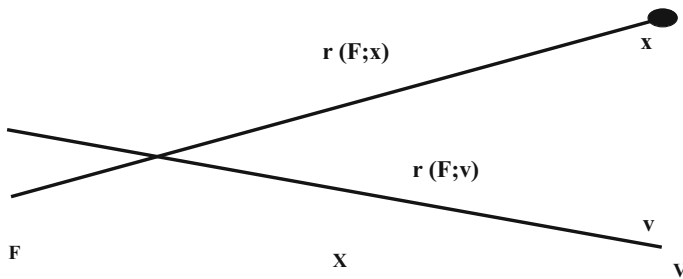


Fig. 2.4 Optimal estimation

We pick an element $x \in X$ (usually, a normed linear space) and ask how bad it is from the point of view of representing the data set F . The measure of *worstness* of x as a representer of F is given by the quantity

$$r(F; x) := \sup\{\|x - y\| : y \in F\} \quad (2.10.1)$$

(in a *worst-case scenario*). In order that this quantity be finite, we must assume that F be bounded. The *intrinsic error* in our estimation problem is then determined by the number

$$\text{rad}(F) := \inf\{r(F; x) : x \in X\},$$

called the *Chebyshev radius* of F . It is impossible for the worstness of x as a representer of F to fall below this number. An element $x_0 \in X$ will be a best representer (or a global approximator) of the data set F if it minimizes the measure of worstness:

$$r(F; x_0) = \min\{r(F; x) : x \in X\}. \quad (2.10.2)$$

An element $x_0 \in X$ satisfying (2.10.2) is called a *Chebyshev center* of F and $\text{Cent}(F)$ denotes (the possibly void) set of all Chebyshev centers of F .

Practical reasons may require us to restrict our search of a best representer of the data set F to another set V which may perhaps be a subspace or a convex set obtained by taking the intersection of a subspace with a set determined by affine constraints, etc. In this case, the intrinsic error in our estimation problem will be determined by the number

$$\text{rad}_V(F) := \inf \{r(F; v) : v \in V\} \quad (2.10.3)$$

called the (*restricted*) *radius* of F in V and a best representer (or global approximator) of F in V will be an element $v_0 \in V$ called (*restricted*) *center* of F in V satisfying

$$r(F; v_0) = \text{rad}_V(F). \quad (2.10.4)$$

For a more detailed exposition of ideas in this direction, the reader may consult Chapter VIII of the book by Mhaskar and Pai [3]. Let us observe that the above stated problem of *Optimal Estimation* is closely related to the so-called problem of *optimal recovery* of functions. Following Micchelli and Rivlin [4, 5], by *optimal recovery* we will mean the problem of estimating some required feature of a function, known to belong to some class of functions prescribed *a priori*, from limited and possibly error-contaminated information about it, as effectively as possible. This problem is again subsumed by what goes on under *information-based complexity* [7] and it has some rich connections with the problem of *image reconstruction* or *image recovery* which, for instance, is important in mathematical studies of computer-assisted tomography. We begin by looking at some simple examples given below, motivating the general theory.

2.10.1 Some Motivating Examples

Example 2.17. Let $X = \mathcal{C}[0, 1]$ and let

$$\mathcal{K} := \{x \in \mathcal{C}^{(n)}[0, 1] : \|x^{(n)}\| \leq 1\}.$$

Suppose we are given:

(i): $x \in \mathcal{K}$, and (ii): the values $x(t_1), x(t_2), \dots, x(t_n)$ sampled at distinct points t_1, t_2, \dots, t_n in $[0, 1]$. Then, we ask the following:

(I): Given $t_0 \in [0, 1] \setminus \{t_1, \dots, t_n\}$, what is the *best possible* estimate of $x(t_0)$, based solely on the information (i) and (ii)? (II): What is the *best possible* estimate of x itself based solely on the information (i) and (ii)?

Here, (I) is a problem of *optimal interpolation* and (II) is a problem of *optimal approximation*. We proceed to answer Question (I) first. Specially, let

$$\mathcal{I} := \{(x(t_1), \dots, x(t_n)) : x \in \mathcal{K}\}.$$

An *algorithm* A is any function of \mathcal{I} into R . Then, the *error* in the algorithm is given by

$$E_A(\mathcal{K}) := \sup_{x \in \mathcal{K}} |x(t_0) - A(x(t_1), \dots, x(t_n))|,$$

and

$$E(\mathcal{K}) := \inf_A E_A(\mathcal{K})$$

denotes the *intrinsic error* in our recovery problem. An algorithm \hat{A} , if one such exists, satisfying $E_{\hat{A}}(\mathcal{K}) = E(\mathcal{K})$ is called an *optimal algorithm*, which is said to effect the *optimal recovery* of $x(t_0)$. It is easily seen that the polynomial

$$\hat{P}(t) := \frac{(t - t_1) \dots (t - t_n)}{n!} \in \mathcal{K}$$

and so also does the polynomial $-\hat{P}$. Thus, if A is any algorithm, then

$$|\hat{P}(t_0) - A(\hat{P}(t_1), \dots, \hat{P}(t_n))| = |\hat{P}(t_0) - A(0, \dots, 0)| \leq E_A(\mathcal{K})$$

and also

$$|-\hat{P}(t_0) - A(0, \dots, 0)| \leq E_A(\mathcal{K}),$$

whence, by triangle inequality, $|\hat{P}(t_0)| \leq E_A(\mathcal{K})$, and we obtain

$$|\hat{P}(t_0)| \leq E(\mathcal{K}), \quad (2.10.5)$$

which gives a lower bound for the intrinsic error. Now, consider the algorithm $\hat{A} : \mathcal{I} \rightarrow R$ given by

$$\hat{A}(x(t_1), \dots, x(t_n)) = L_{t_1, t_2, \dots, t_n}(x; t_0)$$

where $L_{t_1, t_2, \dots, t_n}(x; t)$ denotes the unique Lagrange interpolant in $P_n(R)$ of x on the nodes t_i s. Suppose we show that

$$|x(t_0) - \hat{A}(x(t_1), \dots, x(t_n))| \leq |\hat{P}(t_0)|,$$

for all $x \in \mathcal{K}$, then

$$E_{\hat{A}}(\mathcal{K}) \leq |\hat{P}(t_0)| \leq E(\mathcal{K}),$$

and \hat{A} would be an optimal algorithm sought, with $E(\mathcal{K}) = |\hat{P}(t_0)|$. Assume the contrary that

$$x(t_0) - L_{t_1, t_2, \dots, t_n}(x; t_0) = \alpha \hat{P}(t_0),$$

for some α , $|\alpha| > 1$. Then, the function

$$h(t) := x(t) - L_{t_1, \dots, t_n}(x; t) - \alpha \hat{P}(t)$$

would have $n + 1$ distinct zeros t_0, t_1, \dots, t_n . Since $h \in \mathcal{C}^{(n-1)}[0, 1]$, by Rolle's theorem, $h^{(n)} = x^{(n)} - \alpha$ would have at least one zero $\xi : x^{(n)}(\xi) - \alpha = 0$. Hence,

$$\|x^{(n)}\| \leq 1,$$

would be contradicted. Thus \hat{A} is an optimal algorithm sought.

Remarks 2.19. If our object is to find the best possible estimate of $x^{(m)}(t_j)$ for some fixed $m \leq n$ and j , $1 \leq j \leq n$, based solely on (i) \wedge (ii), then one can show on the same lines as above that $|\hat{P}^{(m)}(t_j)| \leq E(\mathcal{K})$, and that the algorithm $\tilde{A} : \mathcal{I} \rightarrow R$ defined by

$$\tilde{A}(x(t_1), \dots, x(t_n)) = L_{t_1, t_2, \dots, t_n}^{(m)}(x; t_j)$$

is an optimal algorithm.

Next, we address Question (II). An algorithm now is any function $A : \mathcal{I} \rightarrow X$, and

$$E_A(\mathcal{K}) := \sup \|x - A(x(t_1), \dots, x(t_n))\|_\infty$$

is the error in algorithm A . By exactly the same reasoning as above, we obtain $\|\hat{P}\|_\infty \leq E_A(\mathcal{K})$, which yields $\|\hat{P}\|_\infty \leq E(\mathcal{K})$ as a lower bound for the intrinsic error. Consider the algorithm $A^* : \mathcal{I} \rightarrow X$ given by

$$A^*(x(t_1), \dots, x(t_n)) = L_{t_1, t_2, \dots, t_n}(x; t).$$

We claim that

$$\|x - A^*(x(t_1), \dots, x(t_n))\|_\infty \leq \|\hat{P}\|_\infty, \text{ for all } x \in \mathcal{K}.$$

Indeed, if we assume the contrary, then

$$|x(t_0) - L_{t_1, \dots, t_n}(x; t_0)| > |\hat{P}(t_0)|, \text{ for some } t_0 \in [0, 1],$$

which contradicts Rolle's theorem, exactly as before. Thus

$$E_{A^*}(\mathcal{K}) \leq \|\hat{P}\|_\infty \leq E(\mathcal{K}),$$

and we conclude that A^* is an optimal algorithm for the recovery problem involving optimal approximation. The problem of finding *optimal sampling nodes* in this

case amounts to minimizing the quantity $E(\mathcal{K}) = \|\hat{P}\|_\infty$ over all distinct points $t_1, \dots, t_n \in [0, 1]$. This problem is easily solved thanks to two well-known results in classical Approximation Theory, cf., e.g., Theorem 2.4.1 and Corollary 2.2.7, in the book Mhaskar and Pai [3]. Indeed,

$$\min \{E(\mathcal{K}) : t_1, \dots, t_n \in [0, 1] \text{ distinct}\} = \left(\frac{1}{2^{n-1}}\right) \left(\frac{1}{n!}\right),$$

and if we denote by $\tilde{t}_k := \cos\left(\frac{2k-1}{2n}\pi\right)$, $k = 1, 2, \dots, n$ the zeros of the n th Chebyshev polynomial $T_n(t) := \cos(n \cos^{-1}(t))$, then the optimal sampling nodes are precisely the Chebyshev nodes on $[0, 1]$,

$$\hat{t}_k = \frac{1}{2}[1 + \tilde{t}_k], \quad k = 1, 2, \dots, n.$$

Example 2.18. Let T be a compact subset of R^m and let $X = L_\infty(T)$. Let

$$\mathcal{K} := \{x \in L_\infty(T) : |x(t_1) - x(t_2)| \leq \|t_1 - t_2\| \text{ for all } t_1, t_2 \in T\} \quad (2.10.6)$$

denote the set of all *non-expansive functions* in X . We can ask analogous questions as in Example 2.17.

Question 1 (Optimal interpolation): With \mathcal{K} as above and given $t_0 \in T$, what is the best possible estimate of $x(t_0)$ based solely on the information (i) and (ii) (cf. Example 2.17)?

Let $I : X \rightarrow R^n$ be defined by $Ix := (x(t_1), \dots, x(t_n))$, and let

$$\mathcal{I} := I(\mathcal{K}) = \{(x(t_1), \dots, x(t_n)) : x \in \mathcal{K}\}.$$

A lower bound on the intrinsic error of our estimation problem is easily seen to be given by

$$E(\mathcal{K}) \geq \sup \{|x(t_0)| : x \in \mathcal{K}, x|_\Delta = 0\} \quad (2.10.7)$$

where $\Delta := \{t_1, t_2, \dots, t_n\}$. Let

$$T_i := \left\{t \in T : \min_j \|t - t_j\| = \|t - t_i\|\right\},$$

and let $\hat{q}(t) := \min_j \|t - t_j\|$. Clearly, $\hat{q} \in \mathcal{K}$ and $\hat{q}|_\Delta = 0$. Thus, if $t_0 \in T_k$, then $E(\mathcal{K}) \geq \hat{q}(t_0) = \|t_0 - t_k\|$. Define the algorithm $\tilde{A} : \mathcal{I} \rightarrow R$ by

$$\hat{A}(x(t_1), \dots, x(t_n)) = x(t_k).$$

Then,

$$|x(t_0) - \hat{A}(Ix)| = |x(t_0) - x(t_k)| \leq \|t_0 - t_k\| = \hat{q}(t_0) \text{ for all } x \in \mathcal{K}.$$

Consequently, $E_{\hat{A}}(\mathcal{K}) = \hat{q}(t_0)$ and \hat{A} is an optimal algorithm.

Question 2 (Optimal approximation): In the same setting as above with \mathcal{K} as defined in (I), what is the best possible estimate of x itself based solely on the information (i) and (ii)?

Note that an algorithm now is any function $A : \mathcal{I} \rightarrow L_\infty(T)$. An analogue of (2.10.7) is:

$$\begin{aligned} E(\mathcal{K}) &\geq \sup\{\|x\|_\infty : x \in \mathcal{K}, x|_\Delta = 0\} \\ &\geq \|\hat{q}\|_\infty. \end{aligned}$$

Let us denote by s the step function

$$s(t) = x(t_i), \quad t \in \text{int } T_i, \quad i = 1, \dots, n.$$

Then, $s \in L_\infty(T)$ and the algorithm $\tilde{A} : \mathcal{I} \rightarrow L_\infty(T)$ defined by $\tilde{A}(x(t_1), \dots, x(t_n)) = s$ is an optimal algorithm. Indeed, if $t \in \text{int } T_i$ and $x \in \mathcal{K}$ then

$$|x(t) - \tilde{A}(Ix)(t)| = |x(t) - x(t_i)| \leq \|t - t_i\| = \hat{q}(t).$$

Thus $\|x - \tilde{A}(Ix)\|_\infty \leq \|\hat{q}\|_\infty$, and we conclude that

$$E_{\tilde{A}}(\mathcal{K}) \leq \|\hat{q}\|_\infty.$$

Question 3 (Optimal integration): With the same setting and the same information as in the previous two questions, we look for an optimal recovery of $\int_T x dt$.

In this case, a lower bound for the intrinsic error is given by

$$\begin{aligned} E(\mathcal{K}) &\geq \sup \left\{ \left| \int_T x dt \right| : x \in \mathcal{K}, x|_\Delta = 0 \right\} \\ &\geq \int_T \hat{q} dt. \end{aligned}$$

Consider the algorithm $A_0 : \mathcal{I} \rightarrow R$ given by

$$A_0(x(t_1), \dots, x(t_n)) = \int_T s dt = \sum_{i=1}^n x(t_i) \text{vol}(T_i).$$

We have

$$\left| \int_T x dt - A_0(Ix) \right| = \left| \int_T x dt - \int_T s dt \right| \leq \int_T |x - s| dt \leq \int_T \hat{q} dt.$$

Thus $E_{A_0}(\mathcal{K}) \leq \int_T \hat{q} dt$ and we conclude that A_0 is an optimal algorithm.

2.10.2 General Theory

Let X be a linear space and Y, Z be normed linear spaces. Let K be a balanced convex subset of X . Let $T : X \rightarrow Z$ be a linear operator (the so-called *feature operator*). Our object is to estimate Tx for $x \in K$ using “limited information” about x . A linear operator $I : X \rightarrow Y$ called the *information operator* is prescribed and we assume Ix for $x \in K$ is known possibly with some error (Figure 2.5).

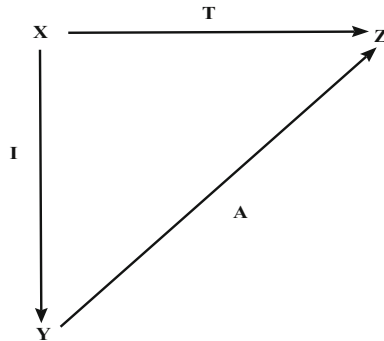


Fig. 2.5 Schematic representation

The algorithm A produces an error. Thus, while attempting to recover Tx for $x \in K$, we only know $y \in Y$ satisfying $\|Ix - y\| \leq \epsilon$ for some preassigned $\epsilon \geq 0$. An *algorithm* is any function—not necessarily a linear one—from $IK + \epsilon U(Y)$ into Z .

Schematically,

$$E_A(K, \epsilon) := \sup\{\|Tx - Ay\| : x \in K \text{ and } \|y - Ix\| \leq \epsilon\},$$

and

$$E(K, \epsilon) := \inf_A E_A(K, \epsilon)$$

is the *intrinsic error* in our estimation problem.

Any algorithm \hat{A} satisfying $E_{\hat{A}}(K, \epsilon) = E(K, \epsilon)$ is called an *optimal algorithm*. When $\epsilon = 0$, which corresponds to the recovery problem with *exact information*, we simply denote $E_A(K, 0)$ and $E(K, 0)$ by $E_A(K)$, $E(K)$ respectively. A lower bound for $E(K, \epsilon)$ is given by the next proposition.

Theorem 2.19. *We have*

$$e(K, \epsilon) := \sup\{\|Tx\| : x \in K, \|Ix\| \leq \epsilon\} \leq E(K, \epsilon). \quad (2.10.8)$$

Proof. For every $x \in K$ such that $\|Ix\| \leq \epsilon$ and any algorithm A , we have

$$\|Tx - A\theta\| \leq E_A(K, \epsilon),$$

as well as

$$\|T(-x) - A\theta\| = \|Tx + A\theta\| \leq E_A(K, \epsilon),$$

whence $\|Tx\| \leq E_A(K, \epsilon)$. This implies $e(K, \epsilon) \leq E_A(K, \epsilon)$, which yields the result.

The next result, due to Micchelli and Rivlin (1977), gives an upper bound for $E(K, \epsilon)$. We omit its proof.

Theorem 2.20. *We have*

$$E(K, \epsilon) \leq 2e(K, \epsilon). \quad (2.10.9)$$

It is important to observe that optimal recovery with error is, at least theoretically, equivalent to recovery with exact information. To see this, let $\hat{X} = X \times Y$ and in \hat{X} let \hat{K} denote the balanced convex set $K \times U(Y)$. We extend T and I to \hat{X} by defining $\hat{T}(x, y) = Tx$, $\hat{I}(x, y) = Ix + \epsilon y$, respectively. Schematically, (Figure 2.6).

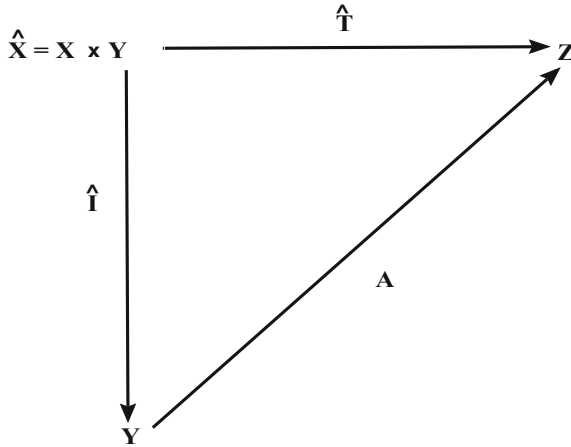


Fig. 2.6 Another schematic representation

Next, we observe that

$$\begin{aligned} E(\hat{K}) &= \inf_A \sup_{(x,y) \in \hat{K}} \|\hat{T}(x, y) - A(\hat{I}(x, y))\| \\ &= \inf_A \sup \{\|Tx - A(Ix + \epsilon y)\| : x \in K, y \in U(Y)\} \end{aligned}$$

$$\begin{aligned}
&= \inf_A \sup \{ \|Tx - Ay\| : x \in K, \|Ix - y\| \leq \epsilon \} \\
&= E(K, \epsilon).
\end{aligned}$$

This justifies the observation made before.

In many situations, it becomes possible to bridge the gap between Theorems 2.19 and 2.20 and thereby solve the optimal recovery problem. Specifically, let

$$K_0 := \{x \in K : Ix = \theta\}, \quad (2.10.10)$$

and let $e(K)$ denote $e(K, 0)$. Then, we have

Theorem 2.21. (Morozov and Grebennikov) *Suppose there exists a transformation $S : IX \rightarrow X$, such that $x - SIx \in K_0$ for all $x \in K$. Then, $e(K) = E(K)$ and $\hat{A} = TS$ is an optimal algorithm.*

Proof: Indeed,

$$\begin{aligned}
E_{\hat{A}}(K) &= \sup_{x \in K} \|Tx - TS(Ix)\| \\
&= \sup_{x \in K} \|T(x - SIx)\| \\
&\leq \sup_{x \in K_0} \|Tx\| = e(K) \leq E(K).
\end{aligned}$$

The above proposition enables us to obtain a variant of Example 2.17 above.

Example 2.19. Let

$$X := W_{\infty}^n[0, 1] := \{x : [0, 1] \longrightarrow \mathbb{R} : x^{(n-1)} \in AC[0, 1] \text{ and } x^{(n)} \in L_{\infty}[0, 1]\}$$

and $K := \{x \in X : \|x^{(n)}\|_{\infty} \leq 1\}$. Let $I : X \rightarrow \mathbb{R}^n$ be defined by $I(x) = (x(t_1), \dots, x(t_n))$ where t_1, t_2, \dots, t_n in $[0, 1]$ are distinct points. Define $S : IX \rightarrow X$ by $S(Ix) = L_{t_1, t_2, \dots, t_n}(x)$, the unique polynomial in $P_n(\mathbb{R})$ interpolating x at the points t_i 's. We have

$$(x - S(Ix))^{(n)} = x^{(n)}, \|x - S(Ix)\|_{\infty}^{(n)} \leq 1 \text{ for all } x \in K.$$

Therefore, $x - S(Ix) \in K_0$ for all $x \in K$. Theorem 2.21 now reveals that $e(K) = E(K)$ and TS is an optimal algorithm. The problem of optimal interpolation which corresponds to $Z = \mathbb{R}$, $Y = \mathbb{R}^{n+r}$, $Tx = x(t_0)$ for some point $t_0 \in [0, 1] \setminus \{t_1, \dots, t_{n+r}\}$, in case the number of sampling nodes $> n$ was treated by Micchelli, Rivlin, and Winograd [6] by making use of perfect spline for bounding the intrinsic error $E(K)$ from below.

Example 2.20. Let $X = W_2^n[0, 1] = Z$, $Y = \mathbb{R}^{n+r}$, $r \geq 0$. Let $\Delta := \{t_i\}_1^{n+r}$ be a strictly increasing sequence of data nodes in $(0, 1)$. For $x \in X$, let

$$Ix = x|_{\Delta} = (x(t_1), \dots, x(t_{n+r})),$$

$$K := \{x \in X : \|x^{(n)}\|_2 \leq 1\}.$$

Let us denote by $\hat{S}_{2n}(\Delta)$ the space of natural splines of order $2n$ with the knot sequence Δ . For $x \in X$, let $s(x)$ denote the unique element of $\hat{S}_{2n}(\Delta)$ interpolating x on $\Delta : x|_{\Delta} = s|_{\Delta}$ (cf. Theorem 6.3.3.3 in Mhaskar and Pai [3]). It follows from Theorem 6.3.3.6 of Mhaskar and Pai [3] that

$$\|x^{(n)} - (s(x))^{(n)}\|_2 \leq \|x^{(n)}\|_2, \text{ for all } x \in X.$$

Thus if we define $S : IX \rightarrow X$ by $S(Ix) = s(x)$, then $I(x - s(x)) = 0$ and $x - SIx \in K$ for all $x \in K$. Theorem 2.21 now reveals that natural spline interpolant is an optimal algorithm.

2.10.3 Central Algorithms

As remarked in the introduction, the problem of optimal recovery is closely related to the notion of Chebyshev center of a set. The reader may recall the notations in the introduction of Section 2.8. In particular, recall that for a bounded subset F of a normed linear space X , $\text{rad}(F) := \inf_{x \in X} r(F; x)$ denotes the Chebyshev radius of F , and

$$\text{Cent}(F) := \{x \in X : r(F; x) = \text{rad}(F)\}$$

denotes the Chebyshev center of F (possibly void). The next two elementary results relate the Chebyshev radius of a bounded set F with its diameter denoted by $\text{diam}(F)$.

Theorem 2.22. *If F is a bounded subset of a normed linear space X , then*

$$\frac{1}{2} \text{diam}(F) \leq \text{rad}(F) \leq \text{diam}(F). \quad (2.10.11)$$

Proof: For $y_1, y_2 \in F$ and $x \in X$ we have

$$\|y_1 - y_2\| \leq \|y_1 - x\| + \|x - y_2\| \leq 2r(F; x).$$

Therefore, $\text{diam}(F) \leq 2r(F; x)$ which entails $\frac{1}{2} \text{diam}(F) \leq \text{rad}(F)$. On the other hand, $\|y_1 - y_2\| \leq \text{diam}(F) \Rightarrow r(F; y_1) \leq \text{diam}(F)$. Hence, $\text{rad}(F) \leq \text{diam}(F)$.

Definition 2.18. A subset F of a linear space X is said to be *symmetric about an element* $x_0 \in X$ if

$$z \in X, \quad x_0 + z \in F \Rightarrow x_0 - z \in F.$$

Theorem 2.23. *If F is a bounded subset of a normed linear space X which is symmetric about an element $x_0 \in X$, then $x_0 \in \text{Cent}(F)$ and*

$$\text{rad}(F) = \frac{1}{2} \text{diam}(F). \quad (2.10.12)$$

Proof: If x_0 were not to belong to $\text{Cent}(F)$, then there would exist $x \in F$ such that $r(F; x) < r(F; x_0)$ and one could pick $y_0 \in F$ such that $r(F; x) < \|x_0 - y_0\|$. Then, $\|x - y\| < \|x_0 - y_0\|$ for all $y \in F$. Let $x_0 - y_0 = z_0$. Then, $x_0 - z_0 \in F$ and since F is symmetric about x_0 , $x_0 + z_0 \in F$. Therefore,

$$\begin{aligned} 2\|z_0\| &= \|(x_0 + z_0) - x - (x_0 - z_0) + x\| \\ &\leq \|(x_0 + z_0) - x\| + \|(x_0 - z_0) - x\| \\ &< 2\|x_0 - y_0\| = 2\|z_0\|, \text{ a contradiction.} \end{aligned}$$

Thus $x_0 \in \text{Cent}(F)$. Let $x \in X$ and $\delta > 0$. Pick $y_1 \in F$ such that $\|x - y_1\| > r(F; x) - \delta$ and put $x - y_1 = z$. Then, $y_1 = x - z \in F$ and symmetry of F about x_0 shows that $y_2 = x + z \in F$. Therefore,

$$\|y_1 - y_2\| = 2\|z\| > 2r(F; x) - 2\delta \geq 2\text{rad}(F) - 2\delta,$$

which yields $\text{diam}(F) \geq 2\text{rad}(F) - 2\delta$. In view of (2.10.11), we obtain

$$\text{rad}(F) = \frac{1}{2} \text{diam}(F).$$

The preceding theorem motivates the next definition.

Definition 2.19. A bounded set F is said to be *centerable* if (2.10.12) holds.

For $y \in IK$, let us denote by K_y the set

$$K_y := \{x \in K : Ix = y\} = I^{-1}\{y\} \cap K,$$

and let the “hypercircle” $H(y)$ be defined by

$$H(y) = T(K_y) = \{Tx : x \in K, Ix = y\}.$$

We assume that $H(y)$ is *bounded* for each $y \in IK$. Let us call

$$r_I(T, K; y) := \text{rad}H(y)$$

the *local radius of information* and

$$r_I(T, K) = \sup_{y \in IK} r_I(T, K; y)$$

the (*global*) *radius of information*. We have

Theorem 2.24. *Under the hypothesis as before,*

$$e(K) = \text{rad}H(\theta) = r_I(T, K; \theta), \quad (2.10.13)$$

and

$$E(K) = \sup_{y \in IK} \text{rad}H(y) = r_I(T, K). \quad (2.10.14)$$

Proof: We have

$$\begin{aligned} e(K) &= e(K, 0) = \sup\{\|Tx\| : x \in K, Ix = \theta\} \\ &= \sup\{\|z\| : z \in H(\theta)\} = r(H(\theta), \theta). \end{aligned}$$

Observe that $H(\theta)$ is symmetric about θ . Indeed, if $z \in H(\theta)$, then $z = Tx$ for some $x \in K$ such that $Ix = \theta$. This implies $-z = T(-x)$, and since $-x \in K$ and $I(-x) = \theta$, we have $-z \in H(\theta)$. Thus $\theta \in \text{Cent}(H(\theta))$ and $r(H(\theta), \theta) = \text{rad}H(\theta)$, which proves (2.10.13).

To prove (2.10.14), observe that since for any algorithm A , we have

$$\begin{aligned} \|Tx - AIx\| &\leq E_A(K) \text{ for all } x \in K, \\ \sup_{z \in H(y)} \|z - Ay\| &\leq E_A(K), \end{aligned}$$

which implies

$$\text{rad}H(y) \leq r(H(y), Ay) \leq E_A(K) \text{ for all } y \in IK.$$

Thus $\sup_{y \in IK} \text{rad}(H(y)) \leq E(K)$. To reverse this inequality $\epsilon > 0$ given, for each $y \in IK$ pick $z_y \in Z$ such that $r(H(y), z_y) < \text{rad}H(y) + \epsilon$. Then, for each $x \in K$ and $y \in Y$ such that $Ix = y$, we have

$$\|Tx - z_y\| < \text{rad}H(y) + \epsilon \leq \sup_{y \in IK} \text{rad}(H(y)) + \epsilon.$$

Now, consider the algorithm $\tilde{A} : IK \rightarrow Z$ defined by $\tilde{A}(y) = z_y$, $y \in IK$. We have

$$\begin{aligned}
E_{\tilde{A}}(K) &= \sup\{\|Tx - \tilde{A}y\| : x \in K, Ix = y\} \\
&= \sup\{\|Tx - z_y\| : x \in K, Ix = y\} \\
&\leq \sup_{y \in IK} \text{rad}(H(y)) + \epsilon.
\end{aligned}$$

Thus $E(K) \leq \sup_{y \in IK} \text{rad}(H(y))$, which establishes (2.10.14).

Corollary 2.4. *Under the hypothesis as before, suppose Z admits centers for the sets $H(y)$, $y \in IK$. Pick $c_y \in \text{Cent}H(y)$ for each $y \in IK$, then $y \rightarrow c_y$ is an optimal algorithm.*

Lastly, aside from Theorem 2.24, we present one more case wherein $e(K) = E(K)$.

Theorem 2.25. *Assume $T(K)$ is bounded and that each set $H(y)$, $y \in IK$ is centerable, then $e(K) = E(K)$.*

Proof: For each $y \in IK$, we have $\text{diam}H(y) = 2\text{rad}H(y)$, and

$$\begin{aligned}
\text{diam}H(y) &= \sup\{\|Tx_1 - Tx_2\| : x_1, x_2 \in K \text{ and } Ix_1 = Ix_2 = y\} \\
&= \sup\{\|T(x_1 - x_2)\| : x_1, x_2 \in K \text{ and } Ix_1 = Ix_2 = y\} \\
&\leq \sup\{\|Tx\| : x \in 2K, Ix = \theta\} \\
&= 2 \sup\{\|Tx\| : x \in K, Ix = \theta\} \\
&= 2e(K).
\end{aligned}$$

Thus $\text{rad}H(y) \leq e(K)$ for all $y \in IK$ and we conclude from (2.10.14) that $E(K) \leq e(K)$.

2.10.4 Notes

One of the first landmark articles devoted to optimal recovery is Golomb and Weinberger [1] and one of the first surveys in this area is Section 33 in Holmes [2]. A more extensive survey of the literature in this direction up to 1977 is contained in Micchelli and Rivlin [4]. An update of this survey has also been contributed by the same authors in 1985. This topic is now encompassed by a wider topic called *information-based complexity*. The interested reader is referred to the book of Traub and Wozniakowski [7] for more details in this direction.

Exercises IV

2.41. (Rivlin and Micchelli): For optimal approximation of non-expansive functions with $m = 1$ in Example II, let $T = [0, 1]$, $0 \leq t_1 < t_2 \dots < t_n \leq 1$, $\xi_i := \frac{1}{2}(t_i +$

$t_{i+1}), i = 1, 2, \dots, n-1, \xi_0 := 0, \xi_n := 1$. Then, $T_1 = [0, \xi_1], T_n = [\xi_{n-1}, 1]$, and $T_i = [\xi_{i-1}, \xi_i], i = 2, \dots, n-1$. Show that with $\Delta := \max_{1 \leq i \leq n-1} \Delta_i, \Delta_i := t_{i+1} - t_i, E(\mathcal{K}) = \|\hat{q}\|_\infty = \max\{t_1, \frac{1}{2}\Delta, 1 - t_n\}$ and that an optimal approximant of $x \in \mathcal{K}$ is given by the step function

$$s(t) = x(t_i), \xi_{i-1} < t < \xi_i, i = 1, \dots, n.$$

Show also that the optimal sampling nodes in this case are given by $t_i = \frac{2i-1}{2n}, i = 1, \dots, n$ with

$$E(\mathcal{K}) = \frac{1}{2n} = \min\{E(\mathcal{K}) : t_1, \dots, t_n \in [0, 1], \text{ distinct}\}.$$

Furthermore, show that for the optimal approximation problem as above another simple optimal algorithm is

$$\mathring{A} : (x(t_1), \dots, x(t_n)) \longrightarrow p(t)$$

where $p(t)$ is the piecewise linear interpolant of x at the nodes t_i s.

2.42. (Rivlin and Micchelli): For the optimal approximation of non-expansive functions with $m = 2, n = 3$ in Example II, take T to be a triangle with vertices $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ (not obtuse angled). Show that another optimal approximant of $x \in \mathcal{K}$ is given by $\mathring{A} = (x(\mathbf{t}_1), x(\mathbf{t}_2), x(\mathbf{t}_3)) \longrightarrow \ell$ where ℓ is the linear interpolant to x ,

$$\ell(\mathbf{t}) = au + bv + c, \mathbf{t} = (u, v)$$

satisfying $\ell(\mathbf{t}_i) = x(\mathbf{t}_i), i = 1, 2, 3$.

References

1. Golomb, M., & Weinberger, H. (1959). Optimal approximation and error bounds. In R. Langer (Ed.), *On numerical approximation* (pp. 117–190). Madison: University of Wisconsin Press.
2. Holmes, R. B. (1972). *A course on optimization and best approximation* (Vol. 257). Lecture notes. New York: Springer.
3. Mhaskar, H. N., & Pai, D. V. (2007). *Fundamentals of approximation theory*. New Delhi: Narosa Publishing House.
4. Micchelli, C. A., & Rivlin, T. J. (1977). A survey of optimal recovery. In C. A. Micchelli & T. J. Rivlin (Eds.), *Optimal estimation in approximation theory* (pp. 1–54). New York: Plenum Press.
5. Micchelli, C. A., & Rivlin, T. J. (1985). *Lectures on optimal recovery* (Vol. 1129, pp. 21–93). Lecture notes in mathematics. Berlin: Springer.
6. Micchelli, C. A., Rivlin, T. J., & Winograd, S. (1976). Optimal recovery of smooth functions. *Numerische Mathematik*, 260, 191–200.
7. Traub, J. F., & Wozniakowski, H. (1980). *A general theory of optimal algorithms*. New York: Academic Press.

Fractional and Multivariable Calculus
Model Building and Optimization Problems

Mathai, A.M.; Haubold, H.

2017, XIII, 234 p. 7 illus., Hardcover

ISBN: 978-3-319-59992-2