Chapter II

METRIC SPACES

§8. Definition and examples of metric spaces

Passage to the limit is one of the most important operations in analysis. The basis of this operation is the fact that the distance between any two points on the real line is defined. A number of fundamental facts from analysis are not connected with the algebraic nature of the set of real numbers (i.e. with the fact that the operations of addition and multiplication, which are subject to known laws, are defined for real numbers), but depend only on those properties of real numbers which are related to the concept of distance. This situation leads naturally to the concept of "metric space" which plays a fundamental role in modern mathematics. Further on we shall discuss the basic facts of the theory of metric spaces. The results of this chapter will play an essential role in all the following discussion.

DEFINITION. A *metric space* is the pair of two things: a set X, whose elements are called points, and a distance, i.e. a single-valued, nonnegative, real function $\rho(x, y)$, defined for arbitrary x and y in X and satisfying the following conditions:

1) $\rho(x, y) = 0$ if and only if $x = y$,
2) (axiom of symmetry) $\rho(x, y) = \rho(y, x)$,
3) (triangle axiom) $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$.

The metric space itself, i.e. the pair X and ρ, will usually be denoted by $R = (X, \rho)$.

In cases where no misunderstanding can arise we shall sometimes denote the metric space by the same symbol X which is used for the set of points itself.

We list a number of examples of metric spaces. Some of the spaces listed below play a very important role in analysis.

1. If we set

$$\rho(x, y) = \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{if } x \neq y, \end{cases}$$

for elements of an arbitrary set, we obviously obtain a metric space.

2. The set D^1 of real numbers with the distance function

$$\rho(x, y) = |x - y|$$

forms the metric space R^1.

3. The set D^n of ordered n-tuples of real numbers $x = (x_1, x_2, \cdots, x_n)$ with distance function

$$\rho(x, y) = \{\sum_{k=1}^{n} (y_k - x_k)^2\}^{\frac{1}{2}}$$

is called Euclidean n-space R^n. The validity of Axioms 1 and 2 for R^n is obvious. To prove that the triangle axiom is also verified in R^n we make use of the Schwarz inequality

$$(1) \qquad (\sum_{k=1}^{n} a_k b_k)^2 \leq \sum_{k=1}^{n} a_k^2 \sum_{k=1}^{n} b_k^2.$$

(The Schwarz inequality follows from the identity

$$(\sum_{k=1}^{n} a_k b_k)^2 = (\sum_{k=1}^{n} a_k^2)(\sum_{k=1}^{n} b_k^2) - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} (a_i b_j - b_i a_j)^2,$$

which can be verified directly.) If

$$x = (x_1, x_2, \cdots, x_n), \quad y = (y_1, y_2, \cdots, y_n) \quad \text{and} \quad z = (z_1, z_2, \cdots, z_n),$$

then setting

$$y_k - x_k = a_k, \qquad z_k - y_k = b_k,$$

we obtain

$$z_k - x_k = a_k + b_k;$$

by the Schwarz inequality

$$\begin{aligned} \sum_{k=1}^{n} (a_k + b_k)^2 &= \sum_{k=1}^{n} a_k^2 + 2\sum_{k=1}^{n} a_k b_k + \sum_{k=1}^{n} b_k^2 \\ &\leq \sum_{k=1}^{n} a_k^2 + 2\{\sum_{k=1}^{n} a_k^2 \sum_{k=1}^{n} b_k^2\}^{\frac{1}{2}} + \sum_{k=1}^{n} b_k^2 \\ &= [(\sum_{k=1}^{n} a_k^2)^{\frac{1}{2}} + (\sum_{k=1}^{n} b_k^2)^{\frac{1}{2}}]^2, \end{aligned}$$

i.e.

$$\rho^2(x, z) \leq \{\rho(x, y) + \rho(y, z)\}^2$$

or

$$\rho(x, z) \leq \rho(x, y) + \rho(y, z).$$

4. Consider the space R_0^n in which the points are again ordered n-tuples of numbers $(x_1, x_2, \cdots, x_n)$, and for which the distance function is defined by the formula

$$\rho_0(x, y) = \max\{|y_k - x_k|;\ 1 \leq k \leq n\}.$$

The validity of Axioms 1–3 is obvious. In many questions of analysis this space is no less suitable than Euclidean space R^n.

Examples 3 and 4 show that sometimes it is actually important to have different notations for the set of points of a metric space and for the metric space itself because the same point set can be metrized in various ways.

5. The set $C[a, b]$ of all continuous real-valued functions defined on the segment $[a, b]$ with distance function

$$(2) \qquad \rho(f, g) = \sup\{|g(t) - f(t)|;\ a \leq t \leq b\}$$

likewise forms a metric space. Axioms 1–3 can be verified directly. This space plays a very important role in analysis. We shall denote it by the same symbol $C[a, b]$ as the set of points of this space. The space of continuous functions defined on the segment [0, 1] with the metric given above will be denoted simply by C.

6. We denote by l_2 the metric space in which the points are all possible sequences $x = (x_1, x_2, \cdots, x_n, \cdots)$ of real numbers which satisfy the condition $\sum_{k=1}^{\infty} x_k^2 < \infty$ and for which the distance is defined by means of the formula

$$\rho(x, y) = \{\sum_{k=1}^{\infty} (y_k - x_k)^2\}^{\frac{1}{2}}. \tag{3}$$

We shall first prove that the function $\rho(x, y)$ defined in this way always has meaning, i.e. that the series $\sum_{k=1}^{\infty} (y_k - x_k)^2$ converges. We have

$$\{\sum_{k=1}^{n} (y_k - x_k)^2\}^{\frac{1}{2}} \leq (\sum_{k=1}^{n} x_k^2)^{\frac{1}{2}} + (\sum_{k=1}^{n} y_k^2)^{\frac{1}{2}} \tag{4$_n$}$$

for arbitrary natural number n (see Example 3).

Now let $n \to \infty$. By hypothesis, the right member of this inequality has a limit. Thus, the expression on the left is bounded and does not decrease as $n \to \infty$; consequently, it tends to a limit, i.e. formula (3) has meaning. Replacing x by $-x$ in (4$_n$) and passing to the limit as $n \to \infty$, we obtain

$$\{\sum_{k=1}^{\infty} (y_k + x_k)^2\}^{\frac{1}{2}} \leq (\sum_{k=1}^{\infty} x_k^2)^{\frac{1}{2}} + (\sum_{k=1}^{\infty} y_k^2)^{\frac{1}{2}}; \tag{4}$$

but this is essentially the triangle axiom. In fact, let

$$a = (a_1, a_2, \cdots, a_n, \cdots),$$

$$b = (b_1, b_2, \cdots, b_n, \cdots),$$

$$c = (c_1, c_2, \cdots, c_n, \cdots)$$

be three points in l_2. If we set

$$b_k - a_k = x_k, \qquad c_k - b_k = y_k,$$

then

$$c_k - a_k = y_k + x_k$$

and, by virtue of (4),

$$\{\sum_{k=1}^{\infty} (c_k - a_k)^2\}^{\frac{1}{2}} \leq \{\sum_{k=1}^{\infty} (b_k - a_k)^2\}^{\frac{1}{2}} + \{\sum_{k=1}^{\infty} (c_k - b_k)^2\}^{\frac{1}{2}},$$

i.e.

$$\rho(a, c) \leq \rho(a, b) + \rho(b, c).$$

7. Consider, as in Example 5, the totality of all continuous functions on the segment $[a, b]$, but now let the distance be defined by setting

$$\rho(x, y) = \left[\int_a^b \{x(t) - y(t)\}^2 \, dt\right]^{\frac{1}{2}}. \tag{5}$$

This metric space is denoted by $C^2[a, b]$ and is called the *space of continuous functions with quadratic metric*. Here again Axioms 1 and 2 in the definition of a metric space are obvious and the triangle axiom follows immediately from the Schwarz inequality

$$\left\{\int_a^b x(t)y(t) \, dt\right\}^2 \leq \int_a^b x^2(t) \, dt \int_a^b y^2(t) \, dt,$$

which can be obtained, for instance, from the following easily-verified identity:

$$\left\{\int_a^b x(t)y(t) \, dt\right\}^2 = \int_a^b x^2(t) \, dt \int_a^b y^2(t) \, dt - \frac{1}{2}\int_a^b \int_a^b [x(s)y(t) - y(s)x(t)]^2 \, ds \, dt.$$

8. Consider the set of all bounded sequences $x = (x_1, x_2, \cdots, x_n, \cdots)$ of real numbers. We obtain the metric space M^∞ if we set

$$\rho(x, y) = \sup |y_k - x_k|. \tag{6}$$

The validity of Axioms 1–3 is obvious.

9. The following principle enables us to write down an infinite number of further examples: if $R = (X, \rho)$ is a metric space and M is an arbitrary subset of X, then M with the same function $\rho(x, y)$, but now assumed to be defined only for x and y in M, likewise forms a metric space; it is called a subspace of the space R.

(1) In the definition of a metric space we could have limited ourselves to two axioms for $\rho(x, y)$, namely:

1) $$\rho(x, y) = 0$$

if, and only if, $x = y$;

2) $$\rho(x, y) \leq \rho(z, x) + \rho(z, y)$$

for arbitrary x, y, z.

It follows that

3) $$\rho(x, y) \geq 0,$$

4) $$\rho(x, y) = \rho(y, x)$$

and consequently Axiom 2 can be written in the form

2′) $$\rho(x, y) \leq \rho(x, z) + \rho(z, y).$$

(2) The set D^n of ordered n-tuples of real numbers with distance

$$\rho_p(x, y) = (\sum_{k=1}^{n} |y_k - x_k|^p)^{1/p} \qquad (p \geq 1)$$

also forms a metric space which we shall denote by R_p^n. Here the validity of Axioms 1 and 2 is again obvious. We shall check Axiom 3. Let

$$x = (x_1, x_2, \cdots, x_n), \qquad y = (y_1, y_2, \cdots, y_n) \quad \text{and} \quad z = (z_1, z_2, \cdots, z_n)$$

be points in R_p^n. If, as in Example 3, we set

$$y_k - x_k = a_k, \qquad z_k - y_k = b_k,$$

then the inequality

$$\rho_p(x, z) \leq \rho_p(x, y) + \rho_p(y, z)$$

assumes the form

$$(7) \qquad (\sum_{k=1}^n | a_k + b_k |^p)^{1/p} \leq (\sum_{k=1}^n | a_k |^p)^{1/p} + (\sum_{k=1}^n | b_k |^p)^{1/p}.$$

This is the so-called Minkowski inequality. Minkowski's inequality is obvious for $p = 1$ (since the absolute value of a sum is less than or equal to the sum of the absolute values) and therefore we can restrict ourselves to considering the case $p > 1$.

In order to prove inequality (7) for $p > 1$ we shall first establish Hölder's inequality:

$$(8) \qquad \sum_{k=1}^n | x_k y_k | \leq (\sum_{k=1}^n | x_k |^p)^{1/p} (\sum_{k=1}^n | y_k |^q)^{1/q},$$

where the number q is defined by the condition

$$(9) \qquad 1/p + 1/q = 1.$$

We note that inequality (8) is homogeneous in the sense that if it is satisfied for any two vectors

$$x = (x_1, x_2, \cdots, x_n) \quad \text{and} \quad y = (y_1, y_2, \cdots, y_n),$$

then it is also satisfied for the vectors λx and μy where λ and μ are arbitrary numbers. Therefore it is sufficient to prove inequality (8) for the case when

$$(10) \qquad \sum_{k=1}^n | x_k |^p = \sum_{k=1}^n | y_k |^q = 1.$$

Thus, we must prove that if Condition (10) is satisfied, then

$$(11) \qquad \sum_{k=1}^n | x_k y_k | \leq 1.$$

Consider in the (ξ, η)-plane the curve defined by the equation $\eta = \xi^{p-1}$, or equivalently by the equation $\xi = \eta^{q-1}$ (see Fig. 7). It is clear from the figure that for an arbitrary choice of positive values for a and b we have $S_1 + S_2 \geq ab$. If we calculate the areas S_1 and S_2, we obtain

$$S_1 = \int_0^a \xi^{p-1}\, d\xi = a^p/p; \qquad S_2 = \int_0^b \eta^{q-1}\, d\eta = b^q/q.$$

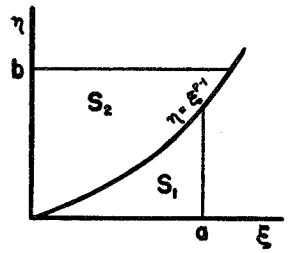

FIG. 7

Thus

$$ab \leq a^p/p + b^q/q.$$

Setting $a = | x_k |$, $b = | y_k |$ and summing with respect to k from 1 to n, we obtain

$$\sum_{k=1}^n | x_k y_k | \leq 1,$$

if we take (9) and (10) into consideration.

Inequality (11) and consequently the more general inequality (8) are thus proved. For $p = 2$ Hölder's inequality (8) becomes the Schwarz inequality (1).

We now proceed to the proof of the Minkowski inequality. Consider the identity

$$(| a | + | b |)^p = (| a | + | b |)^{p-1} | a | + (| a | + | b |)^{p-1} | b |.$$

Setting $a = x_k$, $b = y_k$ in the above identity and summing with respect to k from 1 to n, we obtain

$$\sum_{k=1}^n (| x_k | + | y_k |)^p = \sum_{k=1}^n (| x_k | + | y_k |)^{p-1} | x_k | + \sum_{k=1}^n (| x_k | + | y_k |)^{p-1} | y_k |.$$

If we now apply Hölder's inequality to each of the two sums on the right of the above equality and take into consideration the fact that $(p - 1)q = p$, we obtain

$$\sum_{k=1}^n (| x_k | + | y_k |)^p \leq \{\sum_{k=1}^n (| x_k | + | y_k |)^p\}^{1/q} \{(\sum_{k=1}^n | x_k |^p)^{1/p} + (\sum_{k=1}^n | y_k |^p)^{1/p}\}.$$

Dividing both sides of this inequality by

$$\{\sum_{k=1}^n (| x_k | + | y_k |)^p\}^{1/q},$$

we obtain

$$\{\sum_{k=1}^n (| x_k | + | y_k |)^p\}^{1/p} \leq (\sum_{k=1}^n | x_k |^p)^{1/p} + (\sum_{k=1}^n | y_k |^p)^{1/p},$$

whence inequality (7) follows immediately. This also establishes the triangle axiom for the space R_p^n.

(3) It is possible to show that the metric

$$\rho_0(x, y) = \max \{| y_k - x_k |; 1 \leq k \leq n\}$$

introduced in Example 4 can be defined in the following way:

$$\rho_0(x, y) = \lim_{p\to\infty} \left(\sum_{k=1}^n | y_k - x_k |^p\right)^{1/p}.$$

(4) From the inequality

$$ab \leq a^p/p + b^q/q \qquad (1/p + 1/q = 1)$$

established in Example (2) it is easy to deduce also the integral form of Hölder's inequality

$$\int_a^b x(t)y(t)\, dt \leq \left(\int_a^b | x^p(t) |\, dt\right)^{1/p} \left(\int_a^b | y^q(t) |\, dt\right)^{1/q},$$

which is valid for arbitrary functions $x(t)$ and $y(t)$ for which the integrals on the right have meaning. From this in turn we obtain the integral form of Minkowski's inequality:

$$\left(\int_a^b | x(t) + y(t) |^p\, dt\right)^{1/p} \leq \left(\int_a^b | x(t) |^p\, dt\right)^{1/p} + \left(\int_a^b | y(t) |^p\, dt\right)^{1/p}.$$

(5) We shall point out still another interesting example of a metric space. Its elements are all possible sequences of real numbers

$$x = (x_1, x_2, \cdots, x_n, \cdots)$$

such that $\sum_{k=1}^\infty | x_k |^p < \infty$, where $p \geq 1$ is any fixed number and the distance is defined by means of the formula

$$\rho(x, y) = \left(\sum_{k=1}^\infty | y_k - x_k |^p\right)^{1/p}. \tag{12}$$

We shall denote this metric space by l_p.

By virtue of Minkowski's inequality (7) we have

$$\left(\sum_{k=1}^n | y_k - x_k |^p\right)^{1/p} \leq \left(\sum_{k=1}^n |x_k |^p\right)^{1/p} + \left(\sum_{k=1}^n | y_k |^p\right)^{1/p}$$

for arbitrary n. Since the series

$$\sum_{k=1}^\infty | x_k |^p \quad \text{and} \quad \sum_{k=1}^\infty | y_k |^p$$

converge by assumption, passing to the limit as $n \to \infty$ we obtain

$$\left(\sum_{k=1}^\infty | y_k - x_k |^p\right)^{1/p} \leq \left(\sum_{k=1}^\infty | x_k |^p\right)^{1/p} + \left(\sum_{k=1}^\infty | y_k |^p\right)^{1/p}, \tag{13}$$

and so the series on the left side also converges. This proves that formula (12), which defines distance in l_p, actually has meaning for arbitrary $x, y \in l_p$. At the same time inequality (13) shows that the triangle axiom is satisfied in l_p. The remaining axioms are obvious.

§9. Convergence of sequences. Limit points

In §§9–11 we shall establish some fundamental concepts which we shall frequently use in the sequel.

An *open sphere* $S(x_0, r)$ in the metric space R is the set of all points $x \in R$ which satisfy the condition $\rho(x, x_0) < r$. The fixed point x_0 is called the *center* and the number r is called the *radius* of this sphere.

A *closed sphere* $S[x_0, r]$ is the set of all points $x \in R$ which satisfy the condition $\rho(x, x_0) \leq r$.

An ϵ-neighborhood of the point x, denoted by the symbol $O(x, \epsilon)$, is an open sphere of radius ϵ and center x_0.

A point x is called a *contact point* of the set M if every neighborhood of x contains at least one point of M. The set of all contact points of the set M is denoted by $[M]$ and is called the closure of M. Since every point belonging to M is obviously a contact point of M (each point is contained in every one of its neighborhoods), every set is contained in its closure: $M \subseteq [M]$.

THEOREM 1. *The closure of the closure of M is equal to the closure of M:*

$$[[M]] = [M].$$

Proof. Let $x \in [[M]]$. Then an arbitrary ϵ-neighborhood $O(x, \epsilon)$ of x contains a point $x_1 \in [M]$. Setting $\epsilon - \rho(x, x_1) = \epsilon_1$, we consider the sphere $O(x_1, \epsilon_1)$. This sphere lies entirely in the interior of $O(x, \epsilon)$. In fact, if $z \in O(x_1, \epsilon_1)$, then $\rho(z, x_1) < \epsilon_1$; and since $\rho(x, x_1) = \epsilon - \epsilon_1$, then, by the triangle axiom $\rho(z, x) \leq \epsilon_1 + (\epsilon - \epsilon_1) = \epsilon$, i.e. $z \in O(x, \epsilon)$. Since $x_1 \in [M]$, $O(x, \epsilon_1)$ contains a point $x_2 \in M$. But then $x_2 \in O(x, \epsilon)$. Since $O(x, \epsilon)$ is an arbitrary neighborhood of the point x, we have $x \in [M]$. This completes the proof of the theorem.

The validity of the following assertion is obvious.

THEOREM 2. *If $M_1 \subseteq M$, then $[M_1] \subseteq [M]$.*

THEOREM 3. *The closure of a sum is equal to the sum of the closures:*

$$[M_1 \cup M_2] = [M_1] \cup [M_2].$$

Proof. Let $x \in [M_1 \cup M_2]$, i.e. let an arbitrary neighborhood $O(x, \epsilon)$ contain the point $y \in M_1 \cup M_2$. If it were true that $x \notin [M_1]$ and $x \notin [M_2]$, we could find a neighborhood $O(x, \epsilon_1)$ which does not contain points of M_1 and a neighborhood $O(x, \epsilon_2)$ which does not contain points of M_2. But then the neighborhood $O(x, \epsilon)$, where $\epsilon = \min(\epsilon_1, \epsilon_2)$, would not contain points of $M_1 \cup M_2$. From the contradiction thus obtained it follows that x is contained in at least one of the sets $[M_1]$ and $[M_2]$, i.e.

$$[M_1 \cup M_2] \subseteq [M_1] \cup [M_2].$$

Since $M_1 \subseteq M_1 \cup M_2$ and $M_2 \subseteq M_1 \cup M_2$, the converse inclusion follows from Theorem 2.

The point x is called a *limit point* of the set M if an arbitrary neighborhood of x contains an infinite number of points of M.

A limit point of the set M can either belong to M or not. For example, if M is the set of rational numbers in the closed interval [0, 1], then every point of this interval is a limit point of M.

A point x belonging to the set M is said to be an *isolated point* of this set if x has a neighborhood $O(x, \epsilon)$ which does not contain any points of M different from x.

THEOREM 4. *Every contact point of the set M is either a limit point of the set M or an isolated point of M.*

Proof. Let x be a contact point of the set M. This means that every neighborhood $O(x, \epsilon)$ of x contains at least one point belonging to M. Two cases are possible:

1) Every neighborhood of the point x contains an infinite number of points of the set M. In this case, x is a limit point of M.

2) We can find a neighborhood $O(x, \epsilon)$ of x which contains only a finite number of points of M. In this case, x will be an isolated point of the set M. In fact, let $x_1, x_2, \cdots, x_k$ be the points of M which are distinct from x and which are contained in the neighborhood $O(x, \epsilon)$. Further, let ϵ_0 be the least of the positive numbers $\rho(x, x_i)$, $i = 1, 2, \cdots, k$. Then the neighborhood $O(x, \epsilon_0)$ obviously does not contain any point of M distinct from x. The point x itself in this case must necessarily belong to M since otherwise $O(x, \epsilon_0)$ in general would not contain a single point of M, i.e. x would not be a contact point of the set M. This completes the proof of the theorem.

Thus, the set $[M]$ consists in general of points of three types:

1) Isolated points of the set M;

2) Limit points of the set M which belong to M;

3) Limit points of the set M which do not belong to M.

$[M]$ is obtained by adding to M all its limit points.

Let $x_1, x_2, \cdots$ be a sequence of points in the metric space R. We say that this sequence *converges to the point* x if every neighborhood $O(x, \epsilon)$ contains all points x_n starting with some one of them (i.e. if for every $\epsilon > 0$ we can find a natural number N_ϵ such that $O(x, \epsilon)$ contains all points x_n with $n > N_\epsilon$). The point x is said to be the *limit* of the sequence $\{x_n\}$.

This definition can obviously be formulated in the following form: the sequence $\{x_n\}$ converges to x if $\lim_{n\to\infty} \rho(x, x_n) = 0$.

The following assertions follow directly from the definition of limit: 1) no sequence can have two distinct limits; 2) if the sequence $\{x_n\}$ converges to the point x then every subsequence of $\{x_n\}$ converges to the same point x.

The following theorem establishes the close connection between the concepts of contact point and limit point on the one hand and the concept of limit on the other.

THEOREM 5. *A necessary and sufficient condition that the point x be a contact point of the set M is that there exist a sequence $\{x_n\}$ of points of the set M which converges to x; a necessary and sufficient condition that the point x be a limit point of M is that there exist a sequence of distinct points of the set M which converges to x.*

Proof. Necessity. If x is a contact point of the set M, then every neighborhood $O(x, 1/n)$ contains at least one point x_n of M. These points form a sequence which converges to x. If the point x is a limit point of M, every neighborhood $O(x, 1/n)$ contains a point $x_n \in M$ which is distinct from all the $x_i (i < n)$ (since the number of such points is finite). The points x_n are distinct and form a sequence which converges to x.

Sufficiency is obvious.

Let A and B be two sets in the metric space R. The set A is said to be *dense* in B if $[A] \supseteq B$. In particular, the set A is said to be *everywhere dense* in R if its closure $[A]$ coincides with the entire space R. For example, the set of rational numbers is everywhere dense on the real line.

EXAMPLES OF SPACES CONTAINING AN EVERYWHERE DENSE DENUMERABLE SET. (They are sometimes called "separable." For another definition of such spaces in terms of the concept of basis see §10, Theorem 4.) We shall consider the very same examples which were pointed out in §8.

1. The space described in Example 1, §8, is separable if, and only if, it consists of a denumerable number of points. This follows directly from the fact that in this space $[M] = M$ for an arbitrary set M.

All spaces enumerated in Examples 2–7, §8, are separable. We shall indicate a denumerable everywhere dense set in each of them and leave the details of the proof to the reader.

2. Rational points.

3. The set of all vectors with rational coordinates.

4. The set of all vectors with rational coordinates.

5. The set of all polynomials with rational coefficients.

6. The set of all sequences in each of which all terms are rational and only a finite (but arbitrary) number of terms is distinct from zero.

7. The set of all polynomials with rational coefficients.

The space of bounded sequences (Example 8, §8) is not separable. In fact, let us consider all possible sequences consisting of zeros and ones. They form a set with cardinal number that of the continuum (since each of them can be put into correspondence with the dyadic development of some real number which is contained in the interval [0, 1]). The distance between two such distinct points defined by formula (6), §8, is 1. We surround each of these points with a sphere of radius $\frac{1}{2}$. These spheres do not intersect. If

some set is everywhere dense in the space under consideration, then each of the indicated spheres should contain at least one point of this set and consequently it cannot be denumerable.

(1) Let A be an arbitrary set in the metric space R and let x be a point in R. The distance from the point x to the set A is defined by the number

$$\rho(A, x) = \inf \{\rho(a, x); a \in A\}.$$

If $x \in A$, then $\rho(A, x) = 0$; but the fact that $\rho(A, x) = 0$ does not imply that $x \in A$. From the definition of contact point it follows immediately that $\rho(A, x) = 0$ if, and only if, x is a contact point of the set A.

Thus, the closure $[A]$ of the set A can be defined as the totality of all those points whose distance from the set A is zero.

(2) We can define the distance between two sets analogously. If A and B are two sets in R, then

$$\rho(A, B) = \inf \{\rho(a, b); a \in A, b \in B\}.$$

If $A \cap B \neq \theta$, then $\rho(A, B) = 0$; the converse is not true in general.

(3) If A is a set in the metric space R then the totality A' of its limit points is called its *derived set.*

Although the application to $[M]$ once more of the operation of closure always results again in $[M]$, the equality $(M')' = M'$ does not hold in general. In fact, if we take, for example, the set A of points of the form $1/n$ on the real line, then its derived set A' consists of the single point 0, but the set $A'' = (A')'$ will already be the void set. If we consider on the real line the set B of all points of the form $1/n + 1/(nm)$ $(n, m = 1, 2, \cdots)$, then $B' = A \cup A'$, B'' is the point 0, and B''' is the void set.

§10. Open and closed sets

In this section we shall consider the more important types of sets in a metric space; these are the open and closed sets.

A set M in a metric space R is said to be *closed* if it coincides with its closure: $[M] = M$. In other words, a set is said to be closed if it contains all its limit points.

By Theorem 1, §9, the closure of an arbitrary set M is a closed set. Theorem 2, §9, implies that $[M]$ is the smallest closed set which contains M.

EXAMPLES. 1. An arbitrary closed interval $[a, b]$ on the real line is a closed set.

2. The closed sphere is a closed set. In particular, in the space $C[a, b]$ the set of functions satisfying the condition $|f| \leq K$ is closed.

3. The set of functions satisfying the condition $|f| < K$ (open sphere) is not closed; its closure is the set of functions satisfying the condition $|f| \leq K$.

4. Whatever the metric space R, the void set and the whole space R are closed sets.

5. Every set consisting of a finite number of points is closed.

The fundamental properties of closed sets can be formulated in the form of the following theorem.

THEOREM 1. *The intersection of an arbitrary number and the sum of an arbitrary finite number of closed sets are closed sets.*

Proof. Let $F = \bigcap_\alpha F_\alpha$, where the F_α are closed sets. Further, let x be a limit point of the set F. This means that an arbitrary neighborhood $O(x, \epsilon)$ of x contains an infinite number of points of F. But then $O(x, \epsilon)$ contains an infinite number of points of each F_α and consequently, since all the F_α are closed, the point x belongs to each F_α; thus, $x \in F = \bigcap_\alpha F_\alpha$, i.e. F is closed.

Now let F be the sum of a finite number of closed sets: $F = \bigcup_{i=1}^n F_i$, and let x be a point not belonging to F. We shall show that x cannot be a limit point of F. In fact, x does not belong to any of the closed sets F_i and consequently it is not a limit point of any of them. Therefore for every i we can find a neighborhood $O(x, \epsilon_i)$ of the point x which does not contain more than a finite number of points of F_i. If we take the smallest of the neighborhoods $O(x, \epsilon_1), \cdots, O(x, \epsilon_n)$, we obtain a neighborhood $O(x, \epsilon)$ of the point x which does not contain more than a finite number of points of F.

Thus, if the point x does not belong to F, it cannot be a limit point of F, i.e. F is closed. This completes the proof of the theorem.

The point x is said to be an *interior point* of the set M if there exists a neighborhood $O(x, \epsilon)$ of the point x which is contained entirely in M.

A set all of whose points are interior points is said to be an *open set.*

EXAMPLES. 6. The interval (a, b) of the real line D^1 is an open set; in fact, if $a < \alpha < b$, then $O(\alpha, \epsilon)$, where $\epsilon = \min(\alpha - a, b - \alpha)$, is contained entirely in the interval (a, b).

7. The open sphere $S(a, r)$ in an arbitrary metric space R is an open set. In fact, if $x \in S(a, r)$, then $\rho(a, x) < r$. We set $\epsilon = r - \rho(a, x)$. Then $S(x, \epsilon) \subseteq S(a, r)$.

8. The set of continuous functions satisfying the condition $|f| < K$, where K is an arbitrary number, is an open subset of the space $C[a, b]$.

THEOREM 2. *A necessary and sufficient condition that the set M be open is that its complement $R \setminus M$ with respect to the whole space R be closed.*

Proof. If M is open, then each point $x \in M$ has a neighborhood which belongs entirely to M, i.e. which does not have a single point in common with $R \setminus M$. Thus, no point which does not belong to $R \setminus M$ can be a contact point of $R \setminus M$, i.e. $R \setminus M$ is closed. Conversely, if $R \setminus M$ is closed, an arbitrary point of M has a neighborhood which lies entirely in M, i.e. M is open.

Since the void set and the whole space R are closed and are at the same time complements of each other, the theorem proved above implies the following corollary.

COROLLARY. The void set and whole space R are open sets.

The following important theorem which is the dual of Theorem 1 follows from Theorem 1 and the principle of duality established in §1 (the intersection of complements equals the complement of the sums, the sum of the complements equals the complement of the intersections).

THEOREM 1'. *The sum of an arbitrary number and the intersection of an arbitrary finite number of open sets are open sets.*

A family $\{G_\alpha\}$ of open sets in R is called a *basis* in R if every open set in R can be represented as the sum of a (finite or infinite) number of sets belonging to this family.

To check whether or not a given family of open sets is a basis we find the following criterion useful.

THEOREM 3. *A necessary and sufficient condition that a system of open sets $\{G_\alpha\}$ be a basis in R is that for every open set G and for every point $x \in G$ a set G_α can be found in this system such that $x \in G_\alpha \subset G$.*

Proof. If $\{G_\alpha\}$ is a basis, then every open set G is a sum of G_α's: $G = \bigcup_i G_{\alpha_i}$, and consequently every point x in G belongs to some G_α contained in G. Conversely, if the condition of the theorem is fulfilled, then $\{G_\alpha\}$ is a basis. In fact, let G be an arbitrary open set. For each point $x \in G$ we can find some $G_\alpha(x)$ such that $x \in G_\alpha \subset G$. The sum of these $G_\alpha(x)$ over all $x \in G$ equals G.

With the aid of this criterion it is easy to establish that in every metric space the family of all open spheres forms a basis. The family of all spheres with rational radii also forms a basis. On the real line a basis is formed, for example, by the family of all rational intervals (i.e. intervals with rational endpoints).

We shall say that a set is countable if it is either finite or denumerable.

R is said to be a *space with countable basis* or to satisfy the *second axiom of countability* if there is at least one basis in R consisting of a countable number of elements.

THEOREM 4. *A necessary and sufficient condition that R be a space with countable basis is that there exist in R an everywhere dense countable set.* (A finite everywhere dense set occurs only in spaces consisting of a finite set of points.)

Proof. Necessity. Let R have a countable basis $\{G_n\}$. Choose from each G_n an arbitrary point x_n. The set $\{x_n\}$ obtained in this manner is everywhere dense in R. In fact, let x be an arbitrary point in R and let $O(x, \epsilon)$ be a neighborhood of x. According to Theorem 3, a set G_n can be found such that $x \in G_n \subset O(x, \epsilon)$. Since G_n contains at least one of the points of the set $\{x_n\}$, any neighborhood $O(x, \epsilon)$ of an arbitrary point $x \in R$ contains at least one point from $\{x_n\}$ and this means that $\{x_n\}$ is everywhere dense in R.

Sufficiency. If $\{x_n\}$ is a countable everywhere dense set in R, then the family of spheres $S(x_n, 1/k)$ forms a countable basis in R. In fact, the set of all these spheres is countable (being the sum of a countable family of countable sets). Further, let G be an arbitrary open set and let x be any point in G. By the definition of an open set an $m > 0$ can be found such that the sphere $S(x, 1/m)$ lies entirely in G. We now select a point x_{n_0} from the set $\{x_n\}$ such that $\rho(x, x_{n_0}) < 1/3m$. Then the sphere $S(x_{n_0}, 1/2m)$ contains the point x and is contained in $S(x, 1/m)$ and consequently in G also. By virtue of Theorem 3 it follows from this that the spheres $S(x_n, 1/k)$ form a basis in R.

By virtue of this theorem, the examples introduced above (§ 9) of separable spaces are at the same time examples of spaces with countable basis.

We say that a system of sets M_α is a *covering* of the space R if $\bigcup M_\alpha = R$. A covering consisting of open (closed) sets will be called an *open* (*closed*) covering.

THEOREM 5. *If R is a metric space with countable basis, then we can select a countable covering from each of its open coverings.*

Proof. Let $\{O_\alpha\}$ be an arbitrary open covering of R. Thus, every point $x \in R$ is contained in some O_α.

Let $\{G_n\}$ be a countable basis in R. Then for every $x \in R$ there exists a $G_n(x) \in \{G_n\}$ and an α such that $x \in G_n(x) \subset O_\alpha$. The family of sets $G_n(x)$ selected in this way is countable and covers R. If we choose for each of the $G_n(x)$ one of the sets O_α containing it, we obtain a countable subcovering of the covering $\{O_\alpha\}$.

It was already indicated above that the void set and the entire space R are simultaneously open and closed. A space in which there are no other sets which are simultaneously open and closed is said to be *connected*. The real line R^1 is one of the simplest examples of a connected metric space. But if we remove a finite set of points (for example, one point) from R^1, the remaining space is no longer connected. The simplest example of a space which is not connected is the space consisting of two points which are at an arbitrary distance from one another.

(1) Let M_k be the set of all functions f in $C[a, b]$ which satisfy a so-called Lipschitz condition

$$|f(t_1) - f(t_2)| \leq K|t_1 - t_2|,$$

where K is a constant. The set M_k is closed. It coincides with the closure of the set of all differentiable functions which are such that $|f'(t)| \leq K$.

(2) The set $M = \bigcup_k M_k$ of all functions each of which satisfies a Lipschitz condition for some K is not closed. Since M contains the set of all polynomials, its closure is the entire space $C[a, b]$.

(3) Let distance be defined in the space X in two different ways, i.e. let there be given two distinct metrics $\rho_1(x, y)$ and $\rho_2(x, y)$. The metrics ρ_1 and ρ_2 are said to be *equivalent* if there exist two positive constants a and b such that $a < [\rho_1(x, y)/\rho_2(x, y)] < b$ for all $x \neq y$ in R. If an arbitrary set $M \subseteq X$ is closed (open) in the sense of the metric ρ_1, then it is closed (open) in the sense of an arbitrary metric ρ_2 which is equivalent to ρ_1.

(4) A number of important definitions and assertions concerning metric spaces (for example, the definition of connectedness) do not make use of the concept of metric itself but only of the concept of open (closed) set, or, what is essentially the same, the concept of neighborhood. In particular, in many questions the metric introduced in a metric space can be replaced by any other metric which is equivalent to the initial metric. This point of view leads naturally to the concept of topological space, which is a generalization of metric space.

A *topological space* is a set T of elements of an arbitrary nature (called points of this space) some subsets of which are labeled open sets. In this connection we assume that the following axioms are fulfilled:

1. T and the void set are open;

2. The sum of an arbitrary (finite or infinite) number and the intersection of an arbitrary finite number of open sets are open.

The sets $T \setminus G$, the complements of the open sets G with respect to T, are said to be closed. Axioms 1 and 2 imply the following two assertions.

1′. The void set and T are closed;

2′. The intersection of an arbitrary (finite or infinite) number and the sum of an arbitrary finite number of closed sets are closed.

A neighborhood of the point $x \in T$ is any open set containing x.

In a natural manner we introduce the concepts of contact point, limit point, and closure: $x \in T$ is said to be a contact point of the set M if every neighborhood of the point x contains at least one point of M; x is said to be a limit point of the set M if every neighborhood of the point x contains an infinite number of points of M. The totality of all contact points of the set M is called the closure $[M]$ of the set M.

It can easily be shown that closed sets (defined as the complements of open sets), and only closed sets, satisfy the condition $[M] = M$. As also in the case of a metric space $[M]$ is the smallest closed set containing M.

Similarly, as a metric space is the pair: set of points and a metric, so a topological space is the pair: set of points and a topology defined in this space. To introduce a topology into T means to indicate in T those subsets which are to be considered open in T.

EXAMPLES. (4-a) By virtue of Theorem 1′ open sets in a metric space satisfy Conditions 1 and 2 in the definition of a topological space. Thus, every metric space can be considered as a topological space.

(4-b) Let T consist of two points a and b and let the open sets in T be T, the void set, and the set consisting of the single point b. Axioms 1 and 2 are fulfilled. The closed sets are T, the void set, and the set consisting of the single point a. The closure of the set consisting of the point b is all of T.

(5) A topological space T is said to be *metrizable* if a metric can be introduced into the set T so that the sets which are open in the sense of this metric coincide with the open sets of the initial topological space. The space (4-b) is an example of a topological space which cannot be metrized.

(6) Although many fundamental concepts carry over from metric spaces to topological spaces defined in (4), this concept turns out to be too general in a number of cases. An important class of topological spaces consists of those spaces which satisfy, in addition to Axioms 1 and 2, the Hausdorff separation axiom:

3. Any two distinct points x and y of the space T have disjoint neighborhoods.

A topological space satisfying this axiom is called a *Hausdorff space*. Clearly, every metric space is a Hausdorff space. The space pointed out in Example (4-b) does not satisfy the Hausdorff axiom.

§11. Open and closed sets on the real line

The structure of open and closed sets in an arbitrary metric space can be very complicated. We shall now consider the simplest special case, namely that of open and closed sets on the real line. In this case their complete description does not present much of a problem and is given by the following theorem.

THEOREM 1. *Every open set on the real line is the sum of a countable number of disjoint intervals.*

Proof. [We shall also include sets of the form $(-\infty, \infty)$, (α, ∞), $(-\infty, \beta)$ as intervals.] Let G be an open set and let $x \in G$. Then by the definition of an open set we can find some interval I which contains the point x and belongs entirely to the set G. This interval can always be chosen so that its endpoints are rational. Having taken for every point $x \in G$ a corresponding interval I, we obtain a covering of the set G by means of a denumerable system of intervals (this system is denumerable because the set of all intervals with rational endpoints is denumerable). Furthermore, we shall say that the intervals I' and I'' (from the same covering) belong to one class if there exists a finite chain of intervals:

$$I' = I_1, I_2, \cdots, I_n = I''$$

(belonging to our covering) such that I_k intersects I_{k+1} $(1 \leq k \leq n-1)$. It is clear that there will be a countable number of such classes. Further, the union of all the intervals which belong to the same class obviously again forms an interval U of the same type, and intervals corresponding to distinct classes do not intersect. This completes the proof of the theorem.

Since closed sets are the complements of open sets, it follows that every closed set on the real line is obtained by removing a finite or denumerable number of open intervals on the real line.

The simplest examples of closed sets are segments, individual points, and the sum of a finite number of such sets. We shall now consider a more complicated example of a closed set on the real line, the so-called Cantor set.

Let F_0 be the closed interval [0, 1]. We remove the open interval $(\frac{1}{3}, \frac{2}{3})$ from F_0 and denote the remaining closed set by F_1. Then we remove the open intervals $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$ from F_1 and denote the remaining closed set (consisting of four closed intervals) by F_2. From each of these four intervals we remove the middle interval of length $(\frac{1}{9})^3$, and so forth. If we continue this process, we obtain a decreasing sequence of closed sets F_n. We set $F = \bigcap_{n=0}^{\infty} F_n$; F is a closed set (since it is the intersection of the closed sets F_n). It is obtained from the closed interval [0, 1] by removing a denumerable number of open intervals. Let us consider the structure of the set F. The points

$$0,\ 1,\ \tfrac{1}{3},\ \tfrac{2}{3},\ \tfrac{1}{9},\ \tfrac{2}{9},\ \tfrac{7}{9},\ \tfrac{8}{9},\ \cdots \tag{1}$$

which are the endpoints of the deleted intervals obviously belong to F. However, the set F is not exhausted by these points. In fact, those points of the closed interval [0, 1] which belong to the set F can be characterized in the following manner. We shall write each of the numbers x, $0 \leq x \leq 1$, in the triadic system:

$$x = a_1/3 + a_2/3^2 + \cdots + a_n/3^n + \cdots,$$

where the numbers a_n can assume the values 0, 1, and 2. As in the case of the ordinary decimal expansion, some numbers allow two different developments. For example,

$$\tfrac{1}{3} = \tfrac{1}{3} + (\tfrac{0}{3})^2 + \cdots + (\tfrac{0}{3})^n + \cdots = \tfrac{0}{3} + (\tfrac{2}{3})^2 + (\tfrac{2}{3})^3 + \cdots + (\tfrac{2}{3})^n + \cdots.$$

It is easily verified that the set F contains those, and only those, numbers x, $0 \leq x \leq 1$, which can be written in at least one way in the form of a triadic fraction such that the number 1 does not appear in the sequence $a_1, a_2, \cdots, a_n, \cdots$. Thus, to each point $x \in F$ we can assign the sequence

$$a_1, a_2, \cdots, a_n, \cdots, \tag{2}$$

where a_n is 0 or 2. The set of all such sequences forms a set having the power of the continuum. We can convince ourselves of this by assigning to each sequence (2) the sequence

$$b_1, b_2, \cdots, b_n, \cdots, \tag{2'}$$

where $b_n = 0$ if $a_n = 0$ and $b_n = 1$ if $a_n = 2$. The sequence (2') can be considered as the development of a real number y, $0 \leq y \leq 1$, in the form of a dyadic fraction. We thus obtain a mapping of the set F onto the entire closed interval [0, 1]. This implies that F has the cardinal number of the continuum. [The correspondence established between F and the closed interval [0, 1] is single-valued but it is not one-to-one (because of the fact that the same number can sometimes be formed from distinct fractions). This implies that F has cardinal number not less than the cardinal number of the continuum. But F is a subset of the closed interval [0, 1] and consequently its cardinal number cannot be greater than that of the continuum. (See §5.)] Since the set of points (1) is denumerable, these points cannot exhaust all of F.

EXERCISE. Prove directly that the point $\frac{1}{4}$ belongs to the set F although it is not an endpoint of a single one of the intervals deleted. *Hint*: The point $\frac{1}{4}$ divides the closed interval [0, 1] in the ratio 1:3. The closed interval [0, $\frac{1}{3}$] which remains after the first deletion is also divided in the ratio 1:3 by the point $\frac{1}{4}$, and so on.

The points (1) are said to be points of the first type of the set F and the remaining points are said to be points of the second type.

EXERCISE. Prove that the points of the first type form an everywhere dense set in F.

We have shown that the set F has the cardinal number of the continuum, i.e. that it contains as many points as the entire closed interval [0, 1].

It is interesting to compare this fact with the following result: the sum of the lengths of all the deleted intervals is $\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots$, i.e. exactly 1!

§12. Continuous mappings. Homeomorphism. Isometry

Let $R = (X, \rho)$ and $R' = (Y, \rho')$ be two metric spaces. The mapping f of the space R into R' is said to be *continuous at the point* $x_0 \in R$ if for arbitrary $\epsilon > 0$ a $\delta > 0$ can be found such that

$$\rho'[f(x), f(x_0)] < \epsilon$$

for all x such that

$$\rho(x, x_0) < \delta.$$

In other words, the mapping f is continuous at the point x_0 if an arbitrary

neighborhood $O(f(x_0), \epsilon)$ of the point $f(x_0)$ contains a neighborhood $O(x_0, \delta)$ of the point x_0 whose image is contained in the interior of $O(f(x_0), \epsilon)$.

A mapping f is said to be *continuous* if it is continuous at each point of the space R.

If R' is the real line, then a continuous mapping of R into R' is called a *continuous function* on R.

As in the case of the mapping of arbitrary sets we shall say that f is a mapping of R *onto* R' if every element $y \in R'$ has at least one inverse image.

In analysis, together with the definition of the continuity of a function "in terms of neighborhoods", the definition of continuity "in terms of sequences", which is equivalent to it, is widely used. The situation is analogous also in the case of continuous mappings of arbitrary metric spaces.

THEOREM 1. *A necessary and sufficient condition that the mapping f be continuous at the point x is that for every sequence $\{x_n\}$ which converges to x the corresponding sequence $\{f(x_n)\}$ converge to $y = f(x)$.*

Proof. The necessity is obvious. We shall prove the sufficiency of this condition. If the mapping f is not continuous at the point x, there exists a neighborhood $O(y, \epsilon)$ of the point $y = f(x)$ such that an arbitrary $O(x, \delta)$ contains points whose images do not belong to $O(y, \epsilon)$. Setting $\delta_n = 1/n$ $(n = 1, 2, \cdots)$, we select in each sphere $O(x, 1/n)$ a point x_n such that $f(x_n) \notin O(y, \epsilon)$. Then $x_n \to x$ but the sequence $\{f(x_n)\}$ does not converge to $f(x)$, i.e. the condition of the theorem is not satisfied, which was to be proved.

THEOREM 2. *A necessary and sufficient condition that the mapping f of the space R onto R' be continuous is that the inverse image of each closed set in R' be closed.*

Proof. Necessity. Let $M \subseteq R$ be the complete inverse image of the closed set $M' \subseteq R'$. We shall prove that M is closed. If $x \in [M]$, there exists a sequence $\{x_n\}$ of points in M which converges to x. But then, by Theorem 1, the sequence $\{f(x_n)\}$ converges to $f(x)$. Since $f(x_n) \in M'$ and M' is closed, we have $f(x) \in M'$; consequently $x \in M$, which was to be proved.

Sufficiency. Let x be an arbitrary point in R, $y = f(x)$, and let $O(y, \epsilon)$ be an arbitrary neighborhood of y. The set $R' \setminus O(y, \epsilon)$ is closed (since it is the complement of an open set). By assumption, $F = f^{-1}(R' \setminus O(y, \epsilon))$ is closed, and moreover, $x \notin F$. Thus, $R \setminus F$ is open and $x \in R \setminus F$; consequently, there is a neighborhood $O(x, \delta)$ of the point x which is contained in $R \setminus F$. If $z \in O(x, \delta)$, then $f(z) \in O(y, \epsilon)$, i.e. f is continuous, which was to be proved.

REMARK. The image of a closed set under a continuous mapping is not necessarily closed as is shown by the following example: map the half-open interval $[0, 1)$ onto a circle of the same length. The set $[\frac{1}{2}, 1)$ which is closed in $[0, 1)$, goes over under this mapping into a set which is not closed (see Fig. 8).

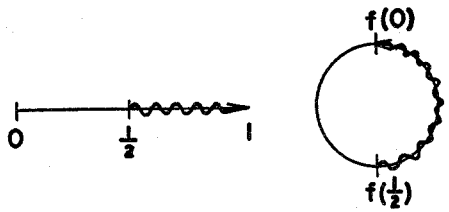

FIG. 8

Since in the case of a mapping "onto" the inverse image of the complement equals the complement of the inverse image, the following theorem which is the dual of Theorem 2 is valid.

THEOREM 2'. *A necessary and sufficient condition that the mapping f of the space R onto R' be continuous is that the inverse image of each open set in R' be open.*

The following theorem which is the analogue of the well-known theorem from analysis on the continuity of a composite function is valid for continuous mappings.

THEOREM 3. *If R, R', R'' are metric spaces and f and φ are continuous mappings of R into R' and R' into R'', respectively, then the mapping $z = \varphi(f(x))$ of the space R into R'' is continuous.*

The proof is carried out exactly as for real-valued functions.

The mapping f is said to be a *homeomorphism* if it is one-to-one and bicontinuous (i.e. both f and the inverse mapping f^{-1} are continuous).

The spaces R and R' are said to be *homeomorphic* if a homeomorphic correspondence can be established between them.

It is easy to see that two arbitrary intervals are homeomorphic, that an arbitrary open interval is homeomorphic to R^1, and so forth.

It follows from Theorems 2 and 2' of this section that a necessary and sufficient condition that a one-to-one mapping be a homeomorphism is that the closed (open) sets correspond to closed (open) sets.

This implies that a necessary and sufficient condition that a one-to-one mapping φ be a homeomorphism is that the equality

$$\varphi([M]) = [\varphi(M)]$$

hold for arbitrary M. (This follows from the fact that $[M]$ is the intersection of all closed sets which contain M, i.e. it is the minimal closed set which contains M.)

EXAMPLE. Consider the spaces $R^n = (D^n, \rho)$ and $R_0^n = (D^n, \rho_0)$ (see §8,

Examples 3 and 4). The following inequalities hold for the mapping which assigns to an element in R^n with coordinates $x_1, x_2, \ldots, x_n$ the element in R_0^n with the same coordinates:

$$\rho_0(x, y) \leq \rho(x, y) \leq n^{\frac{1}{2}}\rho_0(x, y).$$

Consequently an arbitrary ϵ-neighborhood of the point x of the space R^n contains a δ-neighborhood of the same point x considered as an element of the space R_0^n, and conversely. It follows from this that our mapping of R^n onto R_0^n is a homeomorphism.

An important special case of a homeomorphism is an isometric mapping.

We say that a one-to-one mapping $y = f(x)$ of a metric space R onto a metric space R' is *isometric* if

$$\rho(x_1, x_2) = \rho'[f(x_1), f(x_2)]$$

for arbitrary $x_1, x_2 \in R$. The spaces R and R' themselves, between which an isometric correspondence can be established, are said to be *isometric*.

The isometry of two spaces R and R' means that the metric relations between their elements are the same and that they can differ only in the nature of their elements, which is unessential. In the sequel we shall consider two isometric spaces simply as identical.

(1) The concept of continuity of a mapping can be defined not only for metric but also for arbitrary topological spaces. The mapping f of the topological space T into the topological space T' is said to be continuous at the point x_0 if for arbitrary neighborhood $O(y_0)$ of the point $y_0 = f(x_0)$ there exists a neighborhood $O(x_0)$ of the point x_0 such that $f(O(x_0)) \subset O(y_0)$.

Theorems 2 and 3 carry over automatically to continuous mappings of topological spaces.

§13. Complete metric spaces

From the very beginning of our study of mathematical analysis we are convinced of the important role in analysis that is played by the property of completeness of the real line, i.e. the fact that every fundamental sequence of real numbers converges to some limit. The real line represents the simplest example of the so-called complete metric spaces whose basic properties we shall consider in this section.

We shall call a sequence $\{x_n\}$ of points of a metric space R a *fundamental sequence* if it satisfies the Cauchy criterion, i.e. if *for arbitrary* $\epsilon > 0$ *there exists an* N_ϵ *such that* $\rho(x_{n'}, x_{n''}) < \epsilon$ *for all* $n' \geq N_\epsilon$, $n'' \geq N_\epsilon$.

It follows directly from the triangle axiom that every convergent sequence is fundamental. In fact, if $\{x_n\}$ converges to x, then for given $\epsilon > 0$ it is possible to find a natural number N_ϵ such that $\rho(x_n, x) < \epsilon/2$ for all $n \geq N_\epsilon$. Then $\rho(x_{n'}, x_{n''}) \leq \rho(x_{n'}, x) + \rho(x_{n''}, x) < \epsilon$ for arbitrary $n' \geq N_\epsilon$ and $n'' \geq N_\epsilon$.

DEFINITION 1. If every fundamental sequence in the space R converges to an element in R, R is said to be *complete*.

EXAMPLES. All the spaces considered in §8, with the exception of the one given in Example 7, are complete. In fact:

1. In the space consisting of isolated points (Example 1, §8) only those sequences in which there is a repetition of some point, beginning with some index, are fundamental. Clearly, every such sequence converges, i.e. this space is complete.

2. The completeness of the space R^1 of real numbers is known from analysis.

3. The completeness of the Euclidean space R^n follows directly from the completeness of R^1. In fact, let $\{x_i\}$ be a fundamental sequence; this means that for every $\epsilon > 0$ an $N = N_\epsilon$ can be found such that

$$\sum_{k=1}^{n} (x_p^{(k)} - x_q^{(k)})^2 < \epsilon^2$$

for all p, q greater than N. Then for each $k = 1, 2, \cdots, n$

$$| x_p^{(k)} - x_q^{(k)} | < \epsilon$$

for all $p, q > N$, i.e. $\{x_p^{(k)}\}$ is a fundamental sequence of real numbers. We set

$$x^{(k)} = \lim_{p\to\infty} x_p^{(k)},$$

and

$$x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)}).$$

Then it is obvious that

$$\lim_{n\to\infty} x_n = x.$$

4. The completeness of the space R_0^n is proved in an exactly analogous manner.

5. We shall prove the completeness of the space $C[a, b]$. Let $\{x_n(t)\}$ be a fundamental sequence in $C[a, b]$. This means that for each $\epsilon > 0$ there exists an N such that $| x_n(t) - x_m(t) | < \epsilon$ for $n, m > N$ and all t, $a \leq t \leq b$. This implies that the sequence $\{x_n(t)\}$ converges uniformly and that its limit is a continuous function $x(t)$, where

$$| x_n(t) - x(t) | < \epsilon$$

for all t and for all n larger than some N; this means that $\{x_n(t)\}$ converges to $x(t)$ in the sense of the metric of the space $C[a, b]$.

6. The space l_2. Let $\{x^{(n)}\}$, where

$$x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \cdots, x_k^{(n)}, \cdots),$$

be a fundamental sequence in l_2.

For arbitrary $\epsilon > 0$ an N can be found such that

(1) $\rho^2(x^{(n)}, x^{(m)}) = \sum_{k=1}^{\infty} (x_k^{(n)} - x_k^{(m)})^2 < \epsilon$ for $n, m > N$.

It follows from this that for arbitrary k

$$(x_k^{(n)} - x_k^{(m)})^2 < \epsilon,$$

i.e. for each k the sequence of real numbers $\{x_k^{(n)}\}$ converges. Set $\lim_{n\to\infty} x_k^{(n)} = x_k$. Denote the sequence $(x_1, x_2, \cdots, x_n, \cdots)$ by x. We must show that

a) $\sum_{k=1}^{\infty} x_k^2 < \infty$; b) $\lim_{n\to\infty} \rho(x^{(n)}, x) = 0$.

To this end we shall write inequality (1) in the form

$$\sum_{k=1}^{\infty} (x_k^{(n)} - x_k^{(m)})^2 = \sum_{k=1}^{M} (x_k^{(n)} - x_k^{(m)})^2 + \sum_{k=M+1}^{\infty} (x_k^{(n)} - x_k^{(m)})^2 < \epsilon$$

(M arbitrary). Since each of these two sums is nonnegative, each of them is less than ϵ. Consequently

$$\sum_{k=1}^{M} (x_k^{(n)} - x_k^{(m)})^2 < \epsilon.$$

If we fix m in this inequality and pass to the limit as $n \to \infty$, we obtain

$$\sum_{k=1}^{M} (x_k - x_k^{(m)})^2 \leq \epsilon.$$

Since this inequality is valid for arbitrary M, we can pass to the limit as $M \to \infty$. We then obtain

$$\sum_{k=1}^{\infty} (x_k - x_k^{(m)})^2 \leq \epsilon.$$

The inequality thus obtained and the convergence of the series $\sum_{k=1}^{\infty} x_k^{(m)2}$ imply that the series $\sum_{k=1}^{\infty} x_k^2$ converges; consequently x is an element in l_2. Further, since ϵ is arbitrarily small, this inequality means that

$$\lim_{m\to\infty} \rho(x^{(m)}, x) = \lim_{m\to\infty} \{\sum_{k=1}^{\infty} (x_k - x_k^{(m)})^2\}^{\frac{1}{2}} = 0,$$

i.e. $x^{(n)} \to x$.

7. It is easy to convince ourselves of the fact that the space $C^2[a, b]$ is not complete. For example, the sequence of continuous functions

$$\varphi_n(t) = \arctan nt \qquad (-1 \leq t \leq 1)$$

is fundamental, but it does not converge to any continuous function (it converges in the sense of mean square deviation to the discontinuous function which is equal to $-\pi/2$ for $t < 0$, $\pi/2$ for $t > 0$, and 0 for $t = 0$).

EXERCISE. Prove that the space of all bounded sequences (Example 8, §8) is complete.

In analysis the so-called lemma on nested segments is widely used. In the theory of metric spaces an analogous role is played by the following theorem which is called the *principle of nested spheres.*

THEOREM 1. *A necessary and sufficient condition that the metric space R be complete is that every sequence of closed nested spheres in R with radii tending to zero have nonvoid intersection.*

Proof. Necessity. Assume the space R is complete and let $S_1, S_2, S_3, \cdots$ be a sequence of closed nested spheres. Let d_n be the diameter of the sphere S_n. By hypothesis $\lim_{n\to\infty} d_n = 0$. Denote the center of the sphere S_n by x_n. The sequence $\{x_n\}$ is fundamental. In fact, if $m > n$, then obviously $\rho(x_n, x_m) < d_n$. Since R is complete, $\lim_{n\to\infty} x_n$ exists. If we set

$$x = \lim_{n\to\infty} x_n,$$

then $x \in \bigcap_n S_n$. In fact, the sphere S_n contains all the points of the given sequence with the exception perhaps of the points $x_1, x_2, \cdots, x_{n-1}$. Thus, x is a limit point of each sphere S_n. But since S_n is a closed set, we have that $x \in S_n$ for all n.

Sufficiency. To prove the sufficiency we shall show that if the space R is not complete, i.e. if there exists a fundamental sequence in R which does not have a limit, then it is possible to construct a sequence of closed nested spheres in R whose diameters tend to zero and whose intersection is void. Let $\{x_n\}$ be a fundamental sequence of points in R which does not have a limit. We shall construct a sequence of closed spheres S_n in the following way. Let n_1 be such that $\rho(x_{n_1}, x_m) < \frac{1}{2}$ for all $m > n_1$. Denote by S_1 the sphere of radius 1 and center at x_{n_1}. Further, let $n_2 > n_1$ be such that $\rho(x_{n_2}, x_m) < \frac{1}{4}$ for all $m > n_2$. Denote by S_2 the sphere of radius $\frac{1}{2}$ with center x_{n_2}. Since by assumption $\rho(x_{n_1}, x_{n_2}) < \frac{1}{2}$, we have $S_2 \subset S_1$. Now let $n_3 > n_2$ be such that $\rho(x_{n_3}, x_m) < \frac{1}{8}$ for all $m > n_3$ and let S_3 be a sphere of radius $\frac{1}{4}$ with center x_{n_3}, and so forth. If we continue this construction we obtain a sequence of closed nested spheres $\{S_n\}$, where S_n has radius $(\frac{1}{2})^{n-1}$. This sequence of spheres has void intersection; in fact, if $x \in \bigcap_k S_k$, then $x = \lim_{n\to\infty} x_n$. As a matter of fact the sphere S_k contains all points x_n beginning with x_{n_k} and consequently $\rho(x, x_n) < (\frac{1}{2})^{k-1}$ for all $n > n_k$. But by assumption the sequence $\{x_n\}$ does not have a limit. Therefore $\bigcap S_n = \theta$.

If the space R is not complete, it is always possible to embed it in an entirely definite manner in a complete space.

DEFINITION 2. Let R be an arbitrary metric space. A complete metric space R^* is said to be the *completion* of the space R if: 1) R is a subspace of the space R^*; and 2) R is everywhere dense in R^*, i.e. $[R] = R^*$. (Here $[R]$ naturally denotes the closure of the space R in R^*.)

For example, the space of all real numbers is the completion of the space of rationals.

THEOREM 2. *Every metric space has a completion and all of its completions are isometric.*

Proof. We begin by proving uniqueness. It is necessary to prove that if R^* and R^{**} are two completions of the space R, then they are isometric, i.e. there is a one-to-one mapping φ of the space R^* onto R^{**} such that 1) $\varphi(x) = x$ for all $x \in R$; and 2) if $x^* \leftrightarrow x^{**}$ and $y^* \leftrightarrow y^{**}$, then $\rho(x^*, y^*) = \rho(x^{**}, y^{**})$.

Such a mapping φ is defined in the following way. Let x^* be an arbitrary point of R^*. Then, by the definition of completion, there exists a sequence $\{x_n\}$ of points in R which converges to x^*. But the sequence $\{x_n\}$ can be assumed to belong also to R^{**}. Since R^{**} is complete, $\{x_n\}$ converges in R^{**} to some point x^{**}. We set $\varphi(x^*) = x^{**}$. It is clear that this correspondence is one-to-one and does not depend on the choice of the sequence $\{x_n\}$ which converges to the point x^*. This is then the isometric mapping sought. In fact, by construction we have $\varphi(x) = x$ for all $x \in R$. Furthermore, if we let

$$\{x_n\} \to x^* \text{ in } R^* \text{ and } \{x_n\} \to x^{**} \text{ in } R^{**},$$

$$\{y_n\} \to y^* \text{ in } R^* \text{ and } \{y_n\} \to y^{**} \text{ in } R^{**},$$

then

$$\rho(x^*, y^*) = \lim_{n\to\infty} \rho(x_n, y_n)$$

and at the same time

$$\rho(x^{**}, y^{**}) = \lim_{n\to\infty} \rho(x_n, y_n).$$

Consequently

$$\rho(x^*, y^*) = \rho(x^{**}, y^{**}).$$

We shall now prove the existence of the completion. The idea involved in the proof is the same as that in the so-called Cantor theory of real numbers. The situation here is essentially even simpler than in the theory of real numbers since there it is required further that one define all the arithmetic operations for the newly introduced objects—the irrational numbers.

Let R be an arbitrary metric space. We shall say that two fundamental sequences $\{x_n\}$ and $\{x_n'\}$ in R are equivalent (denoting this by $\{x_n\} \sim \{x_n'\}$) if $\lim_{n\to\infty} \rho(x_n, x_n') = 0$. This equivalence relation is reflexive, symmetric and transitive. It follows from this that all fundamental sequences which can be constructed from points of the space R are partitioned into equivalence classes of sequences. We shall now define the space R^* in the following manner. The points of R^* will be all possible equivalence classes of fundamental sequences and the distance between points in R^* will be defined in the following way. Let x^* and y^* be two such classes. We choose one representative from each of these two classes, i.e. we select some fundamental sequence $\{x_n\}$ and $\{y_n\}$ from each, respectively. We set

$$\rho(x^*, y^*) = \lim_{n\to\infty} \rho(x_n, y_n). \tag{2}$$

We shall prove the correctness of this definition of distance, i.e. we shall show that the limit (2) exists and does not depend on the choice of the representatives $\{x_n\} \in x^*$ and $\{y_n\} \in y^*$.

Since the sequences $\{x_n\}$ and $\{y_n\}$ are fundamental, with the aid of the triangle axiom we have for all sufficiently large n', n'':

$$\begin{aligned}
&|\rho(x_{n'}, y_{n'}) - \rho(x_{n''}, y_{n''})| \\
&\quad = |\rho(x_{n'}, y_{n'}) - \rho(x_{n'}, y_{n''}) + \rho(x_{n'}, y_{n''}) - \rho(x_{n''}, y_{n''})| \\
&\quad \le |\rho(x_{n'}, y_{n'}) - \rho(x_{n'}, y_{n''})| + |\rho(x_{n'}, y_{n''}) - \rho(x_{n''}, y_{n''})| \\
&\quad \le \rho(y_{n'}, y_{n''}) + \rho(x_{n'}, x_{n''}) < \epsilon/2 + \epsilon/2 = \epsilon.
\end{aligned}$$

Thus, the sequence of real numbers $s_n = \rho(x_n, y_n)$ satisfies the Cauchy criterion and consequently it has a limit. It remains to prove that this limit does not depend on the choice of $\{x_n\} \in x^*$ and $\{y_n\} \in y^*$. Let

$$\{x_n\}, \{x_n'\} \in x^* \quad \text{and} \quad \{y_n\}, \{y_n'\} \in y^*.$$

Now

$$\{x_n\} \sim \{x_n'\} \quad \text{and} \quad \{y_n\} \sim \{y_n'\}$$

imply that

$$\begin{aligned}
&|\rho(x_n, y_n) - \rho(x_n', y_n')| \\
&\quad = |\rho(x_n, y_n) - \rho(x_n', y_n) + \rho(x_n', y_n) - \rho(x_n', y_n')| \\
&\quad \le |\rho(x_n, y_n) - \rho(x_n', y_n)| + |\rho(x_n', y_n) - \rho(x_n', y_n')| \\
&\quad \le \rho(x_n, x_n') + \rho(y_n, y_n') \to 0,
\end{aligned}$$

i.e.

$$\lim_{n\to\infty} \rho(x_n, y_n) = \lim_{n\to\infty} \rho(x_n', y_n').$$

We shall now show that the metric space axioms are fulfilled in R^*.

Axiom 1 follows directly from the definition of equivalence of fundamental sequences.

Axiom 2 is obvious.

We shall now verify the triangle axiom. Since the triangle axiom is satisfied in the initial space R, we have

$$\rho(x_n, z_n) \le \rho(x_n, y_n) + \rho(y_n, z_n).$$

Passing to the limit as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \rho(x_n, z_n) \leq \lim_{n\to\infty} \rho(x_n, y_n) + \lim_{n\to\infty} \rho(y_n, z_n),$$

i.e.

$$\rho(x, z) \leq \rho(x, y) + \rho(y, z).$$

We shall now prove that R^* is the completion of the space R. (We use *the* keeping in mind that all completions of the space R are isometric.)

To each point $x \in R$ there corresponds some equivalence class of fundamental sequences, namely the totality of all sequences which converge to the point x.

We have:

if

$$x = \lim_{n\to\infty} x_n \text{ and } y = \lim_{n\to\infty} y_n,$$

then

$$\rho(x, y) = \lim_{n\to\infty} \rho(x_n, y_n).$$

Consequently, letting the corresponding class of fundamental sequences converging to x correspond to each point x we embed R isometrically in the space R^*.

In the sequel we shall not have to distinguish between the space R itself and its image in R^* (i.e. the totality of all equivalence classes of convergent sequences) and we can consider R to be a subset of R^*.

We shall now show that R is everywhere dense in R^*. In fact, let x^* be a point in R^* and let $\epsilon > 0$ be arbitrary. We select a representative in x^*, i.e. we choose a fundamental sequence $\{x_n\}$. Let N be such that $\rho(x_n, x_m) < \epsilon$ for all $n, m > N$. Then we have

$$\rho(x_n, x^*) = \lim_{m\to\infty} \rho(x_n, x_m) \leq \epsilon,$$

i.e. an arbitrary neighborhood of the point x^* contains a point of R. Thus we have $[R] = R^*$.

It remains to be proved that the space R^* is complete. We note, first of all, that by the construction of R^* an arbitrary fundamental sequence

$$x_1, x_2, \cdots, x_n, \cdots \tag{3}$$

consisting of points belonging to R, converges in R^* to some point, namely to the point $x^* \in R^*$, defined by the sequence (3). Further, since R is dense in R^*, then for an arbitrary fundamental sequence $x_1^*, x_2^*, \cdots, x_n^*, \cdots$ of points in R^* we can construct an equivalent sequence $x_1, x_2, \cdots, x_n, \cdots$ consisting of points belonging to R. To do this it is sufficient to take for x_n any point in R such that $\rho(x_n, x_n^*) < 1/n$.

The sequence $\{x_n\}$ thus constructed will be fundamental and by what was proved above it will converge to some point $x^* \in R^*$. But then the sequence $\{x_n^*\}$ also converges to x^*. This proves the theorem completely.

§14. Principle of contraction mappings and its applications

As examples of the applications of the concept of completeness we shall consider the so-called contraction mappings which form a useful technique for the proof of various existence and uniqueness theorems (for example, in the theory of differential equations).

Let R be an arbitrary metric space. A mapping A of the space R into itself is said to be a *contraction* if there exists a number $\alpha < 1$ such that

$$\rho(Ax, Ay) \leq \alpha\rho(x, y) \tag{1}$$

for any two points $x, y \in R$. Every contraction mapping is continuous. In fact, if $x_n \to x$, then, by virtue of (1), we also have $Ax_n \to Ax$.

THEOREM (PRINCIPLE OF CONTRACTION MAPPINGS). *Every contraction mapping defined in a complete metric space R has one and only one fixed point (i.e. the equation $Ax = x$ has one and only one solution).*

Proof. Let x_0 be an arbitrary point in R. Set $x_1 = Ax_0$, $x_2 = Ax_1 = A^2x_0$, and in general let $x_n = Ax_{n-1} = A^nx_0$. We shall show that the sequence $\{x_n\}$ is fundamental. In fact,

$$\rho(x_n, x_m) = \rho(A^nx_0, A^mx_0) \leq \alpha^n\rho(x_0, x_{m-n})$$

$$\leq \alpha^n\{\rho(x_0, x_1) + \rho(x_1, x_2) + \cdots + \rho(x_{m-n-1}, x_{m-n})\}$$

$$\leq \alpha^n\rho(x_0, x_1)\{1 + \alpha + \alpha^2 + \cdots + \alpha^{m-n-1}\} \leq \alpha^n\rho(x_0, x_1)\{1/(1 - \alpha)\}.$$

Since $\alpha < 1$, this quantity is arbitrarily small for sufficiently large n. Since R is complete, $\lim_{n\to\infty} x_n$ exists. We set $x = \lim_{n\to\infty} x_n$. Then by virtue of the continuity of the mapping A, $Ax = A \lim_{n\to\infty} x_n = \lim_{n\to\infty} Ax_n = \lim_{n\to\infty} x_{n+1} = x$.

Thus, the existence of a fixed point is proved. We shall now prove its uniqueness. If $Ax = x$, $Ay = y$, then $\rho(x, y) \leq \alpha\rho(x, y)$, where $\alpha < 1$; this implies that $\rho(x, y) = 0$, i.e. $x = y$.

The principle of contraction mappings can be applied to the proof of the existence and uniqueness of solutions obtained by the method of successive approximations. We shall consider the following simple examples.

1. $y = f(x)$, where $f(x)$ is a function defined on the closed interval $[a, b]$ satisfying the Lipschitz condition

$$|f(x_2) - f(x_1)| \leq K|x_2 - x_1|,$$

with $K < 1$, and mapping the closed interval $[a, b]$ into itself. Then f is a contraction mapping and according to the theorem proved above the

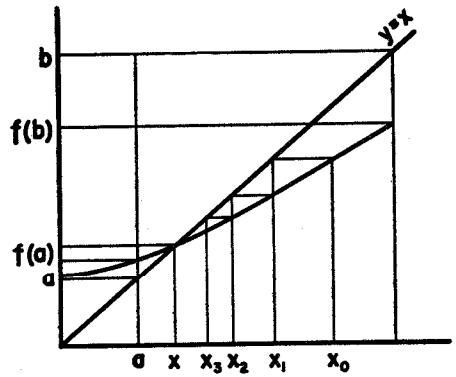

FIG. 9

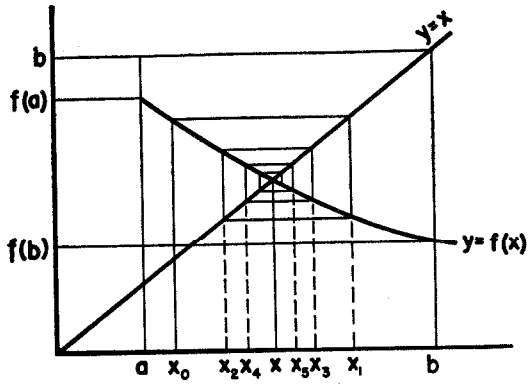

FIG. 10

sequence x_0, $x_1 = f(x_0)$, $x_2 = f(x_1)$, $\cdots$ converges to the single root of the equation $x = f(x)$.

In particular, the condition of contraction is fulfilled if $|f'(x)| \leq K < 1$ on the closed interval $[a, b]$.

As an illustration, Figs. 9 and 10 indicate the course of the successive approximations in the case $0 < f'(x) < 1$ and in the case $-1 < f'(x) < 0$.

In the case where we are dealing with an equation of the form $F(x) = 0$, where $F(a) < 0$, $F(b) > 0$ and $0 < K_1 \leq F'(x) \leq K_2$ on $[a, b]$, a widely used method for finding its root amounts to setting $f(x) = x - \lambda F(x)$ and seeking a solution of the equation $x = f(x)$, which is equivalent to $F(x) = 0$. In fact, since $f'(x) = 1 - \lambda F'(x)$, $1 - \lambda K_2 \leq f'(x) \leq 1 - \lambda K_1$ and it is not difficult to choose λ so that we can apply the method of successive approximations.

2. Let us consider the mapping $y = Ax$ of the space R^n into itself given by the system of linear equations

$$y_i = \sum_{j=1}^{n} a_{ij}x_j + b_i \qquad (i = 1, 2, \cdots, n).$$

If Ax is a contraction mapping, we can apply the method of successive approximations to the solution of the equation $x = Ax$.

Under what conditions then is the mapping A a contraction? The answer to this question depends on the choice of the metric in R^n. (It is easy to see that with the metric (b) R^n is a metric space.)

a) $$\rho(x, y) = \max \{| x_i - y_i |; 1 \leq i \leq n\};$$

$$\rho(y', y'') = \max_i |y_i' - y_i'' | = \max_i | \sum_j a_{ij}(x_j' - x_j'')|$$
$$\leq \max_i \sum_j | a_{ij} || x_j' - x_j'' | \leq \max_i \sum_j | a_{ij} | \max_j | x_j' - x_j'' |$$
$$= \max_i \sum_j | a_{ij} | \rho(x', x'').$$

This yields

(2) $$\sum_{j=1}^{n} | a_{ij} | \leq \alpha < 1$$

as the condition of contraction.

b) $$\rho(x, y) = \sum_{i=1}^{n} | x_i - y_i |;$$

$$\rho(y', y'') = \sum_i | y_i' - y_i'' | = \sum_i | \sum_j a_{ij}(x_j' - x_j'')|$$
$$\leq \sum_i \sum_j | a_{ij} ||x_j' - x_j'' | \leq \max_j \sum_i | a_{ij} | \rho(x', x'').$$

This yields the following condition of contraction:

(3) $$\sum_i | a_{ij} | \leq \alpha < 1.$$

c) $$\rho(x, y) = \{\sum_{i=1}^{n} (x_i - y_i)^2\}^{\frac{1}{2}}.$$

Here

$$\rho^2(y', y'') = \sum_i \{\sum_j a_{ij}(x_j' - x_j'')\}^2 \leq \sum_i \sum_j a_{ij}^2 \rho^2(x', x'')$$

on the basis of the Schwarz inequality.

Then

(4) $$\sum_i \sum_j a_{ij}^2 \leq \alpha < 1$$

is the contraction condition.

Thus, in the case where one of the Conditions (2)–(4) is fulfilled there exists one and only one point $x = (x_1, x_2, \cdots, x_n)$ such that

$$x_i = \sum_{j=1}^{n} a_{ij}x_j + b_i,$$

where the successive approximations to this solution have the form:

$$x^{(0)} = (x_1^0, x_2^0, \cdots, x_n^0);$$
$$x^{(1)} = (x_1^1, x_2^1, \cdots, x_n^1);$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$x^{(k)} = (x_1^k, x_2^k, \cdots, x_n^k);$$
$$\cdots\cdots\cdots\cdots\cdots\cdots,$$

where

$$x_i^k = \sum_{j=1}^n a_{ij} x_j^{k-1} + b_i .$$

(Consequently any one of the Conditions (2)–(4) implies that

$$\begin{vmatrix} a_{11} - 1 & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - 1 & \cdots & a_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - 1 \end{vmatrix} \neq 0.)$$

Each of the Conditions (2)–(4) is *sufficient* in order that the mapping $y = Ax$ be a contraction. As concerns Condition (2) it could have been proved that it is also necessary in order that the mapping $y = Ax$ be a contraction (in the sense of the metric a)).

None of the Conditions (2)–(4) is necessary for the application of the method of successive approximations. Examples can be constructed in which any one of these conditions is fulfilled but the other two are not.

If $|a_{ij}| < 1/n$ (in this case all three conditions are fulfilled), then the method of successive approximations is applicable.

If $|a_{ij}| = 1/n$ (in this case all three sums equal 1), it is easy to see that the method of successive approximations is not applicable.

§15. Applications of the principle of contraction mappings in analysis

In the preceding section there were given some of the simplest examples of the application of the principle of contraction mappings in one- and n-dimensional spaces. However, the most essential applications for analysis of the principle of contraction mappings are in infinite-dimensional function spaces. Further on we shall show how with the aid of this principle we can obtain theorems on the existence and uniqueness of solutions for some types of differential and integral equations.

I. Let

$$dy/dx = f(x, y) \tag{1}$$

be a given differential equation with the initial condition

$$y(x_0) = y_0 , \tag{2}$$

where $f(x, y)$ is defined and continuous in some plane region G which contains the point (x_0, y_0) and satisfies a Lipschitz condition with respect to y:

$$|f(x, y_1) - f(x, y_2)| \leq M |y_1 - y_2|.$$

We shall prove that then on some closed interval $|x - x_0| < d$ there exists a unique solution $y = \varphi(x)$ of the equation (1) satisfying the initial condition (2) (Picard's theorem).

Equation (1) together with the initial condition (2) is equivalent to the integral equation

$$\varphi(x) = y_0 + \int_{x_0}^{x} f(t, \varphi(t))\, dt. \tag{3}$$

Since the function $f(x, y)$ is continuous, we have $|f(x, y)| \leq k$ in some region $G' \subseteq G$ which contains the point (x_0, y_0). Now we select a $d > 0$ such that the following conditions are fulfilled:

1) $(x, y) \in G'$ if $|x - x_0| \leq d$, $|y - y_0| \leq kd$;

2) $Md < 1$.

Denote by C^* the space of continuous functions φ^* which are defined on the closed interval $|x - x_0| \leq d$ and are such that $|\varphi^*(x) - y_0| \leq kd$ with the metric $\rho(\varphi_1, \varphi_2) = \max_x |\varphi_1(x) - \varphi_2(x)|$.

It is easy to see that C^* is a complete space. (This follows, for instance, from the fact that a closed subset of a complete space is a complete space.) Let us consider the mapping $\psi = A\varphi$ defined by the formula

$$\psi(x) = y_0 + \int_{x_0}^{x} f(t, \varphi(t))\, dt,$$

where $|x - x_0| \leq d$. This is a contraction mapping of the complete space C^* into itself. In fact, let $\varphi \in C^*$, $|x - x_0| \leq d$. Then

$$|\psi(x) - y_0| = \left| \int_{x_0}^{x} f(t, \varphi(t))\, dt \right| \leq kd$$

and consequently $A(C^*) \subseteq C^*$. Moreover, we have

$$|\psi_1(x) - \psi_2(x)| \leq \int_{x_0}^{x} |f(t, \varphi_1(t)) - f(t, \varphi_2(t))|\, dt$$
$$\leq Md \max_x |\varphi_1(x) - \varphi_2(x)|.$$

Since $Md < 1$, the mapping A is a contraction.

From this it follows that the operator equation $\varphi = A\varphi$ (and consequently equation (3) also) has one and only one solution.

II. Let

$$\varphi_i'(x) = f_i(x_0, \varphi_1(x), \cdots, \varphi_n(x)); \qquad i = 1, 2, \cdots, n; \tag{4}$$

be a given system of differential equations with initial conditions

$$\varphi_i(x_0) = y_{0i}; \qquad i = 1, 2, \cdots, n; \tag{5}$$

where the functions $f_i(x, y_1, \cdots, y_n)$ are defined and continuous in some region G of the space R^{n+1} such that G contains the point $(x_0, y_{01}, \cdots, y_{0n})$ and satisfy a Lipschitz condition

$$|f_i(x, y_1^{(1)}, \cdots, y_n^{(1)}) - f_i(x, y_1^{(2)}, \cdots, y_n^{(2)})| \leq M \max\{|y_i^{(1)} - y_i^{(2)}|; 1 \leq i \leq n\}.$$

We shall prove that then on some closed interval $|x - x_0| < d$ there exists one and only one system of solutions $y_i = \varphi_i(x)$ satisfying system (4) and the initial conditions (5).

System (4) together with the initial conditions (5) is equivalent to the system of integral equations

$$\varphi_i(x) = y_{0i} + \int_{x_0}^{x} f_i(t, \varphi_1(t), \cdots, \varphi_n(t))\, dt; \quad i = 1, \cdots, n. \tag{6}$$

Since the function f_i is continuous in some region $G' \subseteq G$ containing the point $(x_0, y_{01}, \cdots, y_{0n})$, the inequalities $|f_i(x, y_1, \cdots, y_n)| \leq K$, where K is a constant, are fulfilled.

We now choose $d > 0$ to satisfy the following conditions:

1) $(x, y_1, \cdots, y_n) \in G$ if $|x - x_0| \leq d$, $|y_i - y_{0i}| \leq Kd$,

2) $Md < 1$.

We now consider the space C_n^* whose elements are ordered systems $\bar{\varphi} = (\varphi_1(x), \cdots, \varphi_n(x))$ consisting of n functions which are defined and continuous for all x for which $|x - x_0| \leq d$ and such that $|\varphi_i(x) - y_{0i}| \leq Kd$, with the metric

$$\rho(\bar{\varphi}, \bar{\psi}) = \max_{x,\, i} |\varphi_i(x) - \psi_i(x)|.$$

The mapping $\bar{\psi} = A\bar{\varphi}$, given by the system of integral equations

$$\psi_i(x) = y_{0i} + \int_{x_0}^{x} f(t, \varphi_1(t), \cdots, \varphi_n(t))\, dt,$$

is a contraction mapping of the complete space C_n^* into itself. In fact,

$$\psi_i^{(1)}(x) - \psi_i^{(2)}(x) = \int_{x_0}^{x} [f(t, \varphi_1^{(1)}, \cdots, \varphi_n^{(1)}) - f(t, \varphi_1^{(2)}, \cdots, \varphi_n^{(2)})]\, dt$$

and consequently

$$\max_{x,\, i} |\psi_i^{(1)}(x) - \psi_i^{(2)}(x)| \leq Md \max_{x,\, i} |\varphi_i^{(1)}(x) - \varphi_i^{(2)}(x)|.$$

Since $Md < 1$, A is a contraction mapping.

It follows that the operator equation $\bar{\varphi} = A\bar{\varphi}$ has one and only one solution.

III. We shall now apply the method of contraction mappings to the proof of the existence and uniqueness of the solution of the Fredholm nonhomogeneous linear integral equation of the second kind:

$$f(x) = \lambda \int_a^b K(x, y) f(y)\, dy + \varphi(x), \tag{7}$$

where $K(x, y)$ (the so-called *kernel*) and $\varphi(x)$ are given functions, $f(x)$ is the function sought, and λ is an arbitrary parameter.

We shall see that our method is applicable only in the case of sufficiently small values of the parameter λ.

We shall assume that $K(x, y)$ and $\varphi(x)$ are continuous for $a \leq x \leq b$; $a \leq y \leq b$ and consequently that $|K(x, y)| \leq M$. Consider the mapping $g = Af$, i.e. $g(x) = \lambda \int_a^b K(x, y) f(y)\, dy + \varphi(x)$, of the complete space $C[a, b]$ into itself. We obtain

$$\rho(g_1, g_2) = \max |g_1(x) - g_2(x)| \leq |\lambda| M(b - a) \max |f_1 - f_2|.$$

Consequently, the mapping A is a contraction for $|\lambda| < 1/M(b - a)$.

From this, on the basis of the principle of contraction mappings, we can conclude that the Fredholm equation has a unique continuous solution for every $|\lambda| < 1/M(b - a)$. The successive approximations to this solution: $f_0(x), f_1(x), \cdots, f_n(x), \cdots$ have the form

$$f_n(x) = \lambda \int_a^b K(x, y) f_{n-1}(y)\, dy + \varphi(x).$$

IV. This method is applicable also in the case of nonlinear equations of the form

$$f(x) = \lambda \int_a^b K(x, y, f(y))\, dy + \varphi(x), \tag{8}$$

where K and φ are continuous. Furthermore K satisfies the condition

$$|K(x, y, z_1) - K(x, y, z_2)| \leq M |z_1 - z_2|$$

for $|\lambda| < 1/M(b - a)$ since here again for the mapping $g = Af$ of the complete space $C[a, b]$ into itself given by the formula

$$g(x) = \lambda \int_a^b K(x, y, f(y))\, dy + \varphi(x)$$

the inequality

$$\max | g_1(x) - g_2(x) | \leq | \lambda | M (b - a) \max | f_1 - f_2 |$$

holds.

V. Consider the Volterra type integral equation

$$f(x) = \lambda \int_a^x K(x, y) f(y)\, dy + \varphi(x) \tag{9}$$

which differs from an equation of Fredholm type in that the upper limit in the integral is the variable quantity x. This equation can be considered as a particular case of the Fredholm equation if we complete the definition of the function $K(x, y)$ for $y > x$ by means of the equation $K(x, y) = 0$ (for $y > x$).

In contrast to the Fredholm integral equation for which we were required to limit ourselves to small values of the parameter λ the principle of contraction mappings (and the method of successive approximations based on it) is applicable to Volterra equations for all values of the parameter λ. We note first of all that the principle of contraction mappings can be generalized in the following manner: *if A is a continuous mapping of a complete metric space R into itself such that the mapping A^n is a contraction for some n, then the equation*

$$Ax = x$$

has one and only one solution.

In fact, if we take an arbitrary point $x \in R$ and consider the sequence $A^{kn}x$ ($k = 0, 1, 2, \cdots$), a repetition of the argument introduced in §14 yields the convergence of this sequence. Let $x_0 = \lim_{k\to\infty} A^{kn}x$. Then $Ax_0 = x_0$. In fact, $Ax_0 = \lim_{k\to\infty} A^{kn}Ax$. Since the mapping A^n is a contraction, we have

$$\rho(A^{kn}Ax, A^{kn}x) \leq \alpha\rho(A^{(k-1)n}Ax, A^{(k-1)n}x) \leq \cdots \leq \alpha^k\rho(Ax, x).$$

Consequently,

$$\lim_{k\to\infty} \rho(A^{kn}Ax, A^{kn}x) = 0,$$

i.e. $Ax_0 = x_0$.

Now consider the mapping

$$g(x) = \lambda \int_a^x K(x, y) f(y)\, dy + \varphi(x) = Af.$$

If f_1 and f_2 are two continuous functions defined on the closed interval $[a, b]$, then

$$| Af_1(x) - Af_2(x) | = \left| \lambda \int_a^x K(x, y)[f_1(y) - f_2(y)]\, dy \right| \leq \lambda Mm(x - a),$$

$$(M = \max | K(x, y)|, \qquad m = \max (| f_1 - f_2 |)),$$

$$| A^2f_1(x) - A^2f_2(x) | \leq \lambda^2 Mm(x - a)^2/2, \cdots,$$

$$| A^nf_1(x) - A^nf_2(x)| \leq \lambda^n Mm(x - a)^n/n! \leq \lambda^n Mm(b - a)^n/n!$$

For an arbitrary value of λ the number n can be chosen so large that

$$\lambda^n(b - a)^n/n! < 1,$$

i.e. the mapping A^n will be a contraction. Consequently, the Volterra equation (9) has a solution for arbitrary λ and this solution is unique.

§16. Compact sets in metric spaces

A set M in the metric space R is said to be *compact* if every sequence of elements in M contains a subsequence which converges to some $x \in R$.

Thus, for example, by virtue of the Bolzano-Weierstrass theorem every bounded set on the real line is compact. Other examples of compact sets will be given below. It is clear that an arbitrary subset of a compact set is compact.

The concept of total boundedness which we shall now introduce is closely related to the concept of compactness.

Let M be any set in the metric space R and let ϵ be a positive number. The set A in R is said to be an ϵ-net with respect to M if for an arbitrary point $x \in M$ at least one point $a \in A$ can be found such that

$$\rho(a, x) < \epsilon.$$

For example, the lattice points form a $2^{\frac{1}{2}}$-net in the plane. A subset M of R is said to be *totally bounded* if R contains a finite ϵ-net with respect to M for every $\epsilon > 0$. It is clear that a totally bounded set is bounded since if an ϵ-net A can be found for M consisting of a finite number of points, then A is bounded and since the diameter of M does not exceed diameter $A + 2\epsilon$, M is also bounded; as Example 2 below will show, the converse is not true in general.

The following obvious remark is often useful: if the set M is totally bounded, then its closure $[M]$ is totally bounded.

It follows at once from the definition of total boundedness that every totally bounded metric space R with an infinite number of points is separable. In fact, construct a finite $(1/n)$-net in R for every n. Their sum over all n is a denumerable set which is everywhere dense in R.

EXAMPLES. 1. For subsets of Euclidean n-space total boundedness coincides with ordinary boundedness, i.e. with the possibility of enclosing a given set in the interior of some sufficiently large cube. In fact, if such a cube is subdivided into cubicles with diagonal of length $\epsilon/n^{\frac{1}{2}}$, then the ver-

every point x_i in $x_1, \cdots, x_n$ we can find a point y_j in $y_1, \cdots, y_m$ such that

$$\rho[f(x_i), y_j] < \epsilon.$$

Let the function $g(x) \in L$ be chosen so that $g(x_i) = y_j$. Then

$$\rho[f(x), g(x)] \leq \rho[f(x), f(x_i)] + \rho[f(x_i), g(x_i)] + \rho[g(x), g(x_i)] < 2\epsilon$$

if i is chosen so that $x \in \epsilon_i$.

From this it follows that $\rho(f, g) < 2\epsilon$ and thus the compactness of D in M_{XY} and consequently in C_{XY} also is proved.

§19. Real functions in metric spaces

A real function on a space R is a mapping of R into the space R^1 (the real line).

Thus, for example, a mapping of R^n into R^1 is an ordinary real-valued function of n variables.

In the case when the space R itself consists of functions, the functions of the elements of R are usually called *functionals*. We introduce several examples of functionals of functions $f(x)$ defined on the closed interval $[0, 1]$:

$$F_1(f) = \sup f(x);$$

$$F_2(f) = \inf f(x);$$

$$F_3(f) = f(x_0) \quad \text{where} \quad x_0 \in [0, 1];$$

$$F_4(f) = \varphi[f(x_0), f(x_1), \cdots, f(x_n)] \quad \text{where} \quad x_i \in [0, 1]$$

and the function $\varphi(y_1, \cdots, y_n)$ is defined for all real y_i;

$$F_5(f) = \int_0^1 \varphi[x, f(x)]\, dx,$$

where $\varphi(x, y)$ is defined and continuous for all $0 \leq x \leq 1$ and all real y;

$$F_6(f) = f'(x_0);$$

$$F_7(f) = \int_0^1 [1 + f'^2(x)]^{\frac{1}{2}}\, dx;$$

$$F_8(f) = \int_0^1 | f'(x) |\, dx.$$

Functionals can be defined on all of R or on a subset of R. For example, in the space C the functionals F_1, F_2, F_3, F_4, F_5 are defined on the entire space, $F_6(f)$ is defined only for functions which are differentiable at the point x_0, $F_7(f)$ for functions for which $[1 + f'^2(x)]^{\frac{1}{2}}$ is integrable, and $F_8(f)$ for functions for which $| f'(x) |$ is integrable.

The definition of continuity for real functions and functionals remains the same as for mappings in general (see §12).

For example, $F_1(f)$ is a continuous functional in C since

$$\rho(f, g) = \sup | f - g | \quad \text{and} \quad | \sup f - \sup g | \leq \sup | f - g |;$$

F_2, F_3, F_5 are also continuous functionals in C; F_4 is continuous in the space C if the function φ is continuous for all arguments; F_6 is discontinuous at every point in the space C for which it is defined. In fact, let $g(x)$ be such that $g'(x_0) = 1, | g(x) | < \epsilon$ and $f = f_0 + g$. Then $f'(x_0) = f_0'(x_0) + 1$ and $\rho(f, f_0) < \epsilon$. This same functional is continuous in the space $C^{(1)}$ of functions having a continuous derivative with the metric

$$\rho(f, g) = \sup [| f - g | + | f' - g' |];$$

F_7 is also a discontinuous functional in the space C. In fact, let $f_0(x) \equiv 0$ and $f_n(x) = (1/n) \sin 2\pi n x$. Then $\rho(f_n, f_0) = 1/n \to 0$. However, $F_7(f_n)$ is a constant (it does not depend on n) which is greater than $(17)^{\frac{1}{2}}$ and $F_7(f_0) = 1$.

Consequently, $F_7(f)$ is discontinuous at the point f_0.

By virtue of this same example $F_8(f)$ is also discontinuous in the space C. Both functionals F_7 and F_8 are continuous in the space $C^{(1)}$.

The following theorems which are the generalizations of well-known theorems of elementary analysis are valid for real functions defined on compacta.

THEOREM 1. *A continuous real function defined on a compactum is uniformly continuous.*

Proof. Assume f is continuous but not uniformly continuous, i.e. assume there exist x_n and x_n' such that

$$| x_n - x_n' | < 1/n \quad \text{and} \quad | f(x_n) - f(x_n') | \geq \epsilon.$$

From the sequence $\{x_n\}$ we can choose a subsequence $\{x_{n_k}\}$ which converges to x. Then also $\{x_{n_k}'\} \to x$ and either $| f(x) - f(x_n') | \geq \epsilon/2$ or $| f(x) - f(x_n) | \geq \epsilon/2$, which contradicts the continuity of $f(x)$.

THEOREM 2. *If the function $f(x)$ is continuous on the compactum K, then f is bounded on K.*

Proof. If f were not bounded on K, then there would exist a sequence $\{x_n\}$ such that $f(x_n) \to \infty$. We choose from $\{x_n\}$ a subsequence which converges to x: $\{x_{n_k}\} \to x$. Then in an arbitrarily small neighborhood of x the function $f(x)$ will assume arbitrarily large values which contradicts the continuity of f.

THEOREM 3. *A function f which is continuous on a compactum K attains its least upper and greatest lower bounds on K.*

Proof. Let $A = \sup f(x)$. Then there exists a sequence $\{x_n\}$ such that

$$A > f(x_n) > A - 1/n.$$

We choose a convergent subsequence from $\{x_n\}$: $\{x_{n_k}\} \to x$. By continuity, $f(x) = A$. The proof for $\inf f(x)$ is entirely analogous.

Theorems 2 and 3 allow generalizations to an even more extensive class of functions (the so-called *semicontinuous functions*).

A function $f(x)$ is said to be *lower* (*upper*) *semicontinuous* at the point x_0 if for arbitrary $\epsilon > 0$ there exists a δ-neighborhood of x_0 in which $f(x) > f(x_0) - \epsilon$, $(f(x) < f(x_0) + \epsilon)$.

For example, the function "integral part of x", $f(x) = E(x)$, is upper semicontinuous. If we increase (decrease) the value of $f(x_0)$ of a continuous function at a single point x_0, we obtain a function which is upper (lower) semicontinuous. If $f(x)$ is upper semicontinuous, then $-f(x)$ is lower semicontinuous. These two remarks at once permit us to construct a large number of examples of semicontinuous functions.

We shall also consider functions which assume the values $\pm\infty$. If $f(x_0) = -\infty$, then $f(x)$ will be assumed to be lower semicontinuous at x_0 and upper semicontinuous at x_0 if for arbitrary $h > 0$ there is a neighborhood of the point x_0 in which $f(x) < -h$.

If $f(x_0) = +\infty$, then $f(x)$ will be assumed to be upper semicontinuous at x_0 and lower semicontinuous at x_0 if for arbitrary $h > 0$ there is a neighborhood of the point x_0 in which $f(x) > h$.

The *upper limit* $\bar{f}(x_0)$ of the function $f(x)$ at the point x_0 is the $\lim_{\epsilon\to 0}\{\sup [f(x);\ x \in S(x_0, \epsilon)]\}$. The *lower limit* $\underline{f}(x_0)$ is the $\lim_{\epsilon\to 0}\{\inf [f(x);\ x \in S(x_0, \epsilon)]\}$. The difference $\omega f(x_0) = \bar{f}(x_0) - \underline{f}(x_0)$ is the *oscillation* of the function $f(x)$ at the point x_0. It is easy to see that a necessary and sufficient condition that the function $f(x)$ be continuous at the point x_0 is that $\omega f(x_0) = 0$, i.e. that $\bar{f}(x_0) = \underline{f}(x_0)$.

For arbitrary $f(x)$ the function $\bar{f}(x)$ is upper semicontinuous and the function $\underline{f}(x)$ is lower semicontinuous. This follows easily from the definition of the upper and lower limits.

We now consider several important examples of semicontinuous functionals.

Let $f(x)$ be a real function of a real variable. For arbitrary real a and b such that $f(x)$ is defined on the closed interval $[a, b]$ we define the *total variation* of the function $f(x)$ on $[a, b]$ to be the functional

$$V_a^b(f) = \sup \sum_{i=1}^n |f(x_i) - f(x_{i-1})|$$

where $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ and the least upper bound is taken over all possible subdivisions of the closed interval $[a, b]$.

For a monotone function $V_a^b(f) = |f(b) - f(a)|$. For a piecewise monotone function $V_a^b(f)$ is the sum of the absolute values of the increments on the segments of monotonicity. For such functions

$$\sup \sum_{i=1}^n |f(x_i) - f(x_{i-1})|$$

is attained for some subdivision.

We shall prove that the *functional* $V_a^b(f)$ *is lower semicontinuous in the space* M *of all bounded functions of a real variable with metric* $\rho(f, g) = \sup |f(x) - g(x)|$ (it is clear that C is a subspace of the space M), i.e. that for arbitrary f and $\epsilon > 0$ there exists a δ such that $V_a^b(g) > V_a^b(f) - \epsilon$ for $\rho(f, g) < \delta$.

We choose a subdivision of the closed interval $[a, b]$ such that

$$\sum_{i=1}^n |f(x_i) - f(x_{i-1})| > V_a^b(f) - \epsilon/2.$$

Let $\delta = \epsilon/4n$. Then if $\rho(g, f) < \delta$, we have

$$\sum_{i=1}^n |f(x_i) - f(x_{i-1})| - \sum_{i=1}^n |g(x_i) - g(x_{i-1})| < \epsilon/2$$

and consequently

$$V_a^b(g) \geq \sum_{i=1}^n |g(x_i) - g(x_{i-1})| > V_a^b(f) - \epsilon.$$

In the case $V_a^b(f) = \infty$ the theorem remains valid since then for arbitrary H there exists a subdivision of the closed interval $[a, b]$ such that

$$\sum_{i=1}^n |f(x_i) - f(x_{i-1})| > H$$

and δ can be chosen such that

$$\sum_{i=1}^n |f(x_i) - f(x_{i-1})| - \sum_{i=1}^n |g(x_i) - g(x_{i-1})| < \epsilon.$$

Then $V_a^b(g) > H - \epsilon$, i.e. $V_a^b(g) \geq H$, so that $V_a^b(g) = \infty$.

The functional $V_a^b(f)$ is not continuous as is easily seen from the following example. Let $f(x) \equiv 0$, $g_n(x) = (1/n) \sin nx$. Then $\rho(g_n, f) = 1/n$, but $V_0^\pi(g_n) = 2$ and $V_0^\pi(f) = 0$.

Functions for which $V_a^b(f) < \infty$ are said to be *functions of bounded* (or finite) *variation*. The reader can find more information about the properties of such functions in the books by Aleksandrov and Kolmogorov: *Introduction to the Theory of Functions of a Real Variable*, Chapter 7, §7; Natanson: *Theory of Functions of a Real Variable*, Chapter 8; and Jeffery: *The Theory of Functions of a Real Variable*, Chapter 5.

We shall define the length of the curve $y = f(x)$ $(a \leq x \leq b)$ as the functional

$$L_a^b(f) = \sup \sum_{i=1}^n \{(x_i - x_{i-1})^2 + [f(x_i) - f(x_{i-1})]^2\}^{\frac{1}{2}},$$

where the least upper bound is taken over all possible subdivisions of the closed interval $[a, b]$. This functional is defined on the entire space M. For continuous functions it coincides with the value of the limit

$$\lim \sum_{i=1}^n \{(x_i - x_{i-1})^2 + [f(x_i) - f(x_{i-1})]^2\}^{\frac{1}{2}} \text{ as } \max_i |x_i - x_{i-1}| \to 0.$$

Finally, for functions with continuous derivative it can be written in the form

$$\int_a^b [1 + f'^2(x)]^{\frac{1}{2}}\, dx.$$

The functional $L_a^b(f)$ is lower semicontinuous in M. This is proved exactly as in the case of the functional $V_a^b(f)$.

Theorems 2 and 3 established above generalize to semicontinuous functions.

THEOREM 2a. *A finite function which is lower (upper) semicontinuous on a compactum K is bounded below (above) on K.*

In fact, let f be finite and lower semicontinuous and let $\inf f(x) = -\infty$. Then there exists a sequence $\{x_n\}$ such that $f(x_n) < -n$. We choose a subsequence $\{x_{n_k}\} \to x_0$. Then, by virtue of the lower semicontinuity of f, $f(x_0) = -\infty$, which contradicts the assumption that $f(x)$ is finite.

In the case of an upper semicontinuous function the theorem is proved analogously.

THEOREM 3a. *A finite lower (upper) semicontinuous function defined on a compactum K attains its greatest lower (least upper) bound on K.*

Assume the function f is lower semicontinuous. Then by Theorem 2 it has a finite greatest lower bound and there exists a sequence $\{x_n\}$ such that $f(x_n) \leq \inf f(x) + 1/n$. We choose a subsequence $\{x_{n_k}\} \to x_0$. Then $f(x_0) = \inf f(x)$ since the supposition that $f(x_0) > \inf f(x)$ contradicts the lower semicontinuity of f.

The theorem is proved analogously for the case of an upper semicontinuous function.

Let K be a compact metric space and let C_K be the space of continuous real functions defined on K with distance function $\rho(f, g) = \sup |f - g|$. Then the following theorem is valid.

THEOREM 4. *A necessary and sufficient condition that the set $D \subseteq C_K$ be compact is that the functions belonging to D be uniformly bounded and equicontinuous* (Arzelà's theorem for continuous functions defined on an arbitrary compactum).

The sufficiency follows from the general Theorem 7, §18. The necessity is proved exactly as in the proof of Arzelà's theorem (see §17).

§20. Continuous curves in metric spaces

Let $P = f(t)$ be a given continuous mapping of the closed interval $a \leq t \leq b$ into a metric space R. When t runs through the segment from a to b, the corresponding image point P runs through some "continuous curve" in the space R. We propose to give rigorous definitions connected with the above ideas which were stated rather crudely just now. We shall

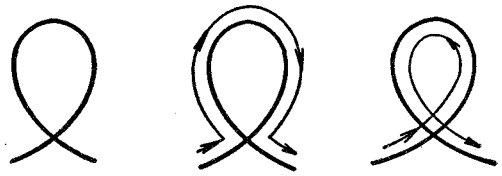

FIG. 11 FIG. 12 FIG. 13

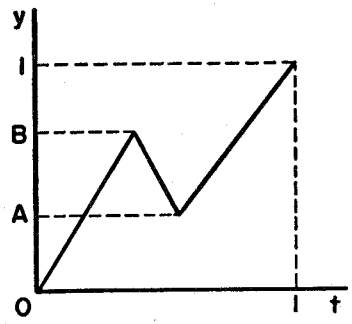

FIG. 14

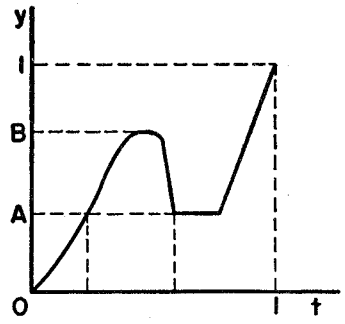

FIG. 15

consider the order in which the point traverses the curve an essential property of the curve itself. The set shown in Fig. 11, traversed in the directions indicated in Figs. 12 and 13, will be considered as distinct curves. As another example let us consider the real function defined on the closed interval [0, 1] which is shown in Fig. 14. It defines a "curve" situated on the segment [0, 1] of the y-axis, distinct from this segment, traversed once from the point 0 to the point 1, since the segment $[A, B]$ is traversed three times (twice upward and once downward).

However, for the same order of traversing the points of the space we shall consider the choice of the "parameter" t unessential. For example, the functions given in Figs. 14 and 15 define the same "curve" over the y-axis although the values of the parameter t corresponding to an arbitrary point of the curve can be distinct in Figs. 14 and 15. For example, in Fig. 14 to the point A there correspond two isolated points on the t-axis, whereas in Fig. 15 to the point A there correspond on the t-axis one isolated point and the segment lying to the right (when t traverses this segment the point A on the curve remains fixed). [Allowing such intervals of constancy of the point $P = f(t)$ is further convenient in the proof of the compactness of systems of curves.]

We pass over to formal definitions. Two continuous functions

$$P = f_1(t') \quad \text{and} \quad P = f_2(t'')$$

defined, respectively, on the closed intervals

$$a' \leq t' \leq b' \quad \text{and} \quad a'' \leq t'' \leq b''$$

are said to be equivalent if there exist two continuous nondecreasing functions

$$t' = \varphi_1(t) \quad \text{and} \quad t'' = \varphi_2(t)$$

defined on a closed interval $a \leq t \leq b$ and possessing the properties

$$\varphi_1(a) = a', \qquad \varphi_1(b) = b',$$
$$\varphi_2(a) = a'', \qquad \varphi_2(b) = b'',$$
$$f_1[\varphi_1(t)] = f_2[\varphi_2(t)]$$

for all $t \in [a, b]$.

It is easy to see that the equivalence property is reflexive (f is equivalent to f), symmetric (if f_1 is equivalent to f_2, then f_2 is equivalent to f_1), and transitive (the equivalence of f_1 and f_2 together with the equivalence of f_2 and f_3 implies the equivalence of f_1 and f_3). Therefore all continuous functions of the type considered are partitioned into classes of equivalent functions. Every such class also defines a continuous curve in the space R.

It is easy to see that for an arbitrary function $P = f_1(t')$ defined on a closed interval $[a', b']$ we can find a function which is equivalent to it and which is defined on the closed interval $[a'', b''] = [0, 1]$. To this end, it is sufficient to set

$$t' = \varphi_1(t) = (b' - a')t + a', \qquad t'' = \varphi_2(t) = t.$$

(We always assume that $a < b$. However we do not exclude "curves" consisting of a single solitary point which is obtained when the function $f(t)$ is constant on $[a, b]$. This assumption is also convenient in the sequel.) Thus, we can assume that all curves are given parametrically by means of functions defined on the closed interval [0, 1].

Therefore it is expedient to consider the space C_{IR} of continuous mappings of the closed interval $I = [0, 1]$ into the space R with the metric $\rho(f, g) = \sup_t \rho[f(t), g(t)]$.

We shall assume that the sequence of curves $L_1, L_2, \ldots, L_n, \ldots$ converges to the curve L if the curves L_n can be represented parametrically in the form

$$P = f_n(t), \qquad 0 \leq t \leq 1,$$

and the curve L in the form

$$P = f(t), \qquad 0 \leq t \leq 1,$$

so that $\rho(f, f_n) \to 0$ as $n \to \infty$.

We obtain Theorem 1 if we apply the generalized Arzelà theorem (Theorem 7, §18) to the space C_{IR}.

THEOREM 1. *If the sequence of curves $L_1, L_2, \cdots, L_n, \cdots$ lying in the compactum K can be represented parametrically by means of equicontinuous functions defined on the closed interval* [0, 1], *then this sequence contains a convergent subsequence.*

We shall now define the length of a curve given parametrically by means of the function $P = f(t)$, $a \leq t \leq b$, as the least upper bound of sums of the form

$$\sum_{i=1}^{n} \rho[f(t_{i-1}), f(t_i)],$$

where the points t_i are subject only to the following conditions:

$$a \leq t_0 \leq t_1 \leq \cdots \leq t_i \leq \cdots \leq t_n = b.$$

It is easy to see that the length of a curve does not depend on the choice of its parametric representation. If we limit ourselves to parametric representations by functions defined on the closed interval [0, 1], then it is easy to prove by considerations similar to those of the preceding section that the length of a curve is a lower semicontinuous functional of f (in the space C_{IR}). In geometric language this result can be expressed in the form of such a theorem on semicontinuity.

THEOREM 2. *If the sequence of curves L_n converges to the curve L, then the length of L is not greater than the greatest lower bound of the lengths of the curves L_n.*

We shall now consider specially *curves of finite length* or *rectifiable curves*. Let the curve be defined parametrically by means of the function $P = f(t)$, $a \leq t \leq b$. The function f, considered only on the closed interval $[a, T]$, where $a \leq T \leq b$, defines an "initial segment" of the curve from the point $P_a = f(a)$ to the point $P_T = f(T)$. Let $s = \varphi(T)$ be its length. It is easily established that

$$P = g(s) = f[\varphi^{-1}(s)]$$

is a new parametric representation of the same curve. In this connection s runs through the closed interval $0 \leq s \leq S$, where S is the length of the entire curve under consideration. This representation satisfies the requirement

$$\rho[g(s_1), g(s_2)] \leq |s_2 - s_1|$$

(the length of the curve is not less than the length of the chord).

Going over to the closed interval [0, 1] we obtain the parametric representation

$$P = F(\tau) = g(s), \qquad \tau = s/S$$

which satisfies the following Lipschitz condition:

$$\rho[F(\tau_1), F(\tau_2)] \leq S \mid \tau_1 - \tau_2 \mid.$$

We thus see that for all curves of length S such that $S \leq M$, where M is a constant, a parametric representation on the closed interval [0, 1] by means of equicontinuous functions is possible. Consequently, Theorem 1 is applicable to such curves.

We shall show the power of the general results obtained above by applying them to the proof of the following important proposition.

THEOREM 3. *If two points A and B in the compactum K can be connected by a continuous curve of finite length, then among all such curves there exists one of minimal length.*

In fact, let Y be the greatest lower bound of the lengths of curves which connect A and B in the compactum K. Let the lengths of the curves L_1, L_2, $\cdots$, L_n, $\cdots$ connecting A with B tend to Y. From the sequence L_n it is possible, by Theorem 1, to select a convergent subsequence. By Theorem 2 the limit curve of this subsequence cannot have length greater than Y.

We note that even when K is a closed smooth (i.e. differentiable a sufficient number of times) surface in three-dimensional Euclidean space, this theorem does not follow directly from the results established in usual differential geometry courses where we restrict ourselves ordinarily to the case of sufficiently proximate points A and B.

All the arguments above would take on great clarity if we formed of the set of all curves of a given metric space R a metric space. This can be done by introducing the distance between two curves L_1 and L_2 by means of the formula

$$\rho(L_1, L_2) = \inf \rho(f_1, f_2),$$

where the greatest lower bound is taken over all possible pairs of parametric representations

$$P = f_1(t), \qquad P = f_2(t) \qquad (0 \leq t \leq 1)$$

of the curves L_1 and L_2, respectively.

The proof of the fact that this distance satisfies the axioms of a metric space is very straightforward with the exception of one point: there is some difficulty in proving that $\rho(L_1, L_2) = 0$ implies that the curves L_1 and L_2 are identical. This fact is an immediate consequence of the fact that the greatest lower bound in the formula which we used in the definition of the distance $\rho(L_1, L_2)$ is attained for a suitable choice of the parametric representations f_1 and f_2. But the proof of this last assertion is also not very straightforward.

Chapter III

NORMED LINEAR SPACES

§21. Definition and examples of normed linear spaces

DEFINITION 1. A set R of elements $x, y, z, \cdots$ is said to be a *linear space* if the following conditions are satisfied:

I. For any two elements $x, y \in R$ there is uniquely defined a third element $z = x + y$, called their sum, such that

1) $x + y = y + x$,

2) $x + (y + z) = (x + y) + z$,

3) there exists an element 0 having the property that $x + 0 = x$ for all $x \in R$, and

4) for every $x \in R$ there exists an element $-x$ such that $x + (-x) = 0$.

II. For an arbitrary number α and element $x \in R$ there is defined an element αx (the product of the element x and the number α) such that

1) $\alpha(\beta x) = (\alpha\beta)x$, and

2) $1 \cdot x = x$.

III. The operations of addition and multiplication are related in the following way:

1) $(\alpha + \beta)x = \alpha x + \beta x$, and

2) $\alpha(x + y) = \alpha x + \alpha y$.

Depending on the numbers admitted (all complex numbers or only the reals), we distinguish between complex and real linear spaces. Unless otherwise stated we shall consider real linear spaces. In a linear space, besides the operations of addition and multiplication by scalars, usually there is introduced in one way or another the operation of passage to the limit. It is most convenient to do this by introducing a norm into the linear space.

A linear space R is said to be *normed* if to each element $x \in R$ there is made to correspond a nonnegative number $\| x \|$ which is called the norm of x and such that:

1) $\| x \| = 0$ if, and only if, $x = 0$,

2) $\| \alpha x \| = | \alpha | \, \| x \|$,

3) $\| x + y \| \leq \| x \| + \| y \|$.

It is easy to see that every normed space is also a metric space; it is sufficient to set $\rho(x, y) = \| x - y \|$. The validity of the metric space axioms follows directly from Properties 1–3 of the norm.

A complete normed space is said to be a *Banach space*, a *space of Banach type*, or, more briefly, a B-space.

EXAMPLES OF NORMED SPACES. 1. The real line with the usual arithmetic definitions is the simplest example of a normed space. In this case the norm is simply the absolute value of the real number.

2. Euclidean n-space, i. e. the space consisting of all n-tuples of real numbers: $x = (x_1, x_2, \cdots, x_n)$ in which the norm (i. e. the length) of the vector is defined to be the square root of its scalar square,

$$\| x \| = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}},$$

is also a normed linear space.

In an n-dimensional linear space the norm of the vector $x = (x_1, x_2, \cdots, x_n)$ can also be defined by means of the formula

$$\| x \| = (\sum_{k=1}^{n} | x_k |^p)^{1/p}, \qquad (p \geq 1).$$

We also obtain a normed space if we set the norm of the vector $x = (x_1, x_2, \cdots, x_n)$ equal to the max $\{|x_k|; 1 \leq k \leq n\}$.

3. The space $C[a, b]$ of continuous functions with the operations of addition and multiplication by a scalar which are usual for functions, in which

$$\| f(t) \| = \max\{| f(t) |; a \leq t \leq b\},$$

is a normed linear space.

4. Let $C^2[a, b]$ consist of all functions continuous on $[a, b]$ and let the norm be given by the formula

$$\| f(t) \| = \left(\int_a^b f^2(t)\, dt \right)^{\frac{1}{2}}.$$

All the norm axioms are satisfied.

5. The space l_2 is a normed linear space if we define the sum of two elements $x = (\xi_1, \xi_2, \cdots, \xi_n, \cdots)$ and $y = (\eta_1, \eta_2, \cdots, \eta_n, \cdots)$ in l_2 to be

$$x + y = (\xi_1 + \eta_1, \xi_2 + \eta_2, \cdots, \xi_n + \eta_n, \cdots),$$

and let

$$\alpha x = (\alpha\xi_1, \alpha\xi_2, \cdots, \alpha\xi_n, \cdots),$$

and

$$\| x \| = (\sum_{n=1}^{\infty} | \xi_n |^2)^{\frac{1}{2}}.$$

6. The space c consisting of all sequences $x = (x_1, x_2, \cdots, x_n, \cdots)$ of real numbers which satisfy the condition $\lim_{n\to\infty} x_n = 0$.

Addition and multiplication are defined as in Example 5 and the norm is set equal to

$$\| x \| = \max\{ | x_n |; 1 \leq n \leq \infty \}.$$

7. The space m of bounded sequences with the same definitions of sum, product, and norm as in the preceding example.

In each of these examples the linear space axioms are verified without difficulty. The fact that the norm Axioms 1–3 are fulfilled in Examples 1–5 is proved exactly as was the validity of the metric space axioms in the corresponding examples in §8, Chapter II.

All the spaces enumerated in the examples, except the space $C^2[a, b]$, are Banach spaces.

DEFINITION 2. A *linear manifold* L in a normed linear space R is any set of elements in R satisfying the following condition: if $x, y \in L$, then $\alpha x + \beta y \in L$, where α and β are arbitrary numbers. A *subspace* of the space R is a closed linear manifold in R.

REMARK 1. In Euclidean n-space R^n the concepts of linear manifold and subspace coincide because every linear manifold in R^n is automatically closed. (Prove this!) On the other hand, linear manifolds which are not closed exist in an infinite-dimensional space. For example, in l_2 the set L of points of the form

$$x = (x_1, x_2, \cdots, x_k, 0, 0, \cdots), \tag{1}$$

i.e. of points which have only a finite (but arbitrary) number of nonzero coordinates, forms a linear manifold which is not closed. In fact, a linear combination of points of form (1) is a point of the same form, i.e. L is a linear manifold. But L is not closed since, for instance, the sequence of points

$$(1, 0, 0, 0, \cdots),$$
$$(1, \tfrac{1}{2}, 0, 0, \cdots),$$
$$(1, \tfrac{1}{2}, \tfrac{1}{4}, 0, \cdots),$$
$$\cdots\cdots\cdots\cdots,$$

belonging to L, converges to the point $(1, \frac{1}{2}, \frac{1}{4}, \cdots, 1/2^n, \cdots)$, which does not belong to L.

REMARK 2. Let $x_1, x_2, \cdots, x_n, \cdots$ be elements of a Banach space R and let M be the totality of elements in R which are of the form $\sum_{i=1}^{n} c_i x_i$ for arbitrary finite n. It is obvious that M is a linear manifold in R. We shall show that $[M]$ is a linear subspace. In view of the fact that $[M]$ is closed it is sufficient to prove that it is a linear manifold.

Let $x \in [M]$, $y \in [M]$. Then in an arbitrary ϵ-neighborhood of x we can find an $x_\epsilon \in M$ and in an arbitrary ϵ-neighborhood of y we can find a $y_\epsilon \in M$. We form the element $\alpha x + \beta y$ and estimate $\| \alpha x + \beta y - \alpha x_\epsilon - \beta y_\epsilon \|$:

$$\| \alpha x + \beta y - \alpha x_\epsilon - \beta y_\epsilon \| \leq | \alpha | \, \| x - x_\epsilon \| + | \beta | \, \| y - y_\epsilon \| \leq (| \alpha | + | \beta |)\epsilon,$$

from which it is clear that $\alpha x + \beta y \in [M]$.

The subspace $L = [M]$ is said to be the *subspace generated by the elements* $x_1, x_2, \cdots, x_n, \cdots$.

§22. Convex sets in normed linear spaces

Let x and y be two points in the linear space R. Then the *segment* connecting the points x and y is the totality of all points of the form $\alpha x + \beta y$, where $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta = 1$.

DEFINITION. A set M in the linear space R is said to be *convex* if, given two arbitrary points x and y belonging to M, the segment connecting them also belongs to M. A convex set is called a *convex body* if it contains at least one interior point, i.e. if it contains some sphere completely.

EXAMPLES. 1. In three-dimensional Euclidean space, the cube, sphere, tetrahedron, and halfspace are convex bodies; but a triangle, plane, and segment are convex sets although they are not convex bodies.

2. A sphere in a normed linear space is always a convex set (and also a convex body). In fact, consider the unit sphere S: $\| x \| \leq 1$.

If x_0, y_0 are two arbitrary points belonging to this sphere: $\| x_0 \| \leq 1$, $\| y_0 \| \leq 1$, then

$$\| \alpha x_0 + \beta y_0 \| \leq \| \alpha x_0 \| + \| \beta y_0 \| = \alpha \| x_0 \| + \beta \| y_0 \| \leq \alpha + \beta = 1,$$

i.e.

$$\alpha x_0 + \beta y_0 \in S \qquad (\alpha \geq 0, \beta \geq 0, \alpha + \beta = 1).$$

3. Let R be the totality of vectors $x = (\xi_1, \xi_2)$ in the plane. Introduce the following distinct norms in R:

$$\| x \|_2 = (\xi_1^2 + \xi_2^2)^{\frac{1}{2}}; \qquad \| x \|_\infty = \max(|\xi_1|, |\xi_2|);$$
$$\| x \|_1 = |\xi_1| + |\xi_2|; \qquad \| x \|_p = (|\xi_1|^p + |\xi_2|^p)^{1/p} \quad (p > 1).$$

Let us see what the unit sphere will be for each of these norms (see Fig. 16).

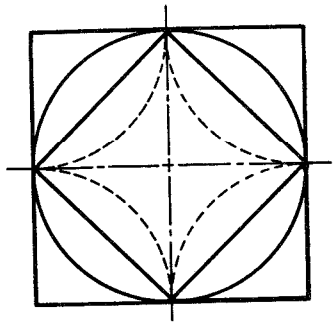

FIG. 16

In the case $\| x \|_2$ it is a circle of radius 1, in the case $\| x \|_\infty$ it is a square with vertices $(\pm 1, \pm 1)$, in the case $\| x \|_1$ it is a square with vertices $(0, 1)$, $(1, 0)$, $(-1, 0)$, $(0, -1)$. If we consider the unit sphere corresponding to the norm $\| x \|_p$, and let p increase from 1 to ∞, then this "sphere" deforms in a continuous manner from the square corresponding to $\| x \|_1$ to the square corresponding to $\| x \|_\infty$. Had we set

$$\| x \|_p = (|\xi_1|^p + |\xi_2|^p)^{1/p} \tag{1}$$

for $p < 1$, then the set $\| x \|_p \leq 1$ would not have been convex (for example, for $p = 2/3$ it would be the interior of an astroid). This is another expression of the fact that for $p < 1$ the "norm" (1) does not satisfy Condition 3 in the definition of a norm.

4. Let us consider a somewhat more complicated example. Let Φ be the set of points $x = (\xi_1, \xi_2, \cdots, \xi_n, \cdots)$ in l_2 which satisfy the condition

$$\sum_{n=1}^{\infty} \xi_n^2 n^2 \leq 1.$$

This is a convex set in l_2 which is not a convex body. In fact, if $x, y \in \Phi$ and $z = \alpha x + \beta y$, where $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$, then by virtue of the Schwarz inequality (Chapter II),

$$\sum_{n=1}^{\infty} n^2(\alpha\xi_n + \beta\eta_n)^2 = \alpha^2 \sum_{n=1}^{\infty} n^2\xi_n^2 + 2\alpha\beta \sum_{n=1}^{\infty} n^2\xi_n\eta_n + \beta^2 \sum_{n=1}^{\infty} n^2\eta_n^2$$
$$\leq \alpha^2 \sum_{n=1}^{\infty} n^2\xi_n^2 + 2\alpha\beta \left(\sum_{n=1}^{\infty} n^2\xi_n^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^{\infty} n^2\eta_n^2\right)^{\frac{1}{2}} + \beta^2 \sum_{n=1}^{\infty} n^2\eta_n^2$$
$$= \left[\alpha\left(\sum n^2\xi_n^2\right)^{\frac{1}{2}} + \beta\left(\sum n^2\eta_n^2\right)^{\frac{1}{2}}\right]^2 \leq (\alpha + \beta)^2 = 1.$$

We shall show that Φ contains no sphere. Φ is symmetric with respect to the origin of coordinates; hence, if Φ contained some sphere S', it would also contain the sphere S'' which is symmetric to S' with respect to the origin. Then Φ, being convex, would contain all segments connecting points of the spheres S' and S'', and consequently it would also contain a sphere S of the same radius as that of S', with the center of S at the origin. But if Φ contained some sphere of radius r with center at the origin, then on every ray emanating from zero there would lie a segment belonging entirely to Φ. However, on the ray defined by the vector $(1, 1/2, 1/3, \cdots, 1/n, \cdots)$ there obviously is no point except zero which belongs to Φ.

EXERCISES. 1. Prove that the set Φ is compact. Prove that no compact convex set in l_2 can be a convex body.

2. Prove that Φ is not contained in any subspace distinct from all of l_2.

3. Prove that the fundamental parallelopiped in l_2 (see Example 3, §16) is a convex set but not a convex body.

We shall now establish the following simple properties of convex sets.

THEOREM 1. *The closure of a convex set is a convex set.*

Proof. Let M be a convex set, $[M]$ its closure and let x, y be two arbitrary

points in $[M]$. Further, let ϵ be an arbitrary positive number. Points a, b can be found in M such that $\rho(a, x) < \epsilon$ and $\rho(b, y) < \epsilon$. Then $\rho(\alpha x + \beta y, \alpha a + \beta b) < \epsilon$ for arbitrary nonnegative α and β such that $\alpha + \beta = 1$, and the point $\alpha a + \beta b$ belongs to M since M is convex. Since $\epsilon > 0$ is arbitrary, it follows that $\alpha x + \beta y \in [M]$, i. e. $[M]$ is convex also.

THEOREM 2. *The intersection of an arbitrary number of convex sets is a convex set.*

Proof. Let $M = \bigcap_\alpha M_\alpha$, where all M_α are convex sets. Further, let x and y be two arbitrary points in M. These points x and y belong to all M_α. Then the segment connecting the points x and y belongs to each M_α and consequently it also belongs to M. Thus, M is in fact convex.

Since the intersection of closed sets is always closed, it follows that *the intersection of an arbitrary number of closed convex sets is a closed convex set.*

Let A be an arbitrary subset of a normed linear space. We define the *convex closure* of the set A to be the smallest closed convex set containing A.

The convex closure of any set can obviously be obtained as the intersection of all closed convex sets which contain the given set.

Consider the following important example of convex closure. Let $x_1, x_2, \cdots, x_{n+1}$ be points in a normed linear space. We shall say that these $n + 1$ points are *in general position* if no three of them lie on one straight line, no four of them lie in one plane, and so forth; in general, no $k + 1$ of these points lie in a subspace of dimension less than k. The convex closure of the points $x_1, x_2, \cdots, x_{n+1}$ which are in general position is called an *n-dimensional simplex* and the points $x_1, x_2, \cdots, x_{n+1}$ themselves are called the *vertices* of the simplex. A zero-dimensional simplex consists of a single point. One-, two-, and three-dimensional simplexes are, respectively, a segment, triangle, tetrahedron.

If the points $x_1, x_2, \cdots, x_{n+1}$ are in general position, then any $k + 1$ of them ($k < n$) also are in general position and consequently they generate a k-dimensional simplex, called a *k-dimensional face* of the given n-dimensional simplex. For example, the tetrahedron with the vertices e_1, e_2, e_3, e_4 has four two-dimensional faces defined respectively by the triples of vertices (e_2, e_3, e_4), (e_1, e_3, e_4), (e_1, e_2, e_4), (e_1, e_2, e_3); six one-dimensional faces; and four zero-dimensional faces.

THEOREM 3. *A simplex with the vertices $x_1, x_2, \cdots, x_{n+1}$ is the totality of all points which can be represented in the form*

$$x = \sum_{k=1}^{n+1} \alpha_k x_k; \qquad \alpha_k \geq 0, \qquad \sum_{k=1}^{n+1} \alpha_k = 1. \tag{2}$$

Proof. In fact, it is easy to verify that the totality of points of the form (2) represents a closed convex set which contains the points $x_1, x_2, \cdots, x_{n+1}$. On the other hand, every convex set which contains the points $x_1, x_2, \cdots, x_{n+1}$ must also contain points of the form (2), and consequently these points form the smallest closed convex set containing the points $x_1, x_2, \cdots, x_{n+1}$.

§23. Linear functionals

DEFINITION 1. A numerical function $f(x)$ defined on a normed linear space R will be called a *functional*. A functional $f(x)$ is said to be *linear* if

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y),$$

where $x, y \in R$ and α, β are arbitrary numbers.

A functional $f(x)$ is said to be *continuous* if for arbitrary $\epsilon > 0$ a $\delta > 0$ can be found such that the inequality

$$| f(x_1) - f(x_2) | < \epsilon$$

holds whenever

$$\| x_1 - x_2 \| < \delta.$$

In the sequel we shall consider only continuous functionals (in particular continuous linear functionals) and for brevity we shall omit the word "continuous".

We shall establish some properties of linear functionals which follow almost directly from the definition.

THEOREM 1. *If the linear functional $f(x)$ is continuous at some point $x_0 \in R$, then it is continuous everywhere in R.*

Proof. In fact, let the linear functional $f(x)$ be continuous at the point $x = x_0$. This is equivalent to the fact that $f(x_n) \to f(x_0)$ when $x_n \to x_0$.

Further, let $y_n \to y$. Then

$$f(y_n) = f(y_n - y + x_0 + y - x_0) = f(y_n - y + x_0) + f(y) - f(x_0).$$

But $y_n - y + x_0 \to x_0$. Consequently, by assumption, $f(y_n - y + x_0) \to f(x_0)$. Thus,

$$f(y_n) \to f(x_0) + f(y) - f(x_0) = f(y).$$

A functional $f(x)$ is said to be *bounded* if there exists a constant N such that

$$| f(x) | < N \| x \| \tag{1}$$

for all $x \in R$.

THEOREM 2. *For linear functionals the conditions of continuity and boundedness are equivalent.*

Proof. We assume that the linear functional $f(x)$ is not bounded. Then for

arbitrary natural number n we can find an element $x_n \in R$ such that $|f(x_n)| > n \|x_n\|$. We shall set $y_n = x_n/(n\|x_n\|)$. Then $\|y_n\| = 1/n$, i.e. $y_n \to 0$. But at the same time

$$|f(y_n)| = |f(x_n/n\|x_n\|)| = (1/n\|x_n\|)|f(x_n)| > 1.$$

Consequently, the functional $f(x)$ is not continuous at the point $x = 0$.

Now let N be a number which satisfies Condition (1). Then for an arbitrary sequence $x_n \to 0$ we have:

$$|f(x_n)| \leq N\|x_n\| \to 0,$$

i.e. $f(x)$ is continuous at the point $x = 0$ and consequently at all the remaining points also. This completes the proof of the theorem.

DEFINITION 2. The quantity

$$\|f\| = \sup\{|f(x)|/\|x\|; x \neq 0\}$$

is called the *norm* of the linear functional $f(x)$.

EXAMPLES OF LINEAR FUNCTIONALS ON VARIOUS SPACES. 1. Let R^n be Euclidean n-space and let a be a fixed nonzero vector in R^n. For arbitrary $x \in R^n$ we set $f(x) = (x, a)$, where (x, a) is the scalar product of the vectors x and a. It is clear that $f(x)$ is a linear functional. In fact,

$$f(\alpha x + \beta y) = (\alpha x + \beta y, a) = \alpha(x, a) + \beta(y, a) = \alpha f(x) + \beta f(y).$$

Further, by virtue of Schwarz's inequality

$$|f(x)| = |(x, a)| \leq \|x\|\,\|a\|. \tag{2}$$

Consequently, the functional $f(x)$ is bounded and is therefore continuous. From (2) we find

$$|f(x)|/\|x\| \leq \|a\|.$$

Since the right member of this inequality does not depend on x, we have

$$\sup |f(x)|/\|x\| \leq \|a\|,$$

i.e. $\|f\| \leq \|a\|$. But, setting $x = a$ we obtain:

$$|f(a)| = (a, a) = \|a\|^2, \qquad \text{i.e. } (|f(a)|/\|a\|) = \|a\|.$$

Therefore $\|f\| = \|a\|$.

If a is zero, then f is the zero linear functional. Hence $\|f\| = \|a\|$ in this case also.

2. The integral

$$I = \int_a^b x(t)\,dt$$

($x(t)$ is a continuous function on $[a, b]$) represents a linear functional on the space $C[a, b]$. Its norm equals $b - a$. In fact,

$$|I| = \left|\int_a^b x(t)\,dt\right| \leq \max|x(t)|\,(b - a),$$

where equality holds when $x =$ constant.

3. Now let us consider a more general example. Let $y_0(t)$ be a fixed continuous function on $[a, b]$. We set, for arbitrary function $x(t) \in C[a, b]$,

$$f(x) = \int_a^b x(t)y_0(t)\,dt.$$

This expression represents a linear functional on $C[a, b]$ because

$$f(\alpha x + \beta y) = \int_a^b (\alpha x(t) + \beta y(t))y_0(t)\,dt$$

$$= \alpha\int_a^b x(t)y_0(t)\,dt + \beta\int_a^b y(t)y_0(t)\,dt = \alpha f(x) + \beta f(y).$$

This functional is bounded. In fact,

$$|f(x)| = \left|\int_a^b x(t)y_0(t)\,dt\right| \leq \|x\|\int_a^b |y_0(t)|\,dt.$$

Thus, the functional $f(x)$ is linear and bounded and consequently it is continuous also. It is possible to show that its norm is exactly equal to $\int_a^b |y_0(t)|\,dt$.

4. We now consider on the same space $C[a, b]$ a linear functional of another type, namely, we set

$$\delta_{t_0}x(t) = x(t_0),$$

i.e. the value of the functional δ_{t_0} for the function $x(t)$ is equal to the value of this function at the fixed point t_0. This functional is frequently encountered, for example, in quantum mechanics where it is usually written in the form

$$\delta_{t_0}x(t) = \int_a^b x(t)\delta(t - t_0)\,dt,$$

where $\delta(t)$ is the "function" equal to zero everywhere except at the point $t = 0$ and such that its integral equals unity (the Dirac δ-function). The δ-function can be thought of as the limit, in some sense, of a sequence of functions $\varphi_n(t)$ each of which assumes the value zero outside some ϵ_n-neighborhood ($\epsilon_n \to 0$ as $n \to \infty$) of the point $t = 0$ and such that the integral of the limiting function equals 1.

5. In the space l_2 we can define a linear functional as in R^n by choosing in l_2 some fixed element $a = (a_1, a_2, \cdots, a_n, \cdots)$ and setting

$$f(x) = \sum_{n=1}^{\infty} x_n a_n . \tag{3}$$

Series (3) converges for arbitrary $x \in l_2$ and

$$|\sum_{n=1}^{\infty} x_n a_n| \leq (\sum_{n=1}^{\infty} x_n^2)^{\frac{1}{2}} (\sum_{n=1}^{\infty} a_n^2)^{\frac{1}{2}} = \|x\| \, \|a\|. \tag{4}$$

Inequality (4) transforms into the identity $\sum_{n=1}^{\infty} a_n^2 \equiv \sum_{n=1}^{\infty} a_n^2$ for $x = a$ and consequently $\|f\| = \|a\|$.

Geometric interpretation of a linear functional. Let $f(x)$ be a linear functional on the space R. We shall assume $f(x)$ is not identically zero. The set L_f of those elements x in R which satisfy the condition $f(x) = 0$ form a subspace. In fact, if $x, y \in L_f$, then

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) = 0,$$

i.e. $\alpha x + \beta y \in L_f$. Further, if $x_n \to x$ and $x_n \in L_f$, then by virtue of the continuity of the functional f,

$$f(x) = \lim_{n \to \infty} f(x_n) = 0.$$

DEFINITION 3. We say that the subspace L of the Banach space R has *index* (or *deficiency*) s if: 1) R contains s linearly independent elements $x_1, x_2, \cdots, x_s$ which do not belong to L with the property that every element $x \in R$ can be represented in the form

$$x = a_1 x_1 + a_2 x_2 + \cdots + a_s x_s + y, \qquad y \in L;$$

and 2) it is impossible to find a smaller number of elements x_i which possess the indicated properties.

In the case of a finite-dimensional space R the index plus the dimension of the subspace L is equal to the dimension of the whole space.

THEOREM 3. *Let $f(x) \not\equiv 0$ be a given functional. The subspace L_f has index equal to unity, i.e. an arbitrary element $y \in R$ can be represented in the form*

$$y = \lambda x_0 + x, \tag{5}$$

where $x \in L_f$, $x_0 \notin L_f$.

Proof. Since $x_0 \notin L_f$, we have $f(x_0) \neq 0$. If we set $\lambda = f(y)/f(x_0)$ and $x = y - \{f(y)/f(x_0)\}x_0$, then $y = \lambda x_0 + x$, where

$$f(x) = f(y) - (f(y)/f(x_0))f(x_0) = 0.$$

If the element x_0 is fixed, then the element y can be represented in the form (5) uniquely. This is easily proved by assuming the contrary. In fact, let

$$y = \lambda x_0 + x,$$
$$y = \lambda' x_0 + x';$$

then

$$(\lambda - \lambda')x_0 = (x' - x).$$

If $\lambda - \lambda' = 0$, then obviously, $x - x' = 0$. But if $\lambda - \lambda' \neq 0$, then $x_0 = (x' - x)/(\lambda - \lambda') \in L_f$, which contradicts the condition that $x_0 \notin L_f$.

Conversely, given a subspace L of R of index 1, L defines a continuous linear functional f which vanishes precisely on L. Indeed, let $x_0 \notin L$. Then for any $x \in R$, $x = y + \lambda x_0$, with $y \in L$, $x_0 \notin L$. Let $f(x) = \lambda$. It is easily seen that f satisfies the above requirements. If f, g are two such linear functionals defined by L, then $f(x) = \alpha g(x)$ for all $x \in R$, α a scalar. This follows because the index of L in R is 1.

We shall now consider the totality M_f of elements in R which satisfy the condition $f(x) = 1$. M_f can be represented in the form $M_f = L_f + x_0$, where x_0 is a fixed element such that $f(x_0) = 1$ and L_f is the totality of elements which satisfy the condition $f(x) = 0$. In analogy with the finite-dimensional case it is natural to call M_f a *hyperplane* in the space R. It is easy to verify that the hyperplanes $f(x) = 1$ and $\varphi(x) = 1$ coincide if, and only if, the functionals f and φ coincide. Thus, it is possible to establish a one-to-one correspondence between all functionals defined on R and all hyperplanes in R which do not pass through the origin of coordinates.

We shall now find the distance from the hyperplane $f(x) = 1$ to the origin. It is equal to

$$d = \inf \{\|x\|; f(x) = 1\}.$$

For all x such that $f(x) = 1$ we have

$$1 \leq \|f\| \, \|x\|, \qquad \text{i.e. } \|x\| \geq 1/\|f\|;$$

therefore $d \geq 1/\|f\|$. Further, since for arbitrary $\epsilon > 0$ an element x satisfying the condition $f(x) = 1$ can be found such that

$$1 > (\|f\| - \epsilon) \, \|x\|,$$

it follows that

$$d = \inf \{\|x\| < 1/(\|f\| - \epsilon); f(x) = 1\}.$$

Consequently,

$$d = 1/\|f\|,$$

i.e. *the norm of the linear functional $f(x)$ equals the reciprocal of the magnitude of the distance of the hyperplane $f(x) = 1$ from the origin of coordinates.*

§24. The conjugate space

It is possible to define the operations of addition and multiplication by a scalar for linear functionals. Let f_1 and f_2 be two linear functionals on a

normed linear space R. Their sum is a linear functional $f = f_1 + f_2$ such that $f(x) = f_1(x) + f_2(x)$ for arbitrary $x \in R$.

The product of a linear functional f_1 by a number α is a functional $f = \alpha f_1$ such that

$$f(x) = \alpha f_1(x)$$

for arbitrary $x \in R$.

It is easy to verify that the operations of addition and multiplication by a scalar of functionals so defined satisfy all the axioms of a linear space. Moreover, the definition we gave above of the norm of a linear functional satisfies all the requirements found in the definition of a normed linear space. In fact,

1) $\| f \| > 0$ for arbitrary $f \not\equiv 0$,

2) $\| \alpha f \| = | \alpha | \, \| f \|$,

3) $$\begin{aligned} \| f_1 + f_2 \| &= \sup \{ | f_1(x) + f_2(x) | / \| x \| \} \\ &\leq \sup \{ (| f_1(x) | + | f_2(x) |) / \| x \| \} \\ &\leq \sup \{ | f_1(x) | / \| x \| \} + \sup \{ | f_2(x) | / \| x \| \} \\ &= \| f_1 \| + \| f_2 \|. \end{aligned}$$

Thus, the totality of all linear functionals on a normed space R itself represents a normed linear space; it is called the conjugate space of R and is denoted by $\bar{R}$.

THEOREM 1. *The conjugate space is always complete.*

Proof. Let $\{f_n\}$ be a fundamental sequence of linear functionals. By the definition of a fundamental sequence, for every $\epsilon > 0$ there exists an N such that $\| f_n - f_m \| < \epsilon$ for all $n, m > N$. Then for arbitrary $x \in R$,

$$| f_n(x) - f_m(x) | \leq \| f_n - f_m \| \, \| x \| < \epsilon \, \| x \|,$$

i.e. for arbitrary $x \in R$ the numerical sequence $f_n(x)$ converges.

If we set

$$f(x) = \lim_{n \to \infty} f_n(x),$$

then $f(x)$ represents a linear functional. In fact,

1) $$\begin{aligned} f(\alpha x + \beta y) &= \lim_{n \to \infty} f_n(\alpha x + \beta y) \\ &= \lim_{n \to \infty} [\alpha f_n(x) + \beta f_n(y)] = \alpha f(x) + \beta f(y). \end{aligned}$$

2) Choose N so that $\| f_n - f_{n+p} \| < 1$ for all $n > N$. Then

$$\| f_{n+p} \| < \| f_n \| + 1$$

for all p. Consequently, $| f_{n+p}(x) | \leq (\| f_n \| + 1) \| x \|$.

Passing to the limit as $p \to \infty$, we obtain

$$\lim_{p \to \infty} | f_{n+p}(x) | = | f(x) | \leq (\| f_n \| + 1) \| x \|,$$

i.e. the functional $f(x)$ is bounded. We shall now prove that the functional f is the limit of the sequence $f_1, f_2, \cdots, f_n, \cdots$. By the definition of the norm, for every $\epsilon > 0$ there exists an element x_ϵ such that

$$\begin{aligned} \| f_n - f \| &\leq \{ (| f_n(x_\epsilon) - f(x_\epsilon) |) / \| x_\epsilon \| \} + \epsilon/2 \\ &= | f_n(x_\epsilon / \| x_\epsilon \|) - f(x_\epsilon / \| x_\epsilon \|) | + \epsilon/2; \end{aligned}$$

since

$$f(x_\epsilon / \| x_\epsilon \|) = \lim_{n \to \infty} f_n(x_\epsilon / \| x_\epsilon \|),$$

it is possible to find an $n_0(\epsilon)$ such that for $n > n_0$

$$| f_n(x_\epsilon / \| x_\epsilon \|) - f(x_\epsilon / \| x_\epsilon \|) | < \epsilon/2,$$

so that for $n > n_0$ the inequality

$$\| f_n - f \| < \epsilon$$

is fulfilled.

This completes the proof of the theorem.

Let us emphasize once more that this theorem is valid independently of whether the initial space R is complete or not.

EXAMPLES. 1. Let the space E be finite-dimensional with basis $e_1, e_2, \cdots, e_n$. Then the functional $f(x)$ is expressible in the form

$$f(x) = \sum_{i=1}^{n} f_i x_i, \tag{1}$$

where $x = \sum_{i=1}^{n} x_i e_i$ and $f_i = f(e_i)$.

Thus, the functional is defined by the n numbers $f_1, \cdots, f_n$ which are the values of f on the basis vectors. The space which is the conjugate of the finite-dimensional space is also finite-dimensional and has the same dimension.

The explicit form assumed by the norm in the conjugate space depends on the choice of norm in E.

a) Let $\| x \| = (\sum x_i^2)^{\frac{1}{2}}$. We have already shown that then

$$\| f \| = (\sum f_i^2)^{\frac{1}{2}},$$

i.e. the conjugate of an Euclidean space is itself Euclidean.

b) Let $\| x \| = \sup_i | x_i |$. Then

$$| f(x) | = | \sum f_i x_i | \leq (\sum | f_i |) \sup_i | x_i | = (\sum | f_i |) \| x \|.$$

From this it follows that

$$\| f \| \leq \sum | f_i |.$$

The norm $\| f \|$ cannot be less than $\sum | f_i |$ since if we set

$$x_i = \begin{cases} +1 \text{ if } f_i > 0, \\ -1 \text{ if } f_i < 0, \\ 0 \text{ if } f_i = 0, \end{cases}$$

then the following equality is valid:

$$| f(x) | = \sum | f_i | = (\sum | f_i |) \| x \|.$$

c) If $\| x \| = (\sum | x_i |^p)^{1/p}$, $p > 1$, then $\| f \| = (\sum | f_i |^q)^{1/q}$, where $1/p + 1/q = 1$. This follows from the Hölder inequality

$$| \sum f_i x_i | \leq (\sum | x_i |^p)^{1/p} (\sum | f_i |^q)^{1/q}$$

and from the fact that the equality sign is attained [for $f_i = (\operatorname{sgn} x_i)(x_i)^{p-1}$].

2. Let us consider the space c consisting of sequences $x = (x_1, x_2, \cdots, x_n, \cdots)$ which are such that $x_n \to 0$ as $n \to \infty$, where $\| x \| = \sup_n x_n$.

If a functional in the space c is expressible by means of the formula

$$f(x) = \sum_{i=1}^{\infty} f_i x_i, \qquad \sum_{i=1}^{\infty} | f_i | < \infty, \tag{2}$$

then it has norm

$$\| f \| = \sum_{i=1}^{\infty} | f_i |.$$

The inequality $\| f \| \leq \sum_{i=1}^{\infty} | f_i |$ is obvious. On the other hand, if $\sum_{i=1}^{\infty} | f_i | = a$, then for every $\epsilon > 0$ it is possible to find an N such that $\sum_{i=1}^{N} | f_i | > a - \epsilon$.

We now set

$$x_n = \left\{ \begin{matrix} +1 \text{ if } f_n > 0 \\ -1 \text{ if } f_n < 0 \\ 0 \text{ if } f_n = 0 \end{matrix} \right\} n \leq N.$$
$$0 \text{ if } n > N$$

Then

$$| f(x) | = \sum_{n=1}^{N} | f_n | > a - \epsilon,$$

whence it follows that $\| f \| = a$.

We shall prove that all functionals in the space c have the form (2). We shall set

$$e_n = (0, 0, \cdots, 1, 0, \cdots),$$

i.e. e_n denotes the sequence in which the n-th entry is unity and the remaining are zeros.

Let the functional $f(x)$ be given; we denote $f(e_n)$ by f_n. If

$$x = (x_1, x_2, \cdots, x_n, 0, \cdots), \tag{3}$$

then

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n \quad \text{and} \quad f(x) = \sum_{i=1}^{n} f_i x_i.$$

The sum $\sum_{n=1}^{\infty} | f_n |$ is $< \infty$ for every bounded linear functional. If $\sum_{n=1}^{\infty} | f_n |$ were $= \infty$, then for every H it would be possible to find an N such that

$$\sum_{n=1}^{N} | f_n | > H.$$

We construct the element x in the following way:

$$x_n = \left\{ \begin{matrix} 1 \text{ if } f_n > 0 \\ -1 \text{ if } f_n < 0 \\ 0 \text{ if } f = 0 \end{matrix} \right\} n \leq N.$$
$$0 \text{ if } n > N$$

The norm of such an element is equal to unity, and

$$| f(x) | = \sum_{i=1}^{N} f_i x_i = \sum_{i=1}^{N} | f_i | > H = H \| x \|,$$

which contradicts the assumption concerning the boundedness of the functional.

The set of elements of the form (3) is everywhere dense in the space c. Therefore the continuous linear functional is uniquely defined by its values on this set. Thus, for every x

$$f(x) = \sum_{n=1}^{\infty} f_i x_i.$$

The space which is conjugate to the space c consists of sequences $(f_1, f_2, \cdots, f_n, \cdots)$ satisfying the condition $\sum_{i=1}^{\infty} | f_n | < \infty$.

3. Let the space consist of sequences

$$x = (x_1, x_2, \cdots, x_n, \cdots), \qquad \sum_{i=1}^{\infty} | x_i | < \infty$$

with norm $\| x \| = \sum_{n=1}^{\infty} | x_n |$.

It can be proved that the space conjugate to this space is the space of bounded sequences

$$f = (f_1, f_2, \cdots, f_n, \cdots)$$

with norm $\| f \| = \sup_n | f_n |$.

In all the examples of finite-dimensional spaces introduced above, the space which is conjugate to the conjugate space coincides with the initial

space. This is always so in the finite-dimensional case. However, as Examples 2 and 3 show, in the infinite-dimensional case the space conjugate to the conjugate space may not coincide with the initial space.

We consider cases when this coincidence holds also in infinite-dimensional space.

4. The space l_2 consists of sequences

$$x = (x_1, x_2, \cdots, x_n, \cdots),$$

with $\sum_{i=1}^{\infty} x_i^2 < \infty$ and norm $\| x \| = (\sum_{i=1}^{\infty} x_i^2)^{\frac{1}{2}}$. All functionals in the space l_2 have the form

$$f(x) = \sum_{i=1}^{\infty} f_i x_i .$$

We shall prove this assertion.

To each functional there is set into correspondence the sequence $f_1, f_2, \cdots, f_n, \cdots$ of its values on the elements $e_1, e_2, \cdots, e_n, \cdots$ defined exactly as in Example 2, above.

If the functional is bounded, then $\sum_{i=1}^{\infty} f_i^2 < \infty$. We shall assume the contrary, i.e. we shall assume that for every H there exists an N such that

$$\sum_{i=1}^{N} f_i^2 = U \geq H.$$

If we apply the functional under consideration to the element

$$x = (f_1, f_2, \cdots, f_N, 0, \cdots), \qquad \| x \| = U^{\frac{1}{2}},$$

we obtain

$$f(x) = \sum_{i=1}^{N} f_i^2 = U \geq H^{\frac{1}{2}} \| x \|,$$

contrary to the assumption that the functional is bounded.

Since the functional f is linear, its values on the elements of the form $x = (x_1, x_2, \cdots, x_n, 0, \cdots)$ are easily found; on all other elements of the space the values of f are found from continuity considerations and we always have

$$f(x) = \sum_{i=1}^{\infty} f_i x_i .$$

The norm of the functional f equals $(\sum_{i=1}^{\infty} f_i^2)^{\frac{1}{2}}$. This is established with the aid of the Schwarz inequality.

5. The space l_p is the space of all sequences of the form

$$x = (x_1, x_2, \cdots, x_n, \cdots), \ (\sum_{i=1}^{\infty} x_i^p)^{1/p} < \infty, \ \| x \| = (\sum_{i=1}^{\infty} x_i^p)^{1/p}.$$

The conjugate space of l_p is the space l_q, where $1/p + 1/q = 1$. The proof is analogous to the preceding proof. *Hint*: Use Hölder's inequality.

§25. Extension of linear functionals

THEOREM (HAHN-BANACH). *Every linear functional $f(x)$ defined on a linear subspace G of a normed linear space E can be extended to the entire space with preservation of norm, i.e. it is possible to construct a linear functional $F(x)$ such that*

1) $F(x) = f(x), \qquad x \in G,$

2) $\| F \|_E = \| f \|_G .$

Proof. The theorem will be proved for a separable space E, although in actuality it is valid also in spaces which are not separable.

First, we shall extend the functional to the linear subspace G_1 obtained by adding to G some element $x_0 \notin G$. An arbitrary element of this subspace is uniquely representable in the form

$$y = tx_0 + x, \qquad x \in G.$$

If the functional sought exists, then

$$F(y) = tF(x_0) + f(x)$$

or, if we set $-F(x_0) = c$, then $F(y) = f(x) - ct$.

In order that the norm of the functional be not increased when it is continued it is necessary to find a c such that the inequality

$$| f(x) - ct | \leq \| f \| \, \| x + tx_0 \| \tag{1}$$

be fulfilled for all $x \in G$.

If we denote the element x/t by z ($z \in G$), the inequality (1) can be rewritten

$$| f(z) - c | \leq \| f \| \, \| z + x_0 \|.$$

This inequality is equivalent to the following inequalities:

$$-\| f \| \, \| z + x_0 \| \leq f(z) - c \leq \| f \| \, \| z + x_0 \|,$$

or, what amounts to the same thing,

$$f(z) + \| f \| \, \| z + x_0 \| \geq c \geq f(z) - \| f \| \, \| z + x_0 \|,$$

for all $z \in G$. We shall prove that such a number c always exists. To do this we shall show that for arbitrary elements $z', z'' \in G$ we always have

$$f(z'') + \| f \| \, \| z'' + x_0 \| \geq f(z') - \| f \| \, \| z' + x_0 \|. \tag{2}$$

But this follows directly from the obvious inequality

$$\begin{aligned} f(z') - f(z'') \leq \| f \| \, \| z' - z'' \| &= \| f \| \, \| z' + x_0 - (z'' + x_0) \| \\ &\leq \| f \| \, \| z' + x_0 \| + \| f \| \, \| z'' + x_0 \|. \end{aligned}$$

We introduce the notation:

$$c' = \inf \{ f(z) + \| f \| \, \| z + x_0 \| ; z \in G \},$$
$$c'' = \sup \{ f(z) - \| f \| \, \| z + x_0 \| ; z \in G \}.$$

It follows from inequality (2) that $c'' \leq c'$.

We take an arbitrary c such that $c'' \leq c \leq c'$. We set

$$F_1(x) = f(x) - ct$$

for the elements of the subspace $G_1 = \{G \cup x_0\}$. We obtain the linear functional F_1, where $\| F_1 \| = \| f \|$.

The separable space E contains a denumerable everywhere dense set $x_1, x_2, \cdots, x_n, \cdots$. We shall construct the linear subspaces

$$\begin{aligned} G_1 &= \{G \cup x_0\}, \\ G_2 &= \{G_1 \cup x_1\}, \\ &\cdots\cdots\cdots\cdots \\ G_{n+1} &= \{G_n \cup x_n\}, \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

and define the functional F on them as follows: we construct functionals F_n which coincide with F_{n-1} on G_{n-1} and which have norm equal to $\| f \|$. Thus, we obtain the functional F defined on a set which is everywhere dense in E. At the remaining points of E the functional is defined by continuity: if $x = \lim_{n\to\infty} x_n$, then $F(x) = \lim_{n\to\infty} F(x_n)$. The inequality $| F(x) | \leq \| f \| \| x \|$ is valid since

$$| F(x) | = \lim_{n\to\infty} | F(x_n) | \leq \lim_{n\to\infty} \| f \| \| x_n \| = \| f \| \| x \|.$$

This completes the proof of the theorem on the extension of a functional.

COROLLARY. Let x_0 be an arbitrary nonzero element in R and let M be an arbitrary positive number. Then there exists a linear functional $f(x)$ in R such that

$$\| f \| = M \quad \text{and} \quad f(x_0) = \| f \| \| x_0 \|.$$

In fact, if we set $f(tx_0) = tM \| x_0 \|$, we obtain a linear functional with norm equal to M which is defined on the one-dimensional subspace of elements of the form tx_0 and then, by the Hahn-Banach theorem, we can extend it to all of R without increasing the norm. The geometric interpretation of this fact is the following: in a Banach space through every point x_0 there can be drawn a hyperplane which is tangent to the sphere $\| x \| = \| x_0 \|$.

§26. The second conjugate space

Inasmuch as the totality $\bar{R}$ of linear functionals on a normed linear space R itself represents a normed linear space it is possible to speak of the space $\bar{\bar{R}}$ of linear functionals on $\bar{R}$, i.e. of the second conjugate space with respect to R, and so forth. We note first of all that every element in R defines a linear functional in $\bar{R}$. In fact, let

$$\psi_{x_0}(f) = f(x_0),$$

where x_0 is a fixed element in R and f runs through all of $\bar{R}$. Thus, to each $f \in \bar{R}$ there is set into correspondence some number $\psi_{x_0}(f)$. In this connection we have

$$\psi_{x_0}(\alpha f_1 + \beta f_2) = \alpha f_1(x_0) + \beta f_2(x_0) = \alpha \psi_{x_0}(f_1) + \beta \psi_{x_0}(f_2)$$

and

$$| \psi_{x_0}(f) | \leq \| f \| \| x_0 \| \qquad \text{(boundedness)},$$

i.e. $\psi_{x_0}(f)$ is a bounded linear functional on $\bar{R}$.

Besides the notation $f(x)$ we shall also use the more symmetric notation:

$$(1) \qquad (f, x)$$

which is analogous to the symbol used for the scalar product. For fixed $f \in \bar{R}$ we can consider this expression as a functional on R and for fixed $x \in R$ as a functional on $\bar{R}$.

From this it follows that the norm of every $x \in R$ is defined in two ways: firstly, its norm is defined as an element in R, and secondly, as the norm of a linear functional on $\bar{R}$, i.e. as an element in $\bar{\bar{R}}$. Let $\| x \|$ denote the norm of x taken as an element in R and let $\| x \|_2$ be the norm of x taken as an element in $\bar{\bar{R}}$. We shall show that in fact $\| x \| = \| x \|_2$. Let f be an arbitrary nonzero element in $\bar{R}$. Then

$$| (f, x) | \leq \| f \| \| x \|, \qquad \| x \| \geq | (f, x) | / \| f \|;$$

since the left member of the last inequality does not depend on f, we have

$$\| x \| \geq \sup \{ | (f, x) | / \| f \|; f \in \bar{R}, f \neq 0 \} = \| x \|_2 .$$

But, according to the corollary to the Hahn-Banach theorem, for every $x \in R$ a linear functional f_0 can be found such that

$$| (f_0 , x) | = \| f_0 \| \| x \|.$$

Consequently,

$$\sup \{ | (f, x) | / \| f \|; f \in R \} = \| x \|,$$

i.e. $\| x \|_2 = \| x \|$.

This proves the following theorem.

THEOREM. *R is isometric to some linear manifold in $\bar{\bar{R}}$.*

Inasmuch as we agreed not to distinguish between isometric spaces this theorem can be formulated as follows: $R \subset \bar{\bar{R}}$.

The space R is said to be *reflexive* in case $\bar{\bar{R}} = R$. If $\bar{\bar{R}} \neq R$, then R is said to be *irreflexive*.

Finite-dimensional space R^n and the space l_2 are examples of reflexive spaces (we even have $\bar{\bar{R}} = R$ for these spaces).

The space c of all sequences which converge to zero is an example of a

complete irreflexive space. In fact, above (§24, Examples 2 and 3) we proved that the conjugate space of the space c is the space l of numerical sequences $(x_1, x_2, \cdots, x_n, \cdots)$ which satisfy the condition $\sum_{n=1}^{\infty} |x_n| < \infty$, to which in turn the space m of all bounded sequences is conjugate. The spaces c and m are not isometric. This follows from the fact that c is separable and m is not. Thus, c is irreflexive.

The space $C[a, b]$ of continuous functions on a closed interval $[a, b]$ is also irreflexive. However, we shall not stop to prove this assertion. (The following stronger assertion can also be proved: *No* normed linear space exists for which $C[a, b]$ is the conjugate space.)

A. I. Plessner established that for an arbitrary normed space R there exist only two possibilities: either the space R is reflexive, i.e. $R = \bar{\bar{R}} = \bar{\bar{\bar{R}}} = \cdots$; $\bar{R} = \bar{\bar{R}} = \cdots$; or the spaces $R, \bar{R}, \bar{\bar{R}}, \cdots$ are all distinct.

The space $l_p (p > 1)$ is an example of a reflexive space (since $\bar{l}_p = l_q$, where $1/p + 1/q = 1$, we have $\bar{\bar{l}}_p = \bar{l}_q = l_p$).

§27. Weak convergence

The concept of so-called weak convergence of elements in a normed linear space plays an important role in many questions of analysis.

DEFINITION. A sequence $\{x_n\}$ of elements in a normed linear space R *converges weakly* to the element x if

1) The norms of the elements x_n are uniformly bounded: $\| x_n \| \leq M$, and

2) $f(x_n) \to f(x)$ for every $f \in \bar{R}$.

(It can be shown that Condition 1 follows from 2; we shall not carry out this proof.)

Condition 2 can be weakened slightly; namely, the following theorem is true.

THEOREM 1. *The sequence $\{x_n\}$ converges weakly to the element x if*

1) $\| x_n \| \leq M$, *and*

2) $f(x_n) \to f(x)$ *for every* $f \in \Delta$, *where* Δ *is a set whose linear hull is everywhere dense in* $\bar{R}$.

Proof. It follows from the conditions of the theorem and the definition of a linear functional that if φ is a linear combination of elements in Δ then $\varphi(x_n) \to \varphi(x)$.

Now let φ be an arbitrary linear functional on R and let $\{\varphi_k\}$ be a sequence of functionals which converges to φ, each of which is a linear combination of elements in Δ. We shall show that $\varphi(x_n) \to \varphi(x)$. Let M be such that $\| x_n \| \leq M$ $(n = 1, 2, \cdots)$ and $\| x \| \leq M$.

Let us evaluate the difference $|\varphi(x_n) - \varphi(x)|$. Since $\varphi_k \to \varphi$, given an arbitrary $\epsilon > 0$ a K can be found such that for all $k > K$,

$$\| \varphi - \varphi_k \| < \epsilon;$$

it follows from this that

$$|\varphi(x_n) - \varphi(x)| \leq |\varphi(x_n) - \varphi_k(x_n)| + |\varphi_k(x_n) - \varphi_k(x)| + |\varphi_k(x) - \varphi(x)| \leq \epsilon M + \epsilon M + |\varphi_k(x_n) - \varphi_k(x)|.$$

But by assumption $|\varphi_k(x_n) - \varphi_k(x)| \to 0$ as $n \to \infty$. Consequently, $|\varphi(x_n) - \varphi(x)| \to 0$ as $n \to \infty$.

If the sequence $\{x_n\}$ converges in norm to x, that is, if $\| x_n - x \| \to 0$ as $n \to \infty$, then such convergence is frequently called strong convergence to distinguish it from weak convergence.

If a sequence $\{x_n\}$ converges strongly to x, it also converges weakly to the same limit. In fact, if $\| x_n - x \| \to 0$, then

$$|f(x_n) - f(x)| \leq \| f \| \, \| x_n - x \| \to 0$$

for an arbitrary linear functional f. The converse is not true in general: strong convergence does not follow from weak convergence. For example, in l_2 the sequence of vectors

$$e_1 = (1, 0, 0, \cdots),$$
$$e_2 = (0, 1, 0, \cdots),$$
$$e_3 = (0, 0, 1, \cdots),$$
$$\cdots\cdots\cdots\cdots$$

converges weakly to zero. In fact, every linear functional f in l_2 can be represented as the scalar product with some fixed vector

$$a = (a_1, a_2, \cdots, a_n, \cdots), \qquad f(x) = (x, a);$$

hence,

$$f(e_n) = (e_n, a) = a_n.$$

Since $a_n \to 0$ as $n \to \infty$ for every $a \in l_2$, we have $\lim f(e_n) = 0$ for every linear functional in l_2.

But at the same time the sequence $\{e_n\}$ does not converge in the strong sense to any limit.

We shall investigate what weak convergence amounts to in several concrete spaces.

EXAMPLES. 1. In *finite-dimensional space* R^n weak and strong convergence coincide. In fact, consider functionals corresponding to multiplication by the elements

$$e_1 = (1, 0, 0, \cdots, 0),$$
$$e_2 = (0, 1, 0, \cdots, 0),$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$e_n = (0, 0, 0, \cdots, 1).$$

If $\{x_k\}$ converges weakly to x, then

$$(x_k , e_i) = x_k^{(i)} \to x^{(i)} \qquad (i = 1, 2, \cdots , n),$$

i.e. the first coordinates of the vectors x_k tend to the first coordinate of the vector x, their second coordinates tend to the second coordinate of the vector x, and so forth. But then

$$\rho(x_k , x) = \{\sum_{i=1}^{n} (x_k^{(i)} - x^{(i)})^2\}^{\frac{1}{2}} \to 0,$$

i.e. $\{x_k\}$ converges strongly to x. Since strong convergence always implies weak convergence, our assertion is proved.

2. *Weak convergence in l_2.* Here we can take for the set Δ, linear combinations of whose elements are everywhere dense in l_2, the totality of vectors

$$e_1 = (1, 0, 0, \cdots),$$
$$e_2 = (0, 1, 0, \cdots),$$
$$e_3 = (0, 0, 1, \cdots),$$
$$\cdots\cdots\cdots\cdots\cdots.$$

If $x = (\xi_1 , \xi_2 , \cdots , \xi_n , \cdots)$ is an arbitrary vector in l_2, then the values assumed at x by the corresponding linear functionals are equal to $(x, e_n) = \xi_n$, i.e. to the coordinates of the vector x. Consequently, weak convergence of the sequence $\{x_n\}$ in l_2 means that the numerical sequence of the k-th coordinates of these vectors ($k = 1, 2, \cdots$) converges. We saw above that this convergence does not coincide with strong convergence in l_2.

3. *Weak convergence in the space of continuous functions.* Let $C[a, b]$ be the space of continuous functions defined on the closed interval $[a, b]$. It can be shown that the totality Δ of all linear functionals δ_{t_0}, each of which is defined as the value of the function at some fixed point t_0 (see Example 4, §23) satisfies the conditions of Theorem 1, i.e. linear combinations of these functionals are everywhere dense in $\bar{C}[a, b]$. For each such functional δ_{t_0}, the condition $\delta_{t_0}x_n(t) \to \delta_{t_0}x(t)$ is equivalent to the condition $x_n(t_0) \to x(t_0)$.

Thus, weak convergence of a sequence of continuous functions means that this sequence is a) uniformly bounded and b) convergent at every point.

It is clear that this convergence does not coincide with convergence in norm in $C[a, b]$, i.e. it does not coincide with uniform convergence of continuous functions. (Give a suitable example!)

§28. Weak convergence of linear functionals

We can introduce the concept of weak convergence of linear functionals as analogous to the concept of weak convergence of elements of a normed linear space R.

DEFINITION. A sequence $\{f_n\}$ of linear functionals converges weakly to the linear functional f if

1) $\| f_n \|$ are uniformly bounded; i.e. $\| f_n \| \leq M$, and

2) $f_n(x) \to f(x)$ for every element $x \in R$.

(This is usually called weak* convergence.—Trans.)

Weak convergence of linear functionals possesses properties which are analogous to properties stated above of weak convergence of elements, namely strong convergence (i.e. convergence in norm) of linear functionals implies their weak convergence, and it is sufficient to require the fulfillment of the condition $f_n(x) \to f(x)$ not of all $x \in R$, but only for a set of elements linear combinations of which are everywhere dense in R.

We shall consider one important example of weak convergence of linear functionals. Above (§23, Example 4), we spoke of the fact that the "δ-function", i.e. the functional on $C[a, b]$, which assigns to every continuous function its value at the point zero, can "in some sense" be considered as the limit of "ordinary" functions, each of which assumes the value zero outside some small neighborhood of zero and has an integral equal to 1. [We assume that the point $t = 0$ belongs to the interval (a, b). Of course, one can take any other point instead of $t = 0$.] Now we can state this assertion precisely. Let $\{\varphi_n(t)\}$ be a sequence of continuous functions satisfying the following conditions:

(1) 1) $\varphi_n(t) = 0$ for $|t| > 1/n$, $\varphi_n(t) \geq 0$,

2) $\int_a^b \varphi_n(t)\, dt = 1.$

Then for an arbitrary continuous function $x(t)$ defined on the closed interval $[a, b]$, we have

$$\int_a^b \varphi_n(t)x(t)\, dt = \int_{-1/n}^{1/n} \varphi_n(t)x(t)\, dt \to x(0) \text{ as } n \to \infty.$$

In fact, by the mean-value theorem,

$$\int_{-1/n}^{1/n} \varphi_n(t)x(t)\, dt = x(\xi_n) \int_{-1/n}^{1/n} \varphi_n(t)\, dt = x(\xi_n), \qquad -1/n \leq \xi_n \leq 1/n;$$

when $n \to \infty$, $\xi_n \to 0$ and $x(\xi_n) \to x(0)$.

The expression

$$\int_a^b \varphi_n(t)x(t)\, dt$$

represents a linear functional on the space of continuous functions. Thus, the result we obtained can be formulated as follows: the δ-function is the limit of the sequence (1) in the sense of weak convergence of linear functionals.

The following theorem plays an important role in various applications of the concept of weak convergence of linear functionals.

THEOREM 1. *If the normed linear space R is separable, then an arbitrary bounded sequence of linear functionals on R contains a weakly convergent subsequence.*

Proof. Choose in R a denumerable everywhere dense set

$$\{x_1, x_2, \cdots, x_n, \cdots\}.$$

If $\{f_n\}$ is a bounded (in norm) sequence of linear functionals on R, then

$$f_1(x_1), f_2(x_1), \cdots, f_n(x_1), \cdots$$

is a bounded numerical sequence. Therefore we can select from $\{f_n\}$ a subsequence

$$f_1', f_2', \cdots, f_n', \cdots$$

such that the numerical sequence

$$f_1'(x_1), f_2'(x_1), \cdots$$

converges. Furthermore, from the subsequence $\{f_n'\}$ we can select a subsequence

$$f_1'', f_2'', \cdots, f_n'', \cdots$$

such that

$$f_1''(x_2), f_2''(x_2), \cdots$$

converges, and so forth. Thus, we obtain a system of sequences

$$\begin{aligned} &f_1', f_2', \cdots, f_n', \cdots, \\ &f_1'', f_2'', \cdots, f_n'', \cdots, \\ &\cdots\cdots\cdots\cdots\cdots \end{aligned}$$

each of which is a subsequence of the one preceding. Then taking the "diagonal" subsequence $f_1', f_2'', f_3''', \cdots$, we obtain a sequence of linear functionals such that $f_1'(x_n), f_2''(x_n), \cdots$ converges for all n. But then $f_1'(x)$, $f_2''(x), \cdots$ also converges for arbitrary $x \in R$. This completes the proof of the theorem.

The last theorem suggests the following question. Is it possible in the space $\bar{R}$, conjugate to a separable space, to introduce a metric so that the bounded subsets of the space $\bar{R}$ become compact with respect to this new metric? In other words, is it possible to introduce a metric in $\bar{R}$ so that convergence in the sense of this metric in $\bar{R}$ coincides with weak convergence of elements in $\bar{R}$ considered as linear functionals. Such a metric can in fact be introduced in $\bar{R}$.

Let $\{x_n\}$ be a denumerable everywhere dense set in R. Set

$$\rho(f_1, f_2) = \sum_{n=1}^{\infty} |f_1(x_n) - f_2(x_n)|/2^n \|x_n\| \tag{2}$$

for any two elements $f_1, f_2 \in \bar{R}$. This series converges since its n-th term does not exceed $(\|f_1\| + \|f_2\|)/2^n$. The quantity (2) possesses all the properties of a distance. In fact, the first two axioms are obviously satisfied. We shall verify the triangle axiom.

Since

$$\begin{aligned} |f_1(x_n) - f_3(x_n)| &= |f_1(x_n) - f_2(x_n) + f_2(x_n) - f_3(x_n)| \\ &\leq |f_1(x_n) - f_2(x_n)| + |f_2(x_n) - f_3(x_n)|, \end{aligned}$$

we have

$$\rho(f_1, f_3) \leq \rho(f_1, f_2) + \rho(f_2, f_3).$$

Direct verification shows that convergence in the sense of this metric is in fact equivalent to weak convergence in $\bar{R}$.

Now Theorem 1 can be formulated in the following manner.

THEOREM 1'. *In a space $\bar{R}$, which is the conjugate of a separable space, with metric* (2), *every bounded subset is compact.*

§29. Linear operators

1. *Definition of a linear operator. Boundedness and continuity.*

Let R and R' be two Banach spaces whose elements are denoted respectively by x and y. Let a rule be given according to which to each x in some set $X \subseteq R$ there is assigned some element y in the space R'. Then we say that an *operator* $y = Ax$ with range of values in R' has been defined on the set X.

DEFINITION 1. An operator A is said to be *linear* if the equality

$$A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A x_1 + \alpha_2 A x_2$$

is satisfied for any two elements $x_1, x_2 \in X$ and arbitrary real numbers α_1, α_2.

DEFINITION 2. An operator A is said to be *bounded* if there exists a constant M such that

$$\|Ax\| \leq M \|x\|$$

for all $x \in X$.

DEFINITON 3. An operator A is said to be *continuous* if for arbitrary $\epsilon > 0$ there exists a number $\delta > 0$ such that the inequality

$$\|x' - x''\|_R < \delta \qquad (x', x'' \in X)$$

implies that

$$\|Ax' - Ax''\|_{R'} < \epsilon.$$

Only linear operators will be considered in the sequel. If the space R' is the real line, the operator $y = A(x)$ is a functional, and the formulated

definitions of linearity, continuity and boundedness go over into the corresponding definitions introduced in §23 for functionals.

The following theorem is a generalization of Theorem 1, §23.

THEOREM 1. *Continuity is equivalent to boundedness for a linear operator.*

Proof. 1. Assume the operator A is bounded. The inequality $\| x' - x'' \| < \delta$ implies that

$$\| Ax' - Ax'' \| = \| A(x' - x'') \| \leq M \| x' - x'' \| \leq M\delta, \tag{1}$$

where M is the constant occurring in the definition of boundedness. If we take $\delta < \epsilon/M$, inequality (1) yields $\| Ax' - Ax'' \| < \epsilon$, i.e. the operator A is continuous.

2. Assume now that the operator A is continuous. We shall prove that A is bounded by contradiction. We assume that A is not bounded. Then there exists a sequence

$$x_1, x_2, \cdots, x_n, \cdots \tag{2}$$

such that

$$\| Ax_n \| > n \| x_n \|.$$

Set $z_n = x_n/n \| x_n \|$; it is obvious that $\| z_n \| = 1/n$, i.e. $z_n \to 0$ as $n \to \infty$. Consider the sequence

$$Az_n = Ax_n/n \| x_n \|$$

which is the map of the sequence $\{z_n\}$ under A. The norm of each element Az_n is not less than 1:

$$\| Az_n \| = \| Ax_n \|/n \| x_n \| \geq n \| x_n \|/ n\| x_n \| = 1.$$

Since for every linear operator, $A(0) = 0$, and $\lim_{n\to\infty} z_n = 0$, we obtain a contradiction of our hypothesis that the operator is continuous. Consequently, the operator A must be bounded.

EXAMPLE. *The general form of a linear operator mapping a finite-dimensional space into a finite-dimensional space.* Given an n-dimensional space R^n with basis $e_1, e_2, \cdots, e_n$, every point of this space can be represented in the form $x = \sum_{i=1}^{n} x_i e_i$.

A linear operator A maps R^n into the finite-dimensional space R^m with the basis $e_1', e_2', \cdots, e_m'$.

Let us consider the representation with respect to this basis of the images of the basis vectors of the space R^n:

$$Ae_i = \sum_{j=1}^{m} a_{ij} e_j'.$$

Now let $y = Ax$,

$$y = Ax = \sum_{i=1}^{n} x_i Ae_i = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} a_{ij} e_j' = \sum_{j=1}^{m} d_j e_j',$$

where

$$d_j = \sum_{i=1}^{n} a_{ij} x_i. \tag{3}$$

It is clear from formula (3) that to determine the operator A it is sufficient to give the coefficient matrix with entries a_{ij}.

A linear operator cannot map a finite-dimensional space into a space of greater dimension since all linear relations among the elements are preserved for their images.

2. *Norm of an operator. Sum and product of operators. Product of an operator by a scalar.*

DEFINITION 4. Let A be a bounded linear operator. This means that there exist numbers M such that

$$\| Ax \| < M \| x \| \tag{4}$$

for all $x \in X$. The *norm* $\| A \|$ of the operator A is the greatest lower bound of the numbers M which satisfy condition (4). It follows from the definition of the norm of an operator that $\| Ax \| \leq \| A \| \| x \|$. But if $M < \| A \|$, then there exists an element x such that $\| Ax \| > M \| x \|$.

THEOREM 2. *If A is an arbitrary linear operator, then*

$$\| A \| = \sup \{ \| Ax \|; \| x \| = 1 \} = \sup \{ \| Ax \|/\| x \|; \| x \| \neq 0 \}.$$

Proof. Introduce the notation

$$\alpha = \sup \{ \| Ax \|; \| x \| = 1 \} = \sup \{ \| Ax \|/\| x \|; \| x \| \neq 0 \}.$$

We shall first prove that $\| A \| \geq \alpha$. Since $\alpha = \sup\{ \| Ax \|/\| x \|; \| x \| \neq 0\}$, for arbitrary $\epsilon > 0$ there exists an element x_1 not zero such that $\| Ax_1 \|/\| x_1 \| > \alpha - \epsilon$ or $\| Ax_1 \| > (\alpha - \epsilon) \| x_1 \|$, which implies that $\alpha - \epsilon < \| A \|$ and hence that $\| A \| \geq \alpha$ because ϵ is arbitrary.

The inequality cannot hold. In fact, if we let $\| A \| - \alpha = \epsilon$, then $\alpha < \| A \| - \epsilon/2$. But this implies that the following inequalities hold for an arbitrary point x:

$$\| Ax \|/\| x \| \leq \alpha < \| A \| - \epsilon/2,$$

or

$$\| Ax \| \leq (\| A \| - \epsilon/2) \| x \|,$$

i.e. $\| A \|$ is not the greatest lower bound of those M for which $\| Ax \| \leq M \| x \|$. It is clear from this contradiction that $\| A \| = \alpha$.

In the sequel we shall make use of the above expression for the norm of an operator as equivalent to the original definition of the norm.

DEFINITION 5. Let A_1 and A_2 be two given continuous linear operators which transform the Banach space E into the Banach space E_1. The *sum*

of these two operators is the operator A which puts the element $y \in E_1$ defined by the formula $y = A_1x + A_2x$ into correspondence with the element $x \in E$. It is easy to verify that $A = A_1 + A_2$ is also a linear operator.

THEOREM 3. *The following relation holds for the norms of the operators* A_1, A_2 *and* $A = A_1 + A_2$:

$$\| A \| \leq \| A_1 \| + \| A_2 \|. \tag{5}$$

Proof. It is clear that

$$\| Ax \| = \| A_1x + A_2x \| \leq \| A_1x \| + \| A_2x \| \leq (\| A_1 \| + \| A_2 \|) \| x \|,$$

whence inequality (5) follows.

DEFINITION 6. Let A_1 and A_2 be continuous linear operators where A_1 transforms the Banach space E into the Banach space E_1 and A_2 transforms the Banach space E_1 into the Banach space E_2. The *product* of the operators A_1 and A_2 (denoted by $A = A_2A_1$) is the operator which sets the element $z \in E_2$ into correspondence with the element $x \in E$, where

$$z = A_2(A_1x).$$

THEOREM 4. *If* $A = A_2A_1$, *then*

$$\| A \| \leq \| A_2 \| \| A_1 \|. \tag{6}$$

Proof. $\| Ax \| = \| A_2(A_1x) \| \leq \| A_2 \| \| A_1x \| \leq \| A_2 \| \| A_1 \| \| x \|,$ whence the assertion of the theorem follows.

The sum and product of three or more operators are defined by iteration. Both operations are associative.

The product of the operator A and the real number k (denoted by kA) is defined in the following manner: the operator kA puts the element $k(Ax)$ of the space E_1 into correspondence with the element $x \in E$.

It is easy to verify that with respect to the operations of addition and multiplication by a scalar introduced above the bounded linear operators form a linear space. If we introduce the norm of the operator in the way indicated above we can form a *normed* linear space.

EXERCISE. Prove that the space of bounded linear operators which transform the space E_1 into a complete space E_2 is complete.

3. *The inverse operator.*

Let us consider the operator T which transforms the Banach space E into the Banach space E_1:

$$Tx = y, \qquad x \in E, \qquad y \in E_1.$$

DEFINITION 7. The operator T is said to have an *inverse* if for every $y \in E_1$ the equation

$$Tx = y \tag{7}$$

has a unique solution.

To each $y \in E_1$ we can put into correspondence the solution of equation (7). The operator which realizes this correspondence is said to be the *inverse* of T and is denoted by T^{-1}.

THEOREM 5. *The operator* T^{-1} *which is the inverse of the linear operator* T *is also linear.*

Proof. To prove the linearity of T^{-1} it is sufficient to verify that the equality

$$T^{-1}(\alpha_1y_1 + \alpha_2y_2) = \alpha_1T^{-1}y_1 + \alpha_2T^{-1}y_2$$

is valid. Denote Tx_1 by y_1 and Tx_2 by y_2. Since T is linear, we have

$$T(\alpha_1x_1 + d_2x_2) = \alpha_1y_1 + \alpha_2y_2. \tag{8}$$

By the definition of the inverse operator,

$$T^{-1}y_1 = x_1, \qquad T^{-1}y_2 = x_2;$$

whence, multiplying these equations by α_1 and α_2 respectively, and adding, we obtain:

$$\alpha_1T^{-1}y_1 + \alpha_2T^{-1}y_2 = \alpha_1x_1 + \alpha_2x_2.$$

On the other hand, from (8) and from the definition of the inverse operator it follows that

$$\alpha_1x_1 + \alpha_2x_2 = T^{-1}(\alpha_1 y_1 + \alpha_2y_2)$$

which together with the preceding equalities yields

$$T^{-1}(\alpha_1y_1 + \alpha_2y_2) = \alpha_1T^{-1}y_1 + \alpha_2T^{-1}y_2.$$

THEOREM 6. *If* T *is a bounded linear operator whose inverse* T^{-1} *exists, then* T^{-1} *is bounded.*

We shall need the following two lemmas in the proof of this theorem.

LEMMA 1. Let M be an everywhere dense set in the Banach space E. Then an arbitrary element $y \in E$, $y \neq 0$, can be developed in the series

$$y = y_1 + y_2 + \cdots + y_n + \cdots,$$

where $y_k \in M$ and $\| y_k \| \leq 3 \| y \|/2^k$.

Proof. We construct the sequence of elements y_k in the following way: we choose y_1 so that

$$\| y - y_1 \| \leq \| y \|/2. \tag{9}$$

This is possible because inequality (9) defines a sphere of radius $\| y \|/2$ with center at the point y, whose interior must contain an element of M (since M is everywhere dense in E). We choose $y_2 \in M$ such that $\| y - y_1 - y_2 \| \leq \| y \|/4$, y_3 such that $\| y - y_1 - y_2 - y_3 \| \leq \| y \|/8$, and in general, y_n such that $\| y - y_1 - \cdots - y_n \| \leq \| y \|/2^n$.

Such a choice is always possible because M is everywhere dense in E. By construction of the elements y_k,

$$\| y - \sum_{k=1}^{n} y_k \| \to 0 \quad \text{as} \quad n \to \infty,$$

i.e. the series $\sum_{k=1}^{\infty} y_k$ converges to y.

To evaluate the norms of the elements y_k we proceed as follows:

$$\| y_1 \| = \| y_1 - y + y \| \leq \| y_1 - y \| + \| y \| = 3 \| y \|/2,$$

$$\| y_2 \| = \| y_2 + y_1 - y - y_1 + y \|$$

$$\leq \| y - y_1 - y_2 \| + \| y - y_1 \| \leq 3 \| y \|/4.$$

Finally, we obtain

$$\| y_n \| = \| y_n + y_{n-1} + \cdots + y_1 - y + y - y_1 - \cdots - y_{n-1} \|$$

$$\leq \| y - y_1 - \cdots - y_n \| + \| y - y_1 - \cdots - y_{n-1} \| = 3 \| y \|/2^n.$$

This proves Lemma 1.

LEMMA 2. If the Banach space E is the sum of a denumerable number of sets: $E = \bigcup_{n=1}^{\infty} M_n$, then at least one of these sets is dense in some sphere.

Proof. Without loss of generality, we can assume that

$$M_1 \subseteq M_2 \subseteq M_3 \subseteq \cdots.$$

We shall assume that all the sets M_i are nowhere dense, i.e. that in the interior of every sphere there exists another sphere which does not contain a single point of M_k, $k = 1, 2, \cdots$.

Take an arbitrary sphere S_0; in it there exists a sphere S_1 which does not contain a single point of M_1; in S_1 there exists a sphere S_2 which does not contain a single point of M_2; and so forth. We obtain a sequence of nested spheres which can be chosen so that the radius of the sphere S_n converges to zero as $n \to \infty$. In a Banach space such spheres have a common point. This point is an element of E but it does not belong to any of the sets M_n. This contradicts the hypothesis of the lemma and proves Lemma 2.

Proof of Theorem 6. In the space E_1 let us consider the sets M_k, where M_k is the set of all y for which the inequality $\| T^{-1}y \| \leq k \| y \|$ holds.

Every element of E_1 is contained in some M_k, i.e. $E_1 = \bigcup_{n=1}^{\infty} M_n$. By Lemma 2, at least one of the M_n is dense in some sphere S_0. In the interior of the sphere S_0 let us consider the spherical shell P consisting of the points z for which

$$\beta < \| z - y_0 \| < \alpha,$$

where

$$0 < \beta < \alpha, \qquad y_0 \in M_n.$$

If we translate the spherical shell P so that its center coincides with the origin of coordinates, we obtain the spherical shell P_0.

We shall show that some set M_n is dense in P_0. Consider $z \in P$; then $z - y_0 \in P_0$. Furthermore, let $z \in M_n$. By virtue of the choice of z and y_0 we obtain:

$$\| T^{-1}(z - y_0) \| \leq \| T^{-1}z \| + \| T^{-1}y_0 \|$$

$$\leq n(\| z \| + \| y_0 \|) \leq n(\| z - y_0 \| + 2 \| y_0 \|)$$

$$= n \| z - y_0 \| [1 + 2 \| y_0 \|/\| z - y_0 \|] \leq n \| z - y_0 \| (1 + 2 \| y_0 \|/\beta).$$

The quantity $n(1 + 2 \| y_0 \|/\beta)$ does not depend on z. Denote $n(1 + 2 \| y_0 \|/\beta)$ by N. Then by definition $z - y_0 \in M_N$ and M_N is dense in P_0 because M_N is obtained from M_n, as was P_0 from P, by means of a translation by y_0 and M_n is dense in P. Consider an arbitrary element y in E_1. It is always possible to choose λ so that $\beta < \| \lambda y \| < \alpha$. For λy we can construct a sequence $y_k \in M_N$ which converges to λy. Then the sequence $(1/\lambda)y_k$ converges to y. (It is obvious that if $y_k \in M_N$, then $(1/\lambda)y_k \in M_N$ for arbitrary real $1/\lambda$.)

We have proved that for arbitrary $y \in E_1$ a sequence of elements of M_N can be found which converges to y, i.e. that M_N is everywhere dense in E_1.

Consider $y \in E_1$; by Lemma 1, y can be developed in a series of elements in M_N:

$$y = y_1 + y_2 + \cdots + y_n + \cdots,$$

where $\| y_n \| < 3 \| y \|/2^n$.

Consider in the space E the series formed from the inverse images of the y_k, i.e. from $x_k = T^{-1}y_k$:

$$\sum_{k=1}^{\infty} x_k = x_1 + x_2 + \cdots + x_n + \cdots.$$

This series converges to some element x since the following inequality holds: $\| x_n \| = \| T^{-1}y_n \| \leq N \| y_n \| < 3N \| y \|/2^n$ and consequently, $\| x \| \leq \sum_{k=1}^{\infty} \| x_k \| \leq 3N \| y \| \sum_{n=1}^{\infty} (\frac{1}{2})^n = 3 \| y \| N$.

By virtue of the convergence of the series $\sum_{n=1}^{\infty} x_n$ and the continuity of T we can apply T to the series. We obtain:

$$Tx = Tx_1 + Tx_2 + \cdots = y_1 + y_2 + \cdots = y,$$

whence $x = T^{-1}y$. We have

$$\| x \| = \| T^{-1}y \| \leq 3N \| y \|$$

and since this estimate is valid for arbitrary y, it follows that T^{-1} is bounded.

THEOREM 7. *An operator which closely approximates an operator whose inverse exists has an inverse, i.e. if T_0 is a linear operator which has an in-*

verse and which maps the space E into the space E_1, and ΔT is an operator which also maps E into E_1, where $\| \Delta T \| < 1/\| T_0^{-1} \|$, then the operator $T = T_0 + \Delta T$ has an inverse.

Proof. Let $y \in E_1$. We wish to find a unique $x_0 \in E$ such that

$$y = Tx_0 = T_0x_0 + \Delta Tx_0 .$$

If we apply the operator T_0^{-1} to this equation, we obtain

$$T_0^{-1}y = x_0 + T_0^{-1}\Delta Tx_0 ; \tag{10}$$

if we denote $T_0^{-1}y$ by $z \in E$ and $T_0^{-1}\Delta T$ by A, then equation (10) can be written in the form

$$z = x_0 + Ax_0 ,$$

where A is an operator which maps the Banach space E into itself and $\| A \| < 1$.

The mapping $x' = z - Ax$ is a contraction mapping of the space E into itself; consequently, it has a unique fixed point which is the unique solution of equation (10) and this means that the operator T has an inverse.

THEOREM 8. *The operator which is the inverse of the operator $T = I - A$, where I is the identity operator and the operator A (of E into E) has norm less than 1 ($\| A \| < 1$), can be written in the form*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k . \tag{11}$$

Proof. Consider the following transformation of the space E into itself:

$$y = Tx, \qquad x \in E, \qquad y \in E.$$

The mapping $x' = y + Ax$ is a contraction mapping of the space E into itself by virtue of the condition that $\| A \| < 1$.

We solve the equation $x = y + Ax$ by means of the iterations: $x_{n+1} = y + Ax_n$. If we set $x_0 = 0$, we obtain $x_1 = y$; $x_2 = y + Ay$; $x_3 = y + Ay + A^2y$; $\cdots$; $x_n = y + Ay + A^2y + \cdots + A^{n-1}y$.

As $n \to \infty$, x_n tends to the unique solution of the equation $x = y + Ax$, i.e. $x = \sum_{k=0}^{\infty} A^k y$, whence

$$(I - A)^{-1}y = \sum_{k=0}^{\infty} A^k y,$$

which yields equation (11).

4. *Adjoint operators.*

Consider the linear operator $y = Ax$ which maps the Banach space E into the Banach space E_1. Let $g(y)$ be a linear functional defined on E_1, i.e. $g(y) \in \bar{E}_1$. Apply the functional g to the element $y = Ax$; $g(Ax)$, as is easily verified, is a linear functional defined on E; denote it by $f(x)$. The functional $f(x)$ is thus an element of the space $\bar{E}$. We have assigned to each

functional $g \in \bar{E}_1$ a functional $f \in \bar{E}$, i.e. we have obtained an operator which maps $\bar{E}_1$ into $\bar{E}$. This operator is called the *adjoint* operator of the operator A and is denoted by A^* or by $f = A^*g$.

If we use the notation (f, x) for the functional $f(x)$, we obtain $(g, Ax) = (f, x)$, or $(g, Ax) = (A^*g, x)$. This relation can be taken for the definition of the adjoint operator.

EXAMPLE. *The expression for the adjoint operator in finite-dimensional space.* Euclidean n-space E^n is mapped by the operator A into Euclidean m-space E^m. The operator A is given by the matrix (a_{ij}).

The mapping $y = Ax$ can be written in the form of the system

$$y_i = \sum_{j=1}^{n} a_{ij}x_j , \qquad i = 1, 2, \cdots, m.$$

The operator $f(x)$ can be written in the form

$$f(x) = \sum_{j=1}^{n} f_j x_j .$$

The equalities

$$f(x) = g(Ax) = \sum_{i=1}^{m} g_i y_i = \sum_{i=1}^{m} \sum_{j=1}^{n} g_i a_{ij} x_j = \sum_{j=1}^{n} x_j \sum_{i=1}^{m} g_i a_{ij}$$

imply that $f_j = \sum_{i=1}^{n} g_i a_{ij}$. Since $f = A^*g$, it follows that the operator A^* is given by the transpose of the matrix for the operator A.

We shall now list the basic properties of adjoint operators.

1. The adjoint operator of the sum of two linear operators is equal to the sum of the adjoint operators:

$$(A + B)^* = A^* + B^*.$$

Let $f_1 = A^*g$, $f_2 = B^*g$, or $f_1(x) = g(Ax)$, $f_2(x) = g(Bx)$; then $(f_1 + f_2)(x) = g(Ax + Bx) = g[(A + B)x]$, whence $(A + B)^* = A^* + B^*$.

2. The adjoint operator of the operator kA, where k is a scalar multiplier, is equal to the adjoint operator of A, multiplied by k:

$$(kA)^* = kA^*.$$

The verification of this property is elementary and is left to the reader.

3. $I^* = I$, i.e. the adjoint of the identity operator on E is the identity operator on $\bar{E}$.

THEOREM 9. *The operator A^*, the adjoint of a linear operator A which maps the Banach space E into the Banach space E_1, is also linear and $\| A^* \| = \| A \|$.*

Proof. The linearity of the operator A^* is obvious. We shall prove the equality of the norms. By virtue of the properties of the norm of an operator we have:

$$| f(x) | = | g(Ax) | \leq \| g \| \| Ax \| \leq \| g \| \| A \| \| x \|,$$

whence $\| f \| \leq \| A \| \| g \|$ or $\| A^*g \| \leq \| A \| \| g \|$ and consequently,

$$\| A^* \| \leq \| A \|. \tag{12}$$

Let $x \in E$ and form $y_0 = Ax/\| Ax \| \in E_1$; it is clear that $\| y_0 \| = 1$. From the corollary to the Hahn-Banach theorem, there exists a functional g such that $\| g \| = 1$ and $g(y_0) = 1$, i.e. $g(Ax) = \| Ax \|$.

From the inequalities $\| Ax \| = | (g, Ax) | = | (A^*g, x) | \leq \| A^*g \| \| x \| \leq \| A^* \| \| g \| \| x \| = \| A^* \| \| x \|$ we obtain $\| A \| \leq \| A^* \|$ which, combined with inequality (12), yields $\| A \| = \| A^* \|$.

ADDENDUM TO CHAPTER III

Generalized Functions

In a number of cases in analysis and in its various applications, for example in theoretical physics, the need arises to introduce various "generalized" functions in addition to the "ordinary" functions. A typical example of this is the well-known δ-function which we have already mentioned above (§23, Example 4).

We wish to emphasize, however, that these concepts, which are discussed briefly in this addendum, did not in any sense originate in an attempt to generalize the concepts of analysis merely for the sake of generalizing. Rather, they were suggested by perfectly concrete problems. Moreover, essentially the same concepts were used by physicists for quite some time before they attracted the attention of mathematicians.

The method of introducing generalized functions which we shall use below originated in the work of S. L. Sobolev, published in 1935–36. Later, these ideas were developed in a somewhat extended form by L. Schwartz.

Consider on the real line the set D of functions $\varphi(x)$ each of which vanishes outside some interval (where for each φ there is a corresponding interval) and has derivatives of all orders. The elements of D can be added and multiplied by a scalar in the usual way. Thus, D is a linear space. We shall not introduce a norm into this space; however, in D one can define in a natural way the convergence of a sequence of elements. We shall say that $\varphi_n \to \varphi$ if: 1) there exists an interval in the exterior of which all φ_n and φ are equal to zero and 2) the sequence of derivatives $\varphi_n^{(k)}$ of order k ($k = 0, 1, 2, \cdots$) (where the derivative of order zero is understood to be as usual the function itself) converges uniformly to $\varphi^{(k)}$ on this interval. The fact that this concept of convergence is not connected with any norm does not give rise to any inconveniences.

We now introduce the concept of generalized function.

DEFINITION 1. *A generalized function* (with values on the real line $-\infty < t < \infty$) is any linear functional $T(\varphi)$ defined on the space D. Thus, $T(\varphi)$ satisfies the following conditions:

1. $T(\alpha\varphi_1 + \beta\varphi_2) = \alpha T(\varphi_1) + \beta T(\varphi_2)$;
2. If $\varphi_n \to \varphi$ (in the sense indicated above), then $T(\varphi_n) \to T(\varphi)$.

We now consider several examples.

1. Let $f(t)$ be an arbitrary continuous function of t. Then, since every function $\varphi(t) \in D$ vanishes outside some finite interval, the integral

$$T(\varphi) = \int_{-\infty}^{\infty} f(t)\varphi(t)\, dt \tag{1}$$

exists for all $\varphi \in D$. The expression (1) represents a linear functional on