THEORETICAL FOUNDATIONS OF THE POTENTIAL FUNCTION METHOD IN PATTERN RECOGNITION LEARNING

(UDC 62-507)

M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 25, No. 6,
pp. 917-936, June, 1964

Algorithms are proposed for teaching automata to recognize classes of input functions based on the construction of the so-called potential functions. A basic hypothesis is introduced concerning the character of the functions distinguishing the ensembles corresponding to various classes of input situations. Using this hypothesis, theorems are proved on the convergence of the algorithm in a finite number of steps.

It is shown that the proposed algorithms may be realized by a broad class of circuits. The characteristics of the elements of which the circuits are constructed are practically arbitrary. It is shown that the Rosenblatt perceptron is related to this class of circuit, i.e., it is proven that the operation of the perceptron can be considered to be a realization of the potential function method. In this connection the theorems demonstrated on the convergence of the method of potential functions also solve the problem of convergence of the perceptron process.

1. Statement of the Problem

1. Partitioning of the input situations into classes. Let us consider an automaton receiving signals from the outside. This input information may contain logical variables (for example, the replies "yes" or "no" to certain questions), continuous or discrete variables (for example, measurement or computation results obtained from analog or digital computers) or even continuous functions * (for example, the characteristics of a system under investigation). The set of information applied in any form to the input of the automaton constitutes the input situation. Input situations can be divided into several classes. The purpose of the automaton is to determine the class to which each newly arising input situation belongs.

Such problems arise, for example, in the development of machines for technical diagnosis (fault detection) or medical diagnosis, in automation of geophysical prospecting based on seismic soundings or electrical core sampling, in the development of optical pattern recognition devices (for example, alphabetic or numerical characters written in different handwritings or printed in different fonts), in recognition of aircraft or ship types from their sound, etc.

In each of the above examples the automaton for classifying the input signals is important in itself. Frequently such automata serve as the primary element of a complex automation system. Indeed, the very problem of automatic control as a whole can be considered to be the problem of assigning the input situation to one or another class, and to generate the optimal response as a function of that class.

2. Teaching by examples. In order that the automaton assign each newly arising input situation to the corresponding class, the simplest procedure would be first to introduce into the automaton an exact description of the "class boundaries," i.e., the rules for the conduct of the automaton. Frequently, however, it is impossible or difficult to program the process for classifying the input signals, regardless of the fact that the person designing the automaton knows himself how to distinguish the classes. This is connected with the fact that the human is often able to distinguish and to recognize classes due to accumulated experience and intuition, but not able to give exact instructions –

*For example, in machines for medical diagnosis the inputs can be temperature curves, cardiograms, encephalograms, etc.; in pattern-recognition machines different types of scans of the image shown to the machine can be introduced, etc.

a program how to do this, to another person lacking experience. Nevertheless, an experienced person can teach an inexperienced one to recognize these classes, i.e., can transmit his experience to him. This is not done by giving instructions, but by showing examples. Thus, in demonstrating examples of machine noise for various defects, it is possible to teach an inexperienced person to distinguish defects by their noise, although each time the new noise differs somewhat from that which was demonstrated during teaching; similarly we teach young doctors to listen to heart or breathing sounds; a teacher can teach the pupil to distinguish letters, showing examples only of their shapes, etc. We shall term such processes teaching by examples.*

Accordingly, by the expression teaching an automaton by examples to distinguish classes we shall have in view the following process. No indications are introduced into the automaton in any form as to the rules or features to be used for classifying the input situations. After the machine has been built, a certain time will be taken up by the teaching process. During this process the only information introduced into the automaton each time a certain input situation has arisen will be the class to which the situation belongs.

After this teaching process has stopped, when these same or new situations appear at the input the automaton must recognize the classes to which they belong (test).

A feature of the problem consists in the fact that during the teaching process a finite (and relatively small) number of situations is presented to the automaton, yet after teaching the machine must know the rules for classifying the infinite (or very large) number of situations that may appear during the test process. This very fact excludes a trivial solution to the problem, the simple memorizing of the situations which have appeared; the design of the automaton should provide for the "extrapolation" of the information obtained in the learning process to new situations which have not appeared at the input during that process.

In this statement of the problem the automaton should classify input situations although (before beginning the teaching process) it was not known exactly which classification would be carried out. For example, in the recognition of visual patterns a given machine should learn to distinguish various numbers or letters of the alphabet, or photographs of individuals, etc. The particular classification to be carried out in a given concrete experiment is defined only by the sequence of situations presented during the learning process. In this sense the automaton capable of learning to distinguish classes must be "universal."

3. Geometrical interpretation of the problem. The method of potential functions. Without loss of generality we shall consider below only automata learning to distinguish two classes: classes A and B.

Let us introduce the input space X, constructed so that to each input situation there correspond in a one-to-one relationship a point of this space. †

By definition the classes A and B do not intersect. This signifies that in the space X there exists at least one separation function $\Psi(x)$ taking on positive values at points corresponding to the class A and negative values at the points corresponding to the class B. The values of $\Psi(x)$ at other points are immaterial. In the general case there can exist many such separation functions.

During the teaching process points in the space X appear successively and information is given as to the class, A or B, to which these points belong. The problem consists in constructing from only this information, some one of the separation functions on the basis of a finite number of examples. Then in the test process the machine can assign the points which appear to the classes A and B according to the sign of the separation function at these points.

The method of solution proposed below is connected with the following procedure. When a certain point x^k appears during teaching, a function $K(x, x^k)$, defined over the entire space X, and depending on x^k as a parameter ("potential function"), is connected with the point. The sequence of potential functions $K(x, x^1)$, $K(x, x^2)$, ... corresponding to the sequence of points $x^1, x^2, \ldots$, appearing during the teaching process, is used to construct the function

* It appears that teaching by examples plays an essential role even in the process of teaching theoretical thinking to a person. Thus, a mathematician can demonstrate new theorems not because he has an universal algorithm how to do this, but rather because, beginning from his school days examples of theorem demonstration have been shown to him.

† If we introduce n logical variables, the space consists of the vertices of the n-dimensional cube. In the presence of m continuous variables, it is the n-dimensional euclidian space. Finally, in the presentation to the automaton of functional relations, functional spaces enter into consideration.

$$\Psi^*(x, x^1, x^2, \ldots).$$

using the rules to be obtained below.

These rules are established so that Ψ^* tends to one of the separation functions as the number of points x^k increases during the teaching process.

A procedure for the successive construction of the separation function from the functions generated by the example points is termed in this paper the method of potential functions.

4. Basic assumptions. The statement of the problem of teaching automata is without sense if no limitations are placed on the set of situations which the automaton is to classify. In effect, in this latter case, regardless of the operating algorithm of the automaton and the separation function generated after the appearance of a finite sequence of points, it would always be possible to name points not yet shown so that the automaton would always err at these points during the test process. It is therefore first necessary to restrict the choice of space X and the class of functions $\Psi(x)$ in a suitable manner. It might be thought that these restrictions must be very severe. From the results of the present work it will be evident that this is not the case, and that the problem posed can be solved under restrictions which cause no practical difficulties. These restrictions are formulated below.

In the entire following discussion we shall assume the existence in the space X of a system of functions $\varphi_i(x)$ $(i = 1, 2, \ldots)$ such that for each pair of separable sets a number N is found (in general different for each pair) for which the separation function can be represented in the form

$$\Psi(x) = \sum_1^N c_i \varphi_i(x). \qquad (!)$$

If in the space X there exists in some class R a complete system of functions, the $\varphi_i(x)$ can be considered to be elements of this system, and any function in R (including each separation function) may be represented in the form of an infinite series The condition (!) requires that the separation functions be expandable in series with finite numbers of elements.*

In the algorithm proposed below the form of the potential function depends essentially on the choice of system $\varphi_i(x)$. It would appear that this signifies that in using the method of potential functions it is necessary to know exactly that system $\varphi_i(x)$ in which the separation function can be expanded. This in turn would signify that to use the algorithm it is necessary to know the characteristics of the sets to be classified, which would practically satisfy assumption (!). In this connection, although formally assumption (!) is sufficient to solve the problem, below we present additional considerations permitting use of the method of potential functions with practically arbitrary choice of the system $\varphi_i(x)$ and, hence, of the form of the potential function.

Let there be defined in the space X the scalar product of two functions $\alpha(x)$, $\beta(x)$ (for example, the integral $\int \bar{\alpha}(x)\beta(x)\,dx$, where the bar indicates the complex conjugate). Then the complete system $\varphi_1(x), \varphi_2(x), \ldots,$ $\varphi_i(x), \ldots$ in which any separation function may be expanded may be considered to be orthonormal, without loss of generality. If in a certain space and for a certain class of functions there exists such a complete orthonormalized system, it is not unique, and many such systems can be introduced.

Mainly in connection with problems of mathematical physics certain special systems of orthonormalized functions (for example, trigonometric functions, the functions of Laguerre, Hermite, Lagrange, etc.) have been introduced and have received general application; they have the following three distinguishing characteristics: a) the number of zeros and the number of extrema of these functions in a finite interval ("oscillation," "frequency") increase monotonically with increase in the number of terms; b) the functions usually encountered in physics are approximated sufficiently well by a finite, and in general small number of elements of the expansion in these systems of functions

*The basic proposition demonstrated below (theorem 1) is true for the broader assumptions that the separation function is representable by an infinite series $\sum_{i=1}^{\infty} c_i \varphi_i(x)$ or even by an integral of the form $\int_\Omega c_\omega \varphi_\omega(x)\,d\omega$. In fact this does not however broaden assumption (!) (for details see remark 2 after the demonstration of theorem 1).

("contain small numbers of harmonics"); c) any function in these systems with a small number of terms can be well approximated by a finite and,generally,a relatively small number of terms in an expansion in any other such system.

We shall term a system of functions having these three properties "ordinary." We shall term "well expandable" any function which with sufficient precision is representable by a finite and relatively small number of terms of a series expansion in any ordinary system of functions. "Good expandability" is closely connected with the properties of "smoothness" of a function, since the functions of an "ordinary" system are ordered in such a manner that the "smoother" functions have lower indices. Due to the properties of "ordinary" systems it is practically unimportant in which of these systems a well expandable, i.e., sufficiently smooth function is represented.

It is assumed about the space X that a similar situation obtains, i.e., that the concept of "ordinary" complete systems can be introduced in X.

The basic assumption of the present work is that for each pair of distinguishable sets A and B there exists a "well expandable" separation function, i.e., there exists a separation function representable by the expansion (!), where $\varphi_1(x), \varphi_2(x), \ldots, \varphi_i(x) \ldots$ is an "ordinary" system. From this it follows that in the organization of the potential function it is immaterial which of the "ordinary" systems is used, i.e., it is not necessary to carry out in advance a detailed study of the problem.

It is to be understood that the concepts of "ordinary system of functions" and "well expandable" functions are not precise concepts. The basic assumption only emphasizes that the separation functions are not very "jagged" or "figured" in X, i.e., do not have a "large number of extrema" in a small region, at nearby points their values usually "differ little," etc.

Many articles have been devoted to the problem of teaching automata to distinguish classes (mainly from the aspect of speed and visual pattern recognition [1-6]). However, the authors of these papers limit themselves to the description of the algorithms proposed by themselves for solution of the problem or of devices constructed for the purpose, and in the best case present examples of pattern recognition teaching by these means. Certain of the proposed algorithms can be considered to be algorithms for construction of potential functions in one or another space of points shown during the teaching process,* and one of them [3] is directly interpreted by the author in these terms. However, none of the publications contains a sharp formulation of the restrictions placed on the concepts of "class" or "pattern"† (i.e., on the choice of space and the character of the sets subject to separation), which, while in practice not restricting the problem, would permit formulation and demonstration of theorems on the convergence of the algorithms in exact terms, i.e., there are no demonstrations that for a broad class of sets the teaching algorithm,after a finite number of examples,will adequately separate the corresponding sets.

An interesting result in this direction has been obtained for the perceptron algorithm by A. Novikoff [7] who, although he did not demonstrate the convergence of thealgorithm in the above sense, showed that for any periodic teaching sequence the perceptron will not commit errors on the elements of the sequence after termination of teaching, if only these elements belong to the sets which can in principle be distinguished by the perceptron. The method ofproof of Novikoff's theorem is used here for the proof of theorem 1.

5. Linearization space. Using the basic concept and starting from (!) we can introduce into consideration an N-dimensional space Z into which the initial space X is mapped. Namely, to each point $x \in X$ we assign a point $z \in Z$ with coordinates $z_i = \varphi_i(x)$ $(i = 1, \ldots, N)$. By virtue of (!) the separation function $\Psi(x)$ in Z is mapped into the linear function $\sum_1^N c_k z_k$. Since

$$\Psi(x) = \sum_1^N c_k z_k \begin{cases} > 0, & x \in A, \\ < 0, & x \in B, \end{cases}$$

*For example, this will be shown for perceptrons in section 4.

†In [6] it was proposed to use the intuitive concept of the "compactness hypothesis" with respect to visual patterns, which in the above terms can be formulated as the assumption of "good expandability" of the separation function directly in the receptor space.

the points belonging to different classes are separated in Z by the hyperplane

$$\sum_{1}^{N} c_k z_k = 0.$$

Since functions expanded in the system $\varphi_i(x)$ (i = 1, ..., N) are linearized in Z, we shall term this a linearization space.

Further, in solving the problem it is frequently found convenient to carry out the analysis in the linearization space, where we sometimes consider the infinite-dimensional space obtained by completing the system of functions $\varphi_1(x), \ldots, \varphi_N(x)$ to a complete system.

2. Algorithm

For the potential function we shall take a function of two variables of the form

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i^2 \varphi_i(x)\, \varphi_i(y), \tag{1}$$

where $\varphi_i(x)$ (i = 1, 2, ...) is a linearly independent system of the functions discussed above in the formulation of the basic assumptions, λ_i are real numbers different from zero for i = 1, 2, ..., N. Further we shall assume that the $\varphi_i(x)$ and the function K(x, x) are bounded for $x \in A \cup B$. The variable y will be identified below with points appearing in the learning process.

Let there appear during the learning process the points $x^1, x^2, \ldots, x^k, \ldots$, each of which belongs to A or B. We assign arbitrarily to these sets the signs + and −, i.e., we name, for example, the set A positive and the set B negative.

At the appearance of the first point x^1 we construct the function $K_1(x)$, equal to the potential of x^1 taken with the sign of the set to which x^1 belongs, i.e.,

$$K_1(x) = \begin{cases} K(x, x^1), & \text{if} \quad x^1 \in A, \\ -K(x, x^1), & \text{if} \quad x^1 \in B. \end{cases}$$

We shall explain the further operation of the algorithm by induction. Let after the rth example the potential $K_r(x)$ be constructed. Let further the point x^{r+1} appear in the following, (r+1)st step of teaching. Then four cases are possible:

$$x^{r+1} \in A, \qquad K_r(x^{r+1}) > 0, \tag{a}$$

$$x^{r+1} \in B, \qquad K_r(x^{r+1}) < 0, \tag{b}$$

$$x^{r+1} \in A, \qquad K_r(x^{r+1}) < 0, \tag{c}$$

$$x^{r+1} \in B, \qquad K_r(x^{r+1}) > 0. \tag{d}$$

In cases (a) and (b) the sign of the set to which the point x^{r+1} belongs and the sign of $K_r(x^{r+1})$ coincide, i.e., "there is no error." In these cases we take

$$K_{r+1}(x) = K_r(x).$$

In cases (c) and (d) there are errors, i.e., the sign of the set to which x^{r+1} belongs and the sign of $K_r(x^{r+1})$ do not coincide. Then the "error is corrected," i.e., in case (c) we take

$$K_{r+1}(x) = K_r(x) + K(x, x^{r+1}),$$

and in case (d)

$$K_{r+1}(x) = K_r(x) - K(x, x^{r+1}).$$

This basic step of the algorithm can be explained as follows: at the appearance of the (r+1)st point the "hypothesis is adopted" that the sign of the potential constructed after the rth point separates the set, i.e., that the function $K_r(x)$ is the required separation function; this hypothesis is tested at the (r+1)st point; if it is found valid, the

potential is not altered at this step, i.e., "the hypothesis is conserved" for the following step, the (r+2)nd point; otherwise the potential is altered by addition to it of the potential of the (r+1)st point with the sign necessary to "correct" the function.

Clearly after r steps the potential can be written in the following form:

$$K_r(x) = \sum_{x^s \in A}' K(x, x^s) - \sum_{x^q \in B}' K(x, x^q). \tag{2}$$

Here the lower indices on the summation signs signify that summation is carried out only over the points shown in r teaching steps belonging to sets A and B respectively, while the primes signify that only those x^s in A (x^q in B) are taken into account whose substitution in the preceding potential "caused error," i.e., gave a sign not agreeing with the sign of the set to which x^s belongs.

We shall now give a convenient interpretation to the algorithm, using for this purpose the process in the infinite-dimensional linearization space Z, with axes $z_i = \lambda_i \varphi_i(x)$ (i = 1, 2, ...). In Z to each point $x \in X$ there correspond two nonintersecting sets in the linearization space; we assign to each the same designation.

If in X there exists a separation function, representable by the expansion

$$\Psi(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x)$$

(according to the basic assumption $c_i = 0$ for $i > N$) such that

$$\Psi(x) \begin{cases} > 0, & \text{if} \quad x \in A, \\ < 0, & \text{if} \quad x \in B, \end{cases}$$

then there exists in the linearization space a separation plane passing through the origin with normal vector γ

$$(\gamma, z) \equiv \sum_{i=1}^{\infty} \gamma_i z_i = 0,$$

where $\gamma_i = c_i/\lambda_i$ is such that

$$(\gamma, z) \begin{cases} > 0, & \text{if} \quad x \in A \\ < 0, & \text{if} \quad x \in B. \end{cases}$$

Let us reflect the set B symmetrically about the origin, i.e., we substitute $-z$ for all vectors $z \in B$. The image set thus obtained we term B' and we consider the union $S = A \cup B'$ (Fig. 1, heavy line).

The condition of separability of sets A and B by a plane with normal vector γ is now written in the form

$$(\gamma, z) \equiv \sum_{i=1}^{\infty} \gamma_i z_i > 0 \quad \text{for} \quad z \in S, \tag{3}$$

i.e., the sets A and B are separated by this plane if the region S lies to one side of it, and vice versa.

Let now to the sequence of points M

$$x^1, x^2, \ldots, x^r, \ldots$$

in X belonging to the sets A and B correspond the sequence M* of points $z^1, z^2, \ldots z^r$ in $S = A \cup B'$ in the linearization space.

The function K(x, y) defined according to (1) can be interpreted in the linearization space Z as the scalar product of two vectors z and u with coordinates $z_i = \lambda_i \varphi_i(x)$ and $u_i = \lambda_i \varphi_i(y)$,

$$K(x, y) = (z, u) \tag{4}$$

Formula (2), taking into account (4) and the definition of M*, can now be rewritten in the form

$$K_r(z) = \sum_{z^q \in M^*}' (z, z^q), \tag{5}$$

where Σ' signifies summation over those points of the sequence M* whose appearance in the teaching process led to "correction of an error." We now remove from M* all points which did not lead to "correction of an error," and we leave those points required for "error correction," renumbering them in order $z^1, z^2, \ldots$ They form a sequence M**. Now expression (5) can be written as follows:

$$K_r(z) = \left(z, \sum_{l=1}^{k_r} z^l\right), \quad z^l \in M^{**}, \tag{6}$$

where k_r is the number of "corrected errors" occurring in the course of the first r examples.

The condition for which "error correction" must be carried out at the point $z \in S$ has the form

$$K_r(z) < 0.$$

Therefore, it follows from equality (6) that the (k+1)st "error correction" occurs if

$$\left(z^{k+1}, \sum_{l=1}^{k} z^l\right) < 0. \tag{7}$$

We shall now describe the algorithm in a "geometric language."

When the first point z^1 in M* appears the application of the algorithm signifies construction in the linearization space of the plane

$$K_1(z) = (z, z^1)$$

with normal vector z^1 (Fig. 2).

If the following point in M* lies in that halfspace to which the normal vector z^1 of the constructed plane is directed, then there is no error; the position of the plane and its normal vector do not change in this case, and the next example is shown. The first time that a point falls in the opposite halfspace, "error correction" occurs, which in this geometric language means the following operation: the normal vector of the plane constructed up to this step is added to the vector of the point requiring the "error correction" and the resultant vector is adopted as the new normal vector of the separation plane and, consequently, the plane itself rotates about the origin of coordinates so as to be perpendicular to the new normal vector. Thus, for example, if error correction is required after the second step, the new normal vector is equal to $z^1 = z^2$ (Fig. 3).

After k error corrections the normal vector of the plane is equal to the sum $\sum_{l=1}^{k} z^l$, $z^l \in M^{**}$, and inequality (7) indicates that the following (k+1)st error correction occurs only if the corresponding point lies in the halfspace opposite the normal vector.

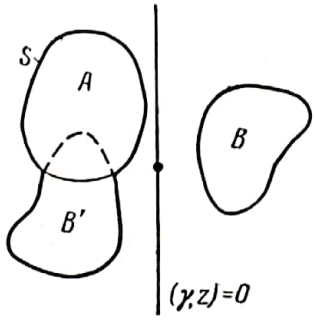

Fig. 1.

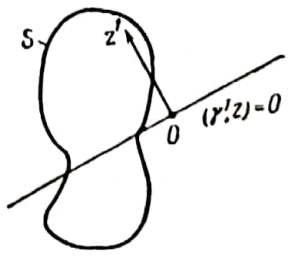

Fig. 2.

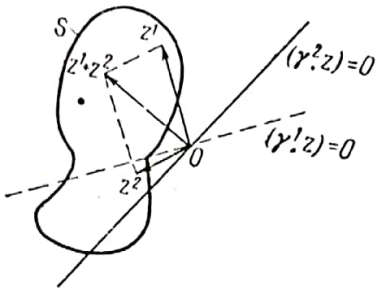

Fig. 3.

The use of the algorithm to construct an automaton for learning to distinguish classes is meant in the following sense. As points are shown to the automaton in the teaching process the machine constructs for the sth step the function $K_s(x)$ according to the algorithm. After a sufficiently long teaching the process is stopped, and the function $K_s(x)$ is adopted as the separation function. In the test process when a new point is presented the automaton computes $K_s(x^*)$ and assigns the point x^* to the class A or B according to the sign of $K_s(x^*)$.

In the geometric language this signifies that after s examples the plane passing through the origin of coordinates in the linearization plane with normal vector $\sum_{l=1}^{k_s} z^l$, $z^l \in M^{**}$ is adopted as the separation plane.

3. Convergence of the Algorithm in a Finite Number of Steps

In this section we shall establish two closely related theorems concerning the algorithm described in section 2: a theorem on the finite number of corrected errors and a theorem on the convergence of the algorithm.

Theorem 1. Let M be an arbitrary infinite sequence of points $x^1, x^2, \ldots, x^k, \ldots$ in the space X, belonging to the sets A and B. Let further there exist a function $\Psi(x)$ rigorously separating the sets A and B, i.e.,

$$\Psi(x) \begin{cases} > \varepsilon, & \text{if} \quad x \in A, \\ < -\varepsilon, & \text{if} \quad x \in B; \end{cases} \tag{8}$$

(where $\varepsilon > 0$), representable by the expansion

$$\Psi(x) = \sum_{i=1}^{N} c_i \varphi_i(x). \tag{9}$$

Let further the function $K(x, x)$ be bounded in $A \cup B$.

Then there exists an integer m independent of the choice of sequence M such that in using the algorithm the number of errors corrected does not exceed m.

Before proceeding to a proof of theorem 1, let us explain it, using the geometric model introduced above. In application to the linearization space theorem 1 states that if there exists a plane such that the entire joint set $S = A \cup B$ lies strictly to one side of it and is bounded, then for any sequence M the algorithm given in section 2 will construct some plane after a finite number of corrected errors such that no further examples taken from the entire infinite "tail" of the sequence will cause the constructed plane to rotate, i. e., no further error correction will occur.

The proof of theorem 1 will be carried out in this "geometric language" applied to the linearization space.

Proof of theorem 1. Let us introduce the following symbolism:

$$a = \inf_{z \in S} \frac{(\gamma, z)}{|\gamma|}, \tag{10}$$

$$b = \sup_{z \in S} |z|. \tag{11}$$

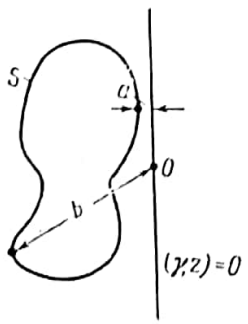

Fig. 4.

According to (3) and (8)

$$a > \frac{\varepsilon}{|\gamma|} > 0.$$

Since

$$|z| = \sqrt{\sum_{i=1}^{\infty} [\lambda_i \varphi_i(x)]^2} = \sqrt{K(x, x)}, \tag{12}$$

$b < \infty$, since $K(x, x)$ is by hypothesis bounded.

In the geometrical language the quantity $\underline{a}$ is equal to the minimum distance from the plane $(\gamma, z) = 0$ to the joint set $S = A \; B'$, and $\underline{b}$ is the distance from the origin to the furthest point of this region (Fig. 4).

From (10) and (11) it follows that

$$(\gamma, z^l) \geqslant a|\gamma|, \tag{13}$$

$$|z^l| \leqslant b. \tag{14}$$

Putting, further,

$$\gamma^k = \sum_{l=1}^{k} z^l, \quad z^l \in M^{**}, \tag{15}$$

i.e., γ^k is the normal vector of the plane after $\underline{k}$ corrected errors. Then inequality (7) may be written:

$$(z^{k+1}, \gamma^k) < 0, \quad z^{k+1} \in M^{**}. \tag{16}$$

Let us examine the change in the normal vector γ^k of the plane constructed by the algorithm. We sum the inequality (13) from l to 1 over $\underline{k}$

$$(\gamma, \gamma^k) \geqslant ka|\gamma|. \tag{17}$$

Using the Cauchy-Bunyakovskii inequality to estimate the left side of (17),

$$|(\gamma, \gamma^k)| \leqslant |\gamma||\gamma^k|.$$

Whence and from (17) after division by $|\gamma|$ we obtain

$$|\gamma^k| > ka. \tag{18}$$

Further, according to the geometrical interpretation of the algorithm

$$\gamma^{k+1} = \gamma^k + z^{k+1}.$$

Therefore

$$|\gamma^{k+1}|^2 = |\gamma^k|^2 + 2(\gamma^k, z^{k+1}) + |z^{k+1}|^2.$$

Using now inequalities (14) and (16) we obtain

$$|\gamma^{k+1}|^2 \leqslant |\gamma^k|^2 + b^2.$$

From this recursive relationship, taking into account that $\gamma^0 = 0$, we find

$$|\gamma^k|^2 \leqslant kb^2. \tag{19}$$

We combine inequalities (18) and (19)

$$k^2a^2 \leqslant |\gamma^k|^2 \leqslant kb^2.$$

Finally we obtain

$$k \leqslant b^2/a^2 = m. \tag{20}$$

Estimate (20) is written "in terms of the linearization space." It can be rewritten "in terms of the space X" if we rewrite in these terms formulas (10) and (11)

$$a = \frac{\inf\limits_{x=A\cup B} \left| \sum\limits_{i=1}^{\infty} c_i \varphi_i(x) \right|}{\sqrt{\sum\limits_{i=1}^{\infty} (c_i^2/\lambda_i^2)}} = \frac{\inf\limits_{x\in A\cup B} |\Psi(x)|}{\sqrt{\sum\limits_{i=1}^{\infty} (c_i^2/\lambda_i^2)}}, \tag{21}$$

$$b = \sup_{x\in A\cup B} \sqrt{K(x, x)}. \tag{22}$$

Therefore

$$k < \left(\frac{\sup_{x \in A \cup B} \sqrt{K(x, x)}}{\inf_{x \in A \cup B} |\Psi(x)|} \sqrt{\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i^2}} \right)^2 \tag{23}$$

and $\underline{k}$ is finite if the series $\sum_{i=1}^{\infty} (c_i^2 / \lambda_i^2)$ converges. This series is known to converge if the basic assumption (:) is satisfied, i.e., if for $i > N$ all $c_i = 0$.

Theorem 1 is thereby proven.

Remark 1. Inequality (23) shows that $\underline{m}$ is smaller if for all other conditions the same, the minimum of function $|\Psi(x)|$ is greater. This reflects the fact that the number of corrected errors decreases as the points of A and B are "further" from the separation plane $\Psi(x) = 0$, i.e., as they are "further" from each other.

2. From the proof of theorem 1 it is easily seen that the theorem remains valid when the separation function is representable by an expansion in an infinite series

$$\Psi(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x) \tag{::}$$

or an expansion of the form

$$\Psi(x) = \int_{\Omega} c_\omega \varphi_\omega(x)\, d\omega, \tag{:::}$$

if the series $\sum_{i=1}^{\infty} (c_i^2 / \lambda_i^2)$ (or the integral $\int_{\Omega} (c_\omega / \lambda_\omega)^2 d\omega$) converges. However, in the proof of the theorem the condition $\inf_{x \in A \cup B} |\Psi(x)| > 0$. is used in an essential manner. At the same time the infinite series (::) and the integral (:::), if they converge uniformly, can be substituted by a finite series of the form (:) such that the function obtained differs by an arbitrarily small quantity from (::) or (:::). From these two facts it follows that a function of the form (:) will also separate. Therefore, in the conditions of the theorem, assumptions on the expandibility of the separation function in (::) or (:::) in fact do not broaden assumption (:).

As important as it is, the proposition of theorem 1 does not yet establish the convergence to the separation function of the $K_r(x)$ constructed by the algorithm. In effect the theorem places no restrictions of any sort on the statistics of the point examples shown in the teaching process. The test result can then contain errors even if an infinite teaching sequence is correctly separated by the automaton (for example, if this sequence consists of an infinite repetition of two points, one of which belongs to A, the other to B). To establish the convergence of $K_r(x)$ to the separation function the sample statistics must be taken into account.

It is obvious that if the automaton is to separate the sets A and B it is necessary that the teaching sequence be "sufficiently representative," its points should be "sufficiently scattered" over the sets A and B. For this it is possible, for example, to require that the points of the teaching sequence appear at random and such that the probability of appearance of a point from any subset of nonzero measure be positive (it is understood that the probability of appearance of points lying outside A and B is equal to zero *). In effect, if this condition is satisfied, for incomplete separation the probability is unity that there will eventually occur the following error correction, and since according to

* For example, for a space consisting of a finite number of points this signifies the existence of a positive probability of appearance of each of the points of the space belonging to A and B; for the n-dimensional euclidian space this signifies that the probability densities in A and B can be zero only in sets smaller than $\underline{n}$, the number of measurements.

theorem 1 there can only be a finite number of these corrections, the separation of the sets will finally take place with a probability of unity. These considerations are confirmed by the following theorem.

Theorem 2. Let the sets in the space X be such that

1) there exists a separation function

$$\Psi(x)\begin{cases}\geq \varepsilon, & \text{if} \quad x \in A,\\ \leq -\varepsilon, & \text{if} \quad x \in B,\end{cases}$$

where $\varepsilon > 0$, representable in the expansion

$$\Psi(x) = \sum_{i=1}^{N} c_i \varphi_i(x);$$

2) the function $K(x, x)$ is bounded for $x \in A \cup B$.

Let further

3) the sample statistics satisfy the following conditions:

a) the appearances of points of the teaching sequence are independent events;

b) for any r, at the rth step of the algorithm there exists a strictly positive probability of correcting an error if at this step complete separation of the sets A and B by the function $K_r(x)$ has not yet occurred.

Then with probability unity for each realization of the algorithm a finite number l is found (in general different for each realization) such that

$$K_l(x)\begin{cases}>0, & \text{if} \quad x \in A,\\ <0, & \text{if} \quad x \in B,\end{cases}$$

i.e., with probability unity the separation of the sets is realized in a finite number of steps.

Before proving the theorem we note that its conditions 1 and 2 ensure applicability of theorem 1, while condition 3 is a condition on the sample statistics which, for example, are guaranteed by the existence of a strictly positive probability of appearance of points from any subset of nonzero measure of the sets A and B. It is easily shown that the proposition of theorem 2 has the following equivalent formulation.

Theorem 2.* In the conditions of theorem 2 there exists for arbitrary ε such an s that the probability of separating the sets A and B at least one of the steps from the 0th to the sth will be greater than $1 - \varepsilon$.

Proof of theorem 2. Let us consider the set of all realizations of the algorithm (corresponding to the set of all teaching sequences). In each realization there exists a last error correction (since from theorem 1 there is only a finite number of these). Let us consider the probability $\Pr(p > \delta)$ that after the last error correction the probability p of appearance of an error on the succeeding step be greater than $\delta \geq 0$*. But the event "after the last error correction $p > \delta$" and the event "for $p > \delta$ no error will occur in any succeeding examples" are identical. The probability of the latter event is equal to zero for each $\delta > 0$, since the probability that errors will not occur in the following L steps is less that $(1 - \delta)^L$, by virtue of the independence of the appearances of the points, and the last expression tends to zero as $L \to \infty$. Consequently $\Pr(p > \delta) = 0$ for any $\delta > 0$, and from this it also follows that $\Pr(p > 0) = 0$. But according to condition 3b of the theorem $\Pr(p > 0)$ is just the probability that separation has not occurred. Therefore, the probability that separation occurs is equal to $1 - \Pr(p > 0) = 1$, Q.E.D.

4. Conditions for Termination of the Algorithm

According to the theorems of the preceding section the algorithm proposed above leads to exact separation of the sets A and B with probability unity for each concrete problem and in each realization in a finite number of steps.

*The condition $p > \delta$ corresponds to the realizations for which, after the last error correction, separation of regions A and B has not yet occurred and the probability of incidence of a point (by virtue of the sample statistics) in that part of the sets A and B which are incorrectly separated, is greater than δ. The probability $\Pr(p > \delta)$ is the measure of this set of realizations, and the aim of the further considerations consists in proving the fact that $\Pr(p > \delta) = 0$. It is assumed that the measure of $\Pr(p > \delta)$ exists. Similar assumptions are also made without explicit statement in the formulation and proof of the following theorems of this work. The conditions guaranteeing the existence of these probabilities are not considered in the present work.

However, it is not possible to give a guarantee that separation of the sets has already occurred at any particular step, regardless of the length of the teaching sequence. Below two variants of the conditions for termination of the algorithm are assumed, for which, although they do not guarantee separation of the sets, the probability of error in the succeeding test is sufficiently small if the statistics of appearance in the teaching process and the test coincide.

Variant 1. Let us supplement the algorithm with the following conditions of termination: the teaching process terminates as soon as no error correction has occurred in the L examples in sequence following an error correction. Here L is an arbitrary prescribed integer. The extended algorithm, by virtue of theorem 1, leads to termination of the teaching process not later than after Lk teaching examples, where k is the maximum number of corrected errors estimated by theorem 1. As soon as the teaching process terminates according to the above condition, the quality of the succeeding test may be guaranteed by the estimate given in the following theorem.

Theorem 3. Let p be the probability of error in the test process carried out after termination of the teaching process. Then, for any $\varepsilon > 0$ and $\delta > 0$, the probability $\Pr(p < \delta)$ of the event $(p < \varepsilon)$ exceeds $1-\delta$ if L satisfies the inequality

$$L > \frac{\ln(\delta / k)}{\ln(1-\varepsilon)}. \tag{24}$$

Proof of theorem 3. Let us consider the event S(s) consisting in the application of the proposed algorithm and the termination condition of a total number of errors corrected not less than a prescribed integer $s \geq 1$. We consider the function $P(w|S)$, $\int_0^1 P(w|S)\,dw = 1$, such that $P(w|S)dw$ signifies the probability of the event "the probability of error in the step after correction of the sth error lies between w and w + dw" on condition that S has occurred.

If at each step the error probability is equal to w, the probability that in the course of L steps in succession an error has not appeared, by virtue of the independence of the examples, is equal to $(1-w)^L$. Therefore the probability that in the course of L examples after the sth error correction there will not occur a new error and that the error probability lies between w and w + dw, is equal to

$$P(w|S)(1-w)^L dw.$$

But by virtue of the termination condition this expression is equal to the probability that the proposed variant of the algorithm leads to termination exactly after s errors corrected, where the error probability in the following test lies between w and w + dw if the event S has occurred. Therefore, the probability P_s that termination occurs after exactly s errors corrected and that the probability of error in the following test is greater than ε is equal to

$$P_{\varepsilon s} = \int_{\varepsilon}^{1} P(w|S)(1-w)^L P(S)\,dw.$$

Termination of the teaching process after different numbers s of errors corrected are incompatible events. Therefore the probability of such a termination that in the succeeding test the error probability exceed ε is equal to

$$P_{\varepsilon} = \sum_{s=1}^{k} P_{\varepsilon s} = \sum_{s=1}^{k} \int_{\varepsilon}^{1} P(w|S)(1-w)^L P(S)\,dw.$$

Let us find an upper bound for this expression:

$$P_{\varepsilon} \leqslant \sum_{s=1}^{k} (1-\varepsilon)^L \int_{\varepsilon}^{1} P(w|S) P(S)\,dw$$

or, considering that

$$\int_{\varepsilon}^{1} P(\omega|S) P(S)\,dw \leqslant 1$$

we obtain

$$P \leqslant k(1-\varepsilon)^L.$$

But

$$k(1-\varepsilon)^L \leqslant \delta,$$

if L satisfies inequality (24). Therefore $P_\varepsilon \leq \delta$.

The quantity $\Pr(p < \varepsilon)$ considered in the theorem is the probability of the event "$p < \varepsilon$" on condition that termination occurred at some step (i.e., the conditional probability). The probability P is the joint probability of the events "$p > \varepsilon$" and "termination has occurred." However, since the latter event always occurs (in no more than Lk steps), then

$$\Pr(p<\varepsilon)=1-P_\varepsilon \geqslant 1-\delta.$$

Q.E.D.

Thus, if from any considerations whatever the estimate can be obtained of k, the maximum possible number of error corrections, theorem 3 permits choosing L such as to guarantee in the use of the first variant of the termination conditions, the required quality of the teaching process. However, usually k is not known in advance; estimate (20) cannot be used in practice, since the sets A and B are also not known in advance. In this consists the inadequacy of the first variant of the termination condition. The second variant described below avoids this difficulty.

Variant 2. In the first variant the reliable number of examples L without error correction after which the teaching process terminates, is independent of the number of errors corrected previously. In the second variant it is assumed that the number $L = L_s = L_0 + s$, where L_0 is a prescribed number, s is the number of previously corrected errors. Thus, in the second variant L increases by unity after each corrected error.

For this definition, by virtue of theorem 1, termination must occur after a finite number of examples not exceeding

$$\sum_{s=1}^{k} L_s = kL_0 + \frac{k(k+1)}{2}.$$

The problem now consists in selecting L_0 so as to guarantee the required quality of the teaching process.

Theorem 4. Let p be the error probability in the test process after termination of teaching.

Then, for any $\varepsilon > 0$ and $\delta > 0$, the probability $\Pr(p < \varepsilon)$ that $p < \varepsilon$ is greater than $1 - \delta$, if

$$L_0 > \frac{\ln \varepsilon\,\delta}{\ln(1-\varepsilon)}. \tag{25}$$

Let us emphasize that the choice of L_0 according to (25) depends only on the values of ε and δ characterizing the quality of the teaching process, and is independent of the forms of A and B and the sample statistics.

Proof of theorem 4. Repeating exactly the first part of the proof of theorem 3, we obtain

$$P_\varepsilon = \sum_{s=1}^{k} \int_{\varepsilon}^{1} P(w \mid s)^{L_s} P(s)\, dw,$$

where $L_s = L_0 + s$.

Let us obtain an upper bound for P_ε:

$$P_\varepsilon \leqslant \sum_{s=1}^{\infty} (1-\varepsilon)^{L_0+s} \int_{\varepsilon}^{1} P(w \mid S)\, P(S)\, dw \leqslant (1-\varepsilon)^{L_0} \sum_{s=0}^{\infty} (1-\varepsilon)^s = \frac{1}{\varepsilon}(1-\varepsilon)^{L_0} \leqslant \delta.$$

Therefore

$$P_\varepsilon \leqslant \delta,$$

if L_0 satisfies inequality (25). Taking this last into account in the proof of theorem 3, we may write $\Pr(p < \varepsilon) = 1 - P_\varepsilon \geq 1 - \delta$, Q.E.D.

In variant 2 of the termination conditions we may adopt in place of $L_s = L_0 + s$ any other monotonically increasing function $L_s = L_0 + \alpha(s)$ as long as the series $\sum_{s=1}^{\infty}(1-\varepsilon)^{L_s}$ converges. The proof remains exactly the same, but the estimate of L_0 depends on the choice of the function α_s.

5. Realization of the Method of Potential Functions. The Perceptron.

Two avenues of realization of the algorithm described in section 2 are open using the means available in computing techniques, distinguished by the methods of storing the data arriving in the machine during the learning process and processed by virtue of the algorithm. The first avenue is specifically for the use of universal computers, the second (which, of course, can also be realized on universal machines) leads to the construction of specialized analog type devices (schemes). Below it will be shown that a particular case of this type of device is the perceptron.

First method. Let us consider the rth step of the algorithm. Up to this step the machine has stored in its memory the coordinates of all those points $x^1, x^2, \ldots, x^l$ shown during the teaching process for which error correction was required before this step, and the numbers $\alpha_1, \alpha_2, \ldots, \alpha_l$ ($\alpha_i = \pm 1$) which indicate by their sign to which of the sets (A or B) these points belong. With the appearance at the $(r+1)$st step of a new point x^* the machine calculates the quantities $K(x^*, x^i)$ ($i = 1, 2, \ldots, l$) and the sum

$$K_r(x^*) = \sum_{i=1}^{l} \alpha_i K(x^*, x^i).$$

If $K_r(x^*) > 0$ and $x^* \in A$ (or $K_r(x^*) < 0$ and $x^* \in B$) the computation results at this step and the point x^* are erased and the following example is considered. If $K_r(x^*) > 0$, and $x^* \in B$ (or $K_r(x^*) < 0$ and $x^* \in A$), the additional point $x^{l+1} = x^*$ is stored in the memory with the number α_{l+1} whose sign indicates to which set x^* belongs, and the other numbers calculated at this step are erased.

Thus at the end of each step (and thus at the end of the entire teaching process) only two series of numbers $x^1, \ldots, x_l$ and $\alpha_1, \ldots, \alpha_l$ are stored in the computer memory. Concerning the values of the potential function $K(x, y)$ and the functions $K_r(x)$ constructed during the process, they are not required in the machine memory, and in testing they are calculated at each step, as convenient, and then erased.

We now make several remarks on the choice of the potential function $K(x, y)$ for practical use of the algorithm. We shall consider that a distance $R(x, y)$ is defined in the space X between two points $\underline{x}$ and $\underline{y}$.

In a number of cases it is found convenient not to be concerned with the choice of the system of functions $\varphi_i(x)$ and constants λ_i, calculating the potential from formula (1), but to give directly the form of the potential $K(x, y)$, according to the following intuitive considerations. Firstly, the function $K(x, y^*)$, considered as a function of $\underline{x}$ for fixed $y = y^*$, should be sufficiently "smooth" and not too "oscillatory," since the sum of potential functions constructed by the algorithm should approximate the separation function $\Psi(x)$ which, according to the basic assumption, is not "excessively jagged." Secondly, it is desirable that the function $K(x, y^*)$ take on a maximum value for $x = y^*$ and decrease with distance of the point $\underline{x}$ from y^* since by virtue of the "smoothness" of the separation function $\Psi(x)$, the closer a point $\underline{x}$ is to y^* the greater is the "basis" for saying that $\underline{x}$ belongs to the same set as y^*. If we take into consideration further the remarks in the discussion of the basic assumption (section 1), the possibility of varying the form of the potential function in wide limits, and the requirement of symmetry $K(x, y) = K(y, x)$ following from (1), in practical use of the method of potential functions it is possible to prescribe $K(x, y)$ in the form of a fairly simple "on the average" decreasing function of the distance $R(x, y)$, for example $e^{-\alpha R^2}$, $\sin(\alpha R/R)$, etc. In the choice of parameter α it is necessary to keep in mind that the more complicated the problem (i.e., the more the sets A and B "interpenetrate" and, therefore, the more "oscillatory" is the separation function), the more rapidly should the potential $K(x, y^*)$ decrease. In practical use of the algorithm the steepness parameter of the potential function is chosen experimentally (cf [8]).

Second method. Let the potential $K(x, y)$ be chosen such that the coefficients $\lambda_i = 0$ for $i > N$, i.e., $K(x, y) = \sum_{i=1}^{N} \lambda_i^2 \varphi_i(x) \varphi_i(y)$. At the $\underline{r}$th step of the algorithm the machine memory has stored the numbers $\gamma_1^r, \gamma_2^r, \ldots, \gamma_N^r$

having the significance of components of the normal vector of the hypersurface in the N-dimensional linearization space. For the (r+1)st example, the point x^*, the quantity

$$z_i^* = \lambda_i \varphi_i(x^*)$$

and the sum

$$K_r(x) = \sum_{i=1}^{N} \gamma_i^r \lambda_i \varphi_i(x^*)$$

are calculated.

Further, we compute the number

$$\delta^r = \begin{cases} 0, & \text{if} \quad K_r(x^*) > 0 \text{ and } x^* \in A, \quad \text{or} \quad K_r(x^*) < 0 \text{ and } x^* \in B, \\ 1, & \text{if} \quad K_r(x^*) < 0 \text{ and } x^* \in A, \\ -1, & \text{if} \quad K_r(x^*) > 0 \text{ and } x^* \in B \end{cases}$$

and new values $\gamma_1^{r+1}, \ldots, \gamma_N^{r+1}$ from the formula

$$\gamma_i^{r+1} = \gamma_i^r + \delta^r \lambda_i \varphi_i(x^*).$$

Then the old values of γ_i^r and all the computations are erased and the machine stores in its memory only γ_i^{r+1} (i = 1, ..., N).

We now turn our attention to the fact that in the second method of realizing the algorithm it is not necessary to store the points shown in the teaching process, but in place of this it is necessary to store N values of γ_i^r. Therefore, the advantage of one or the other realization depends on the relations between the dimensions of the space X and the linearization space and on the length of the teaching sequence.

Let us now consider the diagram of Fig. 5. The diagram contains N nonlinear converters $z_i = \psi_i(x) = \lambda_i \varphi_i(x)$ (i = 1,..., N), a set of multipliers ×, realizing instantaneous multiplication of the signals applied, a summation unit $\sum_i$, which instantaneously produces at its output the sum of the signals applied to the inputs, accumulators having in all a single input and producing at the output the sum of signals applied to this input from the start of operation of the system,* a nonlinear output element sign c with characteristic y = sign v and finally, a nonlinear element $\delta(h, y)$, forming the function δ from the signals y and h, where h is the signal indicating the sign of the set to which the example point belongs (h = 1 if $x \in A$ and h = −1 if $x \in B$):

$$\delta = \begin{cases} 0, & \text{if} \quad y = h, \\ h, & \text{if} \quad y \neq h. \end{cases}$$

It is easily seen that the diagram of Fig. 5 realizes exactly the second of the above described methods of realizing the algorithm, i.e., realizes the teaching process by the method of potentials.

Let us now consider a particular case, where the space X is the space of vertices of the m-dimensional cube, and the system of functions

$$\psi_i(x) = \lambda_i \varphi_i(x) \qquad (i = 1, \ldots, N)$$

is a function of the form

$$\psi_i(x) = \mathrm{Sg}\left(\sum_{s=1}^{m} \mu_i^s x_s + \mu_i^0\right), \tag{2.6}$$

where $(x_1, \ldots, x_m)$ is the set of coordinates of the vertices of the m-cube, μ_i^0 is a prescribed constant, and the constants

*In Fig. 5 the circuit generating the coefficients γ_i is shown only for the first converter $z_1 = \psi_1(x)$ (for γ_1). Similar circuits containing accumulators Σ_t are connected after each nonlinear converter.

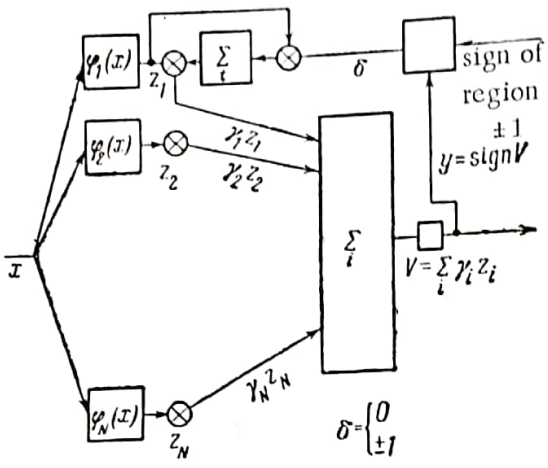

Fig. 5.

μ_i^s (s = 1, ..., m) have the values 0, 1 or −1. Then, as is easily grasped, Fig. 5 is exactly the perceptron scheme of Rosenblatt's Mark 1 (cf. [9]). The nonlinear converters

$$z_i = \psi_i(x)$$

play here the role of the associative elements of the perceptron (A-elements).

Therefore

1. The Mark 1 perceptron is a particular case of the above described class of schemes (Fig. 5).

2. The operation of the perceptron can be understood to be a realization of the method of potential functions and the characteristics of the perceptron A-elements are "harmonics" of the function system in which the potential is expanded.

3. The above theorems on the convergence of the teaching process for pattern recognition (theorems 1 and 2) are applicable to the Mark 1 perceptron and solve the problem of convergence of the process in it.

4. From the viewpoint of convergence of the process it is not at all necessary that the A-elements be threshold elements; any functional converters $\psi_i(x)$ are suitable if only the separation function $\Psi(x)$ is expandable in the system $\psi_1(x), \ldots, \psi_N(x)$.*

Let us now consider the form of the potential which is in fact realized in the Mark 1 perceptron, i.e., in the case where the functional converters are threshold elements.

Since the μ_i^s are chosen at random in the perceptron, it is possible to consider both concrete realizations of the perceptron (if these numbers have already been chosen) and the statistical properties of the ensemble of perceptrons. Let us begin with a concrete realization of the perceptron. We shall consider the potential function

$$K(x, y) = \sum_{i=1}^{N} \psi_i(x)\psi_i(y). \tag{27}$$

In the euclician m-dimensional space E_m we define the vertices of the m-dimensional cube, forming the space X. Each A-element defines in E_m a plane

$$\sum_{s=1}^{m} \mu_i^s x^s + \mu_i^0 = 0, \tag{28}$$

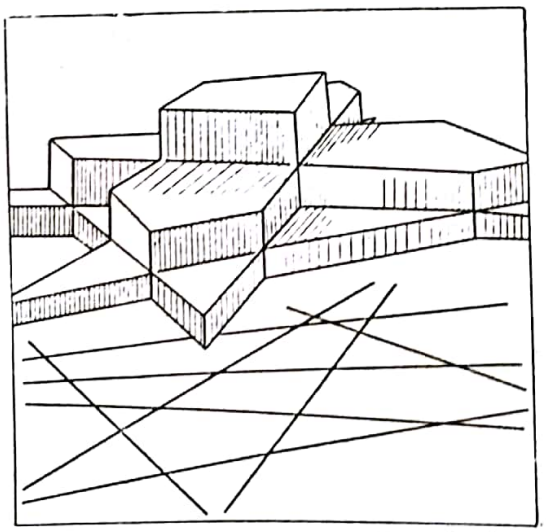

Fig. 6.

while the set of these planes divides E_m into polyhedra. The vertices of the cube X are divided into sets corresponding to the polyhedron in which they occur. Let us consider (27) for fixed $y = y^*$. The value of the function $K(x, y^*)$ does not change if x is identified with the vertices of the cube located in the same polyhedron. In this sense $K(x, y^*)$ is a piecewise-constant function, given on the polyhedra. Let us consider further the value of $K(y^*, y^*)$. This number is obviously equal to the number of excited A-elements for $x = y^*$. Let us plot in E_m an arbitrary straight line passing through the point $x = y^*$ and shift x along this line from the point y^*. Up to the first intersection of this line with a boundary of the polyhedron containing the vertex y^* the value of $K(x, y^*)$ does not change and is equal to $K(y^*, y^*)$.

*Independently of us V. A. Iakubovich has directed attention to this in an article whose manuscript came to our attention during preparation of the present article for publication.

At each intersection with the constructed planes (i.e., the boundaries of the polyhedra) the value of $K(x, y^*)$ can only decrease. In effect, at the intersection of the line with the ith plane in (28) two cases are possible: $\psi_i(y^*) = 0$ (the ith A-element is not excited at the point y^*) or $\psi_i(y^*) = 1$ (the ith A-element is excited at the point y^*). In intersection of the line with the ith plane the components $\psi_k(x)\psi_k(y^*)$ in (27) for $k \neq i$ do not change and only the component $\psi_i(x)\psi_i(y^*)$ can change. But in the case $\psi_i(y^*) = 0$ this term is equal to zero and the value of the potential does not change. If $\psi_i(y^*) = 1$, then $\psi_i(x)$ changes from 1 to 0 at the intersection with this plane, so that the term $\psi_i(x)\psi_i(y)$ which was equal to 1 becomes equal to 0, and the potential decreases by unity.

From the above it follows that for any concrete realization of the perceptron the potential $K(x, y^*)$ represents a function which does not increase in any direction from the "source" of potential and reaches a maximum at $x=y^*$. The form of the potential for $m = 2$ is shown in Fig. 6.

Concerning the statistical ensemble of perceptrons, for it the functions $\psi_i(x)$ are random functions, and so the potential is also a random function. It is easily shown that for any pair of fixed points x and y the value of the potential $\overline{K(x, y)}$ averaged over the ensemble of perceptrons can be expressed by the formula

$$\overline{K(x, y)} = N(p(x) - p(x, y)), \tag{29}$$

where $p(x)$ is the probability that a randomly chosen A-element be excited at the point x, and $p(x, y)$ is the probability that it be excited at the point x and not at y. Since, in general, the probability that a random plane separates two points x and y increases with increase of the distance between them, the mean potential $\overline{K(x, y)}$ is a function decreasing with increase of distance from the source.

LITERATURE CITED

1. F. Rosenblatt, Perceptron Simulation Experiments, Proc. IRE, 48, 3 (1960).
2. M. M. Bongard, Modelling the recognition process on a digital computer, Biofizika, 6, 2 (1961).
3. G. S. Sebestyen, Decision-Making Processes in Pattern Recognition, The Macmillan Company, N. Y. (1962).
4. L. A. Kamensky and C. N. Liu, Computer-automated design of multifont print recognition logic, IBM J. Research and Development 7, 1, January, 1963.
5. A. Borsellino and A. Gamba, An outline of mathematical theory of Papa Nuova cimento, 20, Ser. X, 2 (1961).
6. E. M. Braverman. Experiments in teaching a machine to recognize visual patterns. Avtomatika i telemekhanika, 23, No. 3 (1962).
7. A. B. J. Novikoff, On convergence proofs for perceptrons, Report at the Symposium on Mathematical Theory of Automata, Politechn. Inst. Brooklyn April, 24-26, 1962.
8. O. A. Bashkirov, E. M. Braverman, and I. B. Muchnik, Algorithms for teaching a machine to recognize visual patterns, bassed on the use of potential functions, Avtomatika i telemekhanika, 25, 5 (1964).
9. J. S. Hay, F. S. Martin, and S. V. Whitman, The Mark-1 perceptron, its design and characteristics, Cybernetic Coll., 4, Izd. inostr. lit (1962).

All abbreviations of periodicals in the above bibliography are letter-by-letter transliterations of the abbreviations as given in the original Russian journal. *Some or all of this periodical literature may well be available in English translation.* A complete list of the cover-to-cover English translations appears at the back of this issue.
