Teaching Space: A Representation Concept for Adaptive Pattern Classification

by

Stevo Bozinovski

Department of Computer and Information Science
University of Massachusetts
Amherst, Massachusetts 01003

COINS Technical Report No. 81-28

Acknowledgement. This work was prepared during my visit to the Computer and Information Science Department, University of Massachusetts at Amherst from September 1980 to December 1981. I wish to thank Prof. Spinelli for enabling me to work with the Adaptive Network Group, which has provided an excellent environment for studying the adaptive properties of biological and artificial systems. Prof. Spinelli and Prof. Barto have given valuable comments on the earlier drafts of this paper.

My stay at this Department was supported by a Fulbright grant, and this particular project by a Sigma Xi grant.

ABSTRACT

This work considers the problem of the pattern recognition training from the teacher point of view. The approach, denoted as teaching space approach is proposed, which makes explicite two problems of the pattern recognition training: the problem of interpattern similarity and the problem of a transfer of some previous training of the learner. Particular attention is paid on the problem of interpattern similarity. As a result, a measure of similarity is derived from the behavior of an artificial classifier, which turns out not to be considered so far in any area dealing with similarity judgment. Some implications to the theory of adaptive linear classifiers, cluster analysis, and neural modelling are discussed. The goal-seeking nature of the teacher during the pattern recognition training is emphasized.

INTRODUCTION

It is a well known fact that perceptron type machines cannot form a concept of similarity different from some feature matching procedure [6] This is consistent with the studies of their inability to extract some complex features such as connectivity [28]. The recognition of translation, rotation and change in shape is also not an ability of these machines although some attempts have been made in that direction [36]. The notion of similarity, although originally present in peceptron theory [35], gradually has dissappeared from theories of adaptive pattern recognition. For example, contemporary theory of linear machines [31] does not include that notion at leas not explicitely.

From the other side,other areas dealing with pattern recognition, such as cluster analysis techniques, neccesarily includes the notion of similarity [38],[41]. Similarity is also a natural concept in human pattern recognition and is actively studied in the latest psychological research [21],[44],[26],[33]. The importance of similarity in connection with learning was also pointed out in artificial intelligence [14] . Some attempts for adoption of this notion in neural modelling have recently been shown [19],[46].

This work studies in some detail the "trivial" concept of similarity built by a linear machine during the training process. As a result, this work has derived a measure of similarity from the pattern classification process performed by the linear classifiers. This measure is used by the machine

during the training process, that is, when the machine learns to distinguish between the reference patterns. It turns out that the derived measure of similarity has not been proposed so far among the measures of similarity in cluster analysis and psychology. Let us emphasize here that it is not an intention to compare the measure of similarity derived here with other measures proposed so far. But we do emphasize that the measure described bellow is derived from the process of the pattern classification performed by an artificial system, rather then proposed heuristically on the basis on the observations over some data structure.

The approach taken here is influenced by the original works of Glushkov [18], Block [5], and Rosenblatt [37]. The problem of learning is considered as a control problem, and is viewed from the teacher side. Attention is placed on the generation of the training sequence rather then on the trajectory of the learner's memory vectors in the weight space, which is the approach taken in the classical theory of linear machines. This approach develops the model of the training process in which the interpattern similarity and transfer of training can be viewed explicitely as the parameters of the model. This is consistent with the pedagogical and psychological findings about pattern recognition training [15],[45],[17],[40].

The approach introduces two concepts, teaching space and similarity space which are described in the following sections. Some implications of the results obtained using the teaching space approach are discussed in the later chapters of this paper.

THE TEACHING SPACE CONCEPT

A classifier is given by the set of n real valued functions $g_i(\underline{x})$, i=1,..,n often called discriminant functions, defined over the set of vectors named patterns the typical element being $\underline{x}$. The pattern $\underline{x}$ is said to belong to the i-th class if

$$g_i(\underline{x}) > g_j(\underline{x}) \text{ for all } j \neq i. \qquad (1)$$

The classifier is trainable if it is possible to modify the parameters of the functions $g_i(\underline{x})$ using the set of reference patterns from the given reference set X . The modification is usually performed by introducing an external control system, or teacher. Figure 1 shows such a classifier. The control vector $U^{\sim}=[u_1,\ldots,u_n]$("~" denotes transposition) modifies the parameters of the classifier in order to achieve some predefined desired classification, i.e. canonical surjection, over the set X . The maximum selector element generates the signal $Y^{\sim}=[y_1,..,y_n]$ where $y_i=1$ if and only if condition (1) is satisfied, otherwise $y_i=0$, i=1,..,n.

Figure 1

In the simplest case of a linear classifier the discriminant functions are chosen to be linear forms

$$g_i(\underline{x})=\langle \underline{w}_i,\underline{x}\rangle + \theta_i \qquad (2)$$

(<.,.> denotes inner product) where $\underline{w}_i$ is the memory vector associated with the i-th class, and θ_i is some threshold value. The elements which compute the functions $g_i(\underline{x})$ are denoted as dot product units [32],[23]. It is sometimes convenient to introduce the vectors $\underline{w}_i'=[\theta_i,\underline{w}_i]$ and $\underline{x}'=[1,\underline{x}]$ in order to write the functions (2) in the form $gi(\underline{x})=\langle \underline{wi}',\underline{x}'\rangle$. We will not use

that notation, but let us note that the usage of the vectors $\underline{x}'$ instead of $\underline{x}$ in the training procedure affects the results obtained. A discussion about that is given later in this paper.

The approach taken here will be introduced by assuming the simplest possible task for this classifier: the single-sample dichotomy, in which the classifier is trained to distinguish, i.e. separate and correctly classify two reference patterns $\underline{x}_1$ and $\underline{x}_2$. Let us define the desired classification as $\underline{x}_i$ belonging to i-th class, for i=1,2. Then in order for the classification performed by this classifier to be correct it is necessary and sufficient to find $\underline{w}_1$ and $\underline{w}_2$ such that:

$$g_1(\underline{x}_1) > g_2(\underline{x}_1) \qquad (3.1)$$

$$g_2(\underline{x}_2) > g_1(\underline{x}_2) \qquad (3.2)$$

As a procedure of searching for the feasible $\underline{w}_1$ and $\underline{w}_2$ using a training paradigm, we choose the fixed increment rule with arbitrary nonnegative constants of "reward" c_r and "punishment" c_p, not both equal to zero. Two types of trials are distinguished, training trials and examination trials. The examination trial does not change the memory of the classifier; it just evaluate the classification performed by the classifier. If $\underline{x}_i$, i=1,2, was correctly classified in the previous examination trial, there will be no training trial. If a misclassification for $\underline{x}_i$ occurs, in the training trial:

$$\underline{w}_i(t+1) = \underline{w}_i(t) + c_r \cdot \underline{x}_i \qquad (4.1)$$

$$\underline{w}_k(t+1) = \underline{w}_k(t) - c_p \cdot \underline{x}_i \quad \text{for } i \neq k. \qquad (4.2)$$

In other words, the elements of the classifier which have voted for the incorrect classification will be punished and the elements which have voted for the desired classification will be rewarded. This training procedure is the (generalized) $\alpha(c_r, c_p)$ learning law used in the perceptron training

experiments [18]. The laws $\alpha(c,c)$, $\alpha(c,0)$, and $\alpha(1,0)$ are most commonly used.

Let T be a training sequence, consisting of the patterns $\underline{x}_1$ and $\underline{x}_2$ which appear in the training trials. Let pi be the number of appearances of the pattern $\underline{x}i$ within T. The sequence T is said to be successful if after training with that sequence, the classifier correctly classifies the patterns $\underline{x}_1$ and $\underline{x}_2$. A successful sequence is denoted by T*. The length of the training sequence $L(T^*)=p_1+p_2$ can be used as a measure of the optimality of T. Observe that given a successful training sequence T* , any permutation of that sequence is also a successful training sequence for that system [18].

Let $\underline{w}_{1o}$ and $\underline{w}_{2c}$ be the initial values of the vectors $\underline{w}_1$ and $\underline{w}_2$ at the begining of the training process. Then after training with some T, the values of the vectors $\underline{w}_1$ and $\underline{w}_2$ will be

$$\underline{w}_1 = p_1c_r\underline{x}_1 - p_2c_p\underline{x}_2 + \underline{w}_{1o} \qquad (5.1)$$

$$\underline{w}_2 = p_2c_r\underline{x}_2 - p_1c_p\underline{x}_1 + \underline{w}_{2o} \qquad (5.2)$$

Now, replacing (1) and (5) in (3) and introducing the notation

$$a_{ik} = \langle \underline{x}_i, \underline{x}_k \rangle \qquad (6)$$

$$b_{ik} = \langle (\underline{w}_{io}-\underline{w}_{ko}), \underline{x}_k \rangle \qquad (7)$$

$$d_{ik} = \theta_i-\theta_k \qquad (8)$$

we obtain the pair of inequalities

$$a_{11}p_1 > a_{21}p_2 + (b_{21}+d_{21})/(c_r+c_p) \qquad (9.1)$$

$$a_{22}p_2 > a_{12}p_1 + (b_{12}+d_{12})/(c_r+c_p) \qquad (9.2)$$

Thus, the problem of training for distinguishing, i.e. separating, the patterns $\underline{x}_1$ and $\underline{x}_2$ is equivalent to the problem of the solution of the inequalities (9). These inequalities are the basis of the approach taken here. They describe the model of the teaching process over the $\alpha(c_r,c_p)$

system whose threshold does not change during the training.

Let us give an interpretation of the parameters that appear in the inequalities (9). Note that the a-parameters a_{ik}, i,k=1,2, which appear in the relations (9) are functions of the input patterns only. They represent the degree of matching between the features of the patterns $\underline{x}_1$ and $\underline{x}_2$. Thus we will refer to the a-parameters as to the matching coefficients. The influence of the state of memory of the learner at the begining of the training experiments is considered through b-parameters, b_{ik}. In other words, they represent the influence of the transfer of some previous training upon the training process. The d-parameters are results of the set of threshold values for the linear classifier at the begining of the training process. Here it is assumed that the d-parameters are not changing during the training process. The b- and d-parameters are in general assumed to be unknown, in contrast to the a-parameters which are always computable from the reference patterns. In the case where $b_{ik}=0$ and $d_{ik}=0$ we say that the learner starts with the homogeneous initial conditions. It is easy to see that in this case the problem of solution of (9) is a linear programming problem. Otherwise it is search under uncertainity since b- and d-parameters are unknown. The p-parameters p_1 and p_2, which are searched for, represent the distribution of the patterns within the training sequence. Another interesting interpretation can be obtained by dividing the inequalities (9) by the length of the teaching sequence L(P). Then the parameters $p_i/L(P)$ i=1,2 are the probabilities of the appearence of the patterns in the training sequence.

The geometrical interpretation of this approach is given in Figure 2. Two cases are shown: $a_{ik} > 0$,(Fig. 2.1) and $a_{ik} < 0$ (Fig. 2.2)

Figure 2

In Figure 2 the problem is represented in (p_1, p_2)-space, which we call teaching space or P-space. This space consists of all the points (p_1, p_2) where p_1 and p_2 are nonnegative integers. It must be searched in order to find a point satisfying (9). These points are marked in Figure 2. Such points can be found in the solution region which is shaded on the figure. This region is always convex polyhidron. In the case of homogeneous initial conditions it is convex polyhidral cone.

It is important that the basic relations (9) of this approach do not explicitly include the memory matrix $\underline{W}=[\underline{w}_1, \underline{w}_2]$, but rather the teaching sequence through the vector $P=[p_1, p_2]$. One must search for the nonnegative integers p_1 and p_2, parameters which are directly controllable and observable in every training process: they are parameters of the control system (teacher). In general, the parameters of the learner, such as the state of its memory, are observable and controllable only indirectly, if at all.

Some results concerning the influence of the learning constants c_r and c_p upon the training process are immediately evident from the inequalities (9). Since learning constants are assumed positive and not both equal to zero in the case of homogeneous initial conditions they have no effect upon the training process. If the initial conditions are not homogeneous, the effect of the learning constants can be summarized in the statement that the bigger the sum c_r+c_p, the less the influence of the initial conditions upon the training process. Certainly, this is not always desirable. The system which could show a positive transfer of some previous training by chosing large cr and cp can become equivalent to a system which has no knowledge about the "lessons". In that case the convergence of the training of such a system will

be slowed using the large learning constants.

Further in the sequel we will be mainly interested in the influence of the chosen reference patterns upon the teaching process.

SIMILARITY SPACE CONCEPT AND THE MEASURE OF INTERPATTERN SIMILARITY DURING THE TRAINING PROCESS

Consider the construction of the solution region in the teaching space (Figure 2). Given inequalities (9), we first assume homogeneous initial conditions. Then a convex polyhidral cone is defined in the teaching space whose vertex is at the coordinate origin. This cone is characterised by its vertex angle. Now, taking into account the given initial conditions, the cone is translated relative to the coordinate origin. It is important that no rotation or scaling is performed due to the influence of the initial conditions. As a result of that construction, the intersection between the cone defined by the reference patterns, and the positive quadrant of the coordinate system is the region where the solution points can be found.

Thus, given arbitrary reference patterns $\underline{x}_1$ and $\underline{x}_2$, in the teaching space (p_1, p_2) is defined an angle θ (Fig. 2)

$$\beta = 90° - \alpha_1 - \alpha_2 \qquad (10)$$

where

$$\alpha_1 = \operatorname{arctg}(a_{21}/a_{11}) \qquad (11.1)$$

$$\alpha_2 = \operatorname{arctg}(a_{12}/a_{22}) \qquad (11.2)$$

Now considering the Figure 2, it is intuitively clear that the greater the angle, the easier it should be to search for a goal point in the teaching space. Thus, the angle β can be viewed as a measure of distinguishability or the distance between the patterns $\underline{x}_1$ and $\underline{x}_2$ assumed in the internal model of the linear classifier. This motivates definition of a measure of similarity between the patterns to be the cosine of the angle β.

Note that this measure of similarity is naturaly built into a linear classifier, and used during the learning for pattern recognition. This measure is derived from pattern recognition training of an artificial classifier rather than from the analysis of the patterns in the feature space. As in the experiments performed in psychology where some physiological parameter is used , for example reaction time[34], as a measure of similarity, here it is done with an artificial pattern recognizer: the similarity measure is derived from the analysis of the pattern recognition process in the teaching space.

The function $\cos\beta$ where β is defined in (10) can be expressed in several ways. The most convenient for our further discussion is the form used in the following observation. This observation sumarizes the discussion above.

Observation 1. Given two reference patterns $\underline{x}_1$ and $\underline{x}_2$ which a linear classifier is trained to distinguish and correctly classify using $\alpha(c_r, c_p)$ fixed increment rule. Then the function

$$s_2(\underline{x}_1, \underline{x}_2) = \cos \text{ arc } \left(\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} \right) \qquad (12)$$

is a measure of similarity between the patterns $\underline{x}_1$ and $\underline{x}_2$ assumed by that system during the training period.

Proof. Replacing (10) in (12)

$$s_2(\underline{x}_1,\underline{x}_2) = \cos\beta = \cos(90°-\alpha_1-\alpha_2) = \sin(\alpha_1+\alpha_2) \ .$$

By definition (11) , $tg\alpha_1=a_{21}/a_{11}$ and $tg\alpha_2=a_{12}/a_{22}$, which substituting above gives

$$\cos\beta = \frac{a_{11}a_{12} + a_{22}a_{21}}{\sqrt{a_{11}^2+a_{21}^2}\cdot\sqrt{a_{22}^2+a_{12}^2}}$$

where we recognoze the expression stated in the Observation. Q.E.D.

Observation 1 shows that similarity analysis is performed over the vectors of the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad (13)$$

which we call <u>feature matching matrix</u>. Note that cosine of the angle between either the row- or the column vectors of A will give the similarity measure (12).

The matrix A is a Gramian matrix which means that the properties of an inner product space will be preserved in the similarity analysis. The properties which are refered to in the sequel are:

$$a_{11} \geq 0 \quad \text{and} \quad a_{22} \geq 0 \qquad (14.1)$$

$$a_{12} = a_{21} \qquad (14.2)$$

$$a_{11}a_{22} \geq a_{12}^2 \qquad (14.3)$$

$$a_{11}+a_{22} \geq 2a_{12} \qquad (14.4)$$

$$(a_{11}+1)(a_{22}+1) \geq (a_{12}+1)^2 \qquad (14.5)$$

Strict inequality in (14.1) holds if and only if $\underline{x}_1=\underline{x}_2=0$. Relation (14.2) is a tautology in an inner product space. The relation (14.3) is the Cauchy-Schwartz inequality for which the sign of strict inequality is valid assuming that the patterns are not colinear. The relation (14.4) is the

consequence of the positiveness of the euclidean distance and strict inequality holds assuming the patterns are not identical. Relation (14.5) is the consequence of the all others, and the strict inequality holds if and only if the patterns are not identical.

A geometrical interpretation of the relation (12) is given on Figure 3. Figure 3 represents the space where the pair of vectors (a_{11},a_{21}) and (a_{12},a_{22}) is considered. This space we denote as similarity space or S-space.

Figure 3

Similarity space shows actualy the vertex angle of the cone in the twodimensional teaching space . However, those are different spaces. Given $\underline{x}_1$ and $\underline{x}_2$, the similarity space is a continuous and bounded , whereas the teaching space is a unbounded discrete space in the positive quadrant.

Now the analysis of the training problem turns out to be the analysis of the interpattern similarity. The following is the result of that analysis.

Lemma 1. The measure of similarity (12) takes its extremal values iff the patterns $\underline{x}_1$ and $\underline{x}_2$ are colinear, i.e. if there exists some scalar k for which $\underline{x}_1=k\ \underline{x}_2$. The minimal value -1 is obtained iff $k<0$, and the maximal value 1 iff $k>0$.

Proof. The extremal values of a cosine function are -1 and 1. Suppose $\underline{x}_1=k.\underline{x}_2$ and $k>0$. Then it is directly evident that that gives $\cos\beta=1$. Analogously, for $k<0$, implies $\cos\beta=-1$. Now let us prove the converse. Observe that by definition $a_{12} = |\underline{x}_1||\underline{x}_2|\cos\psi$ where $|.|$ denotes norm and ψ is

the angle between the patterns. Replacing in (12) we obtain

$$\cos\beta = \frac{(a_{11}+a_{22})\cos\psi}{\sqrt{a_{11}+ a_{22}\cos^2\psi}\ \sqrt{a_{22}+ a_{11}\cos^2\psi}}$$

which means that $\cos\beta$ and $\cos\psi$ have always the same sign. Having that, let $\cos\beta = 1$. Replacing in relation above we obtain $\cos^2\psi = 1$ which gives two solutions, $\cos\psi = 1$ and $\cos\psi = -1$. Since the sign must be equal to the assumed sign of $\cos\beta$, the solution is $\psi = 0$, i.e. patterns are positively colinear. Analogously, for assumption $\cos\beta = -1$, we obtain that the patterns are anticolinear. Q.E.D.

Corolary 1. Two nonzero binary patterns will be assumed by a linear classifier during the training process to be maximaly similar iff they are identical, and minimaly similar iff they are orthogonal.

Proof. The proof is straight forward observing that two binary patterns have only nonnegative components, which means their inner product is always nonnegative. Minumum value of the similarity measure is 0, which replacing in (12) yields that the patterns are orthogonal. Further, since two binary patterns are colinear iff k=1, i.e. if they are identical. Q.E.D.

n-DIMENSIONAL TEACHING SPACE

Let us consider the single-prototype multiple category problem, or pair-association problem in which n samples are learned to be distinguished and classified into n different classes. In that case the system of $n.(n-1)/2$ pairs of inequalities

$$a_{ii}p_i > a_{ki} + r_{ki} \qquad (15.1)$$

$$a_{kk}p_k > a_{ik} + r_{ik} \qquad (15.2)$$

is to be considered, where $r_{ik}=(b_{ik}+d_{ik})/(c_r+c_p)$, for all $i,k = 1,..,n$.

Using the terminology developed above, it can be shown [11] that a solution invariant to the initial conditions exists whenever there are no two patterns $\underline{x}_i$ and $\underline{x}_k$ i,k=1,..,n within the samples which are maximaly similar i.e. positively colinear.

It is worthy to mention that the generalization of the notion of colinearity, linear dependence, is not a restriction. We will show an example, which will illustrate that fact and also illustate the application of the teaching space approach.

Example 1. Three reference patterns are given: $\underline{x}_1=[0.1,0.0]$, $\underline{x}_2=[0.0,0.2]$, and $\underline{w}_3=[0.1,0.2]$. Clearly, they are linearly dependent, $\underline{x}_3=\underline{x}_1+\underline{x}_2$. Let a $\alpha(1,0)$ system starts from the homogeneous initial conditions. We will find a solution vector $P=[p_1,p_2,p_3]$ which will set the system into a state being in which it distinguishes $\underline{x}_1$, $\underline{x}_2$, and $\underline{x}_3$.

First we compute the feature matching matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 0.01 & 0.00 & 0.01 \\ 0.00 & 0.04 & 0.04 \\ 0.01 & 0.04 & 0.05 \end{bmatrix}$$

The inequalities (15) give

$p_1 > 0$ $\quad p_2 > 0$ $\quad 5p_3 > p_1$

$p_1 > p_3$ $\quad p_2 > p_3$ $\quad 5p_3 > 4p_2$

which has, for example, a nonnegative integer solution at P=[7,6,5]. (A careful reader will notice a "hidden" inequality in this problem, which often appears in the problems of linear inequalities given two unknowns pro inequality.) Thus, any teaching sequence in which $\underline{x}_1$ appears 7 times, $\underline{x}_2$ 6 times and $\underline{x}_3$ 5 times will solve the of separability between the three linearly dependent vectors. Since the solution region is convex cone, any other vector k.P, where k is natural number, is also a solution .

A graphical illustration is given in Figure 4, where the convex polyhidral cone in three dimensional space is shown. The projections to the principal coordinate surfaces are the geometrical interpretaions of the each pair of inequalities above.

figure 4

The problem of linear dependence is of particular importance in the associative memory paradigm [1],[2],[25],[30],[42]. It has been emphasized that the associative memory cannot form appropriate associative mapping if the patterns are linearly dependent [25]. Note that the machines which are known as linear classifiers [31] are actualy nonlinear devices. They contain a maximum selector, which is nonlinear element and makes the classifier to be nonlinear even in the case when the threshold values are homogenious . This is why these machines are capable of solving problem which associative memories are not able to solve.

Figure 5

In fact, a linear classifier with isothreshold values, i.e by which $d_{ik}=0$ for all i and k, can be represented as a system consisting of an adaptive associative memory and a maximum selector [11]. This is shown on Figure 5. Such a system called Linear Adaptive Array (LAA) was used in our experimental investigation which is described bellow.. Let us remark that the Steinbuch's Learning matrix [38] is actualy such a system.

THE INFLUENCE OF SIMILARITY UPON SEARCH THROUGH THE TEACHING SPACE

In the previous sections we have considered the problem of existence of the solution of the problem of training a linear classifier to perform a pair associaton, and the influence of the interpattern similarity upon that problem. Now assuming that the solution exists, we are interested in the influence of the interpattern similarity upon the search for the feasible solution in the teaching space. An interpretation of the problem of search through the training space can be given using Figure 6.

Figure 6

Given two patterns, a cone is defined in the teaching space. The system which performs the search is assumed to start in the origin of the coordinate system. Figure 6 shows four different starting points for four different systems. They are searching for the solution point using their own searching strategies. The coordinat system is not shown: it is relative to a particular system which performs the search, i.e., to the initial conditions of that system. We can imagine that each system is in the origin of its own coordinate system and that its coordinate system moves through the space.

Figure 6 shows also that from the point of view of hill-climbing, the peak defined in this kind of pattern recognition problem is characrerised by a plateau rather then by a isolated spike [39]. Moreover, using the $\alpha(1,0)$-learning rule, the only allowed movements in this climbing are "east" and "north".

The applicability of some general types of searching strategies on this problem are considered in the sequel. Two training procedures will be of interest, which we describe by the following informal code (n is the number of patterns):

```
Procedure OPENLOOP
begin i:= 0
while i ≤ n do
  i:=i+1
  call TEACH(X(i))
enddo
end
```

```
Procedure PERC
begin i:=0
while i ≤ n do
  i:=i+1
  call EXAM(X(i),grade)
  if grade='nosatisfy' call TEACH(X(i))
enddo
end
```

The first strategy is open loop, all-or-none strategy, which, once started, will perform the training trials over all the reference patterns. For our discussion is important that the probability of appearence of a pattern within the teaching sequence is uniform. The second strategy is the wellknown Perceptron strategy, which is responce senzitive, closed loop strategy: before applying the training trial TEACH in which twe weights are updated according to the learning rule of the $\alpha(1,0)$ systems , an examination trial EXAM checks the necessity of the training. This strategy produces the

teaching sequence in which probability of apppearence of a reference pattern is in general not uniform.

It turns out that another aspect of similarity becomes salient in the analysis of the speed of the convergence of the training process using the fixed increment rule: the inclusion of one pattern into an other, or in other terms, the dominance of one pattern over an other. We say that the pattern $\underline{x}_i$ dominates over the pattern $\underline{x}_k$ if and only if $a_{ik} \geq a_{kk}$. In that case we also say that $\underline{x}_k$ is included in $\underline{x}_i$. If the patterns are binary, the definition is intuitively clear: the set of features of $\underline{x}_k$ is included in the set of features of $\underline{x}_i$ if and only if $a_{ik}=a_{kk}$. Note that in binary case $\sup(a_{ik})=\min(a_{ii},a_{kk})$. If the condition of inclusion is not satisfied we say that the patterns contain mutualy distinctive features. The following is a result of the influence of the inclusion condition patterns upon the choice of a training strategy.

Theorem 1. Given two patterns $\underline{x}_1$ and $\underline{x}_2$ which are not maximaly similar in the sense of (12) and for which the condition of inclusion is not satisfied. Then the teaching procedure which iterates the subroutine OPENLOOP will reach the solution of the teaching problem for a $\alpha(c_r,c_p)$ learning system in finite number of iterations, for arbitrary initial conditions of the learner.

Proof. The strategy OPENLOOP always generate the sequences in which $p_1 \neq p_2$, which in teaching space produce a trajectory which travels along the symmetral line of the positive quadrant. The solution will exist if the symmetral line enters the solution region. But since inclusion condition is not satisfyed, i.e. since $a_{ii} \geq a_{ik}$ for i,k=1,2, and the angles α_1 and α_2 are smaller than 45°. Thus, regardless initial conditions, the symmetral line of

the positive quadrant enters the solution region. Q.E.D.

In fact, that is true for any strategy in which $p_1=p_2$. Thus, the inclusion condition can be a test of applicability of a training strategy in which the probability of the appearence of the reference patterns is uniform. In the case of $n > 2$ classes , the n-dimensional matching matrix should be considered: if the diagonal elements are the maximal elements in each row, then the strategy OPENLOOP is applicable and the training process using that strategy will converge, regardless the initial conditions.

The other, perceptron strategy is always applicable in the case when solution exists, which is welknown result from the perceptron convergence theorem [31],[28],[3],[12],[20]. In general, any strategy in which before applying a training trial, an examination of neccessity of such a trial is performed, will reach the goal point in the teaching space regardless the initial conditions providing that solution exists. That is illustrated on Figure 7.

Figure 7

On Figure 7 the inclusion condition is fulfilled and one of the boundary lines has slope 1. The solution region is so translated due to the initial conditions, that the open-loop strategy in which $p_1=p_2$ will never reach the goal (search A). An other strategy, which checks the justification of the teaching trial before its implementation, will reach a goal point (search B).

The influence of the mutual distinctivity is remarkable in the application of the perceptron strategy in the multiple category case: Let us assume homogeneous initial conditions: If all the patterns contains mutualy distinctive set of features it is sufficient to show the patterns once and the system will learn to distinguish them; i.e. one-trial learning will occur. That can be done by an open-loop strategy as well. If there are patterns $\underline{x}_i$ which satisfy the inclusion condition, $a_{ik} \geq a_{ii}$ for some k, then such patterns neccesarily appear more than once within the training sequence. But that will usualy produce the neccessity fo other patterns to appear more than once within the teaching sequence. As effect, we have prolongation of the teaching process regardless the strategy used. Thus, not only applicability of a strategy is affected by the inclusion condition, but also the speed of the convergence of the training process. In the following section we give experimental evidence of that phenomenon.

SIMULATION EXPERIMENTS

In this section we describe the results of our simulation experiments on the training the $\alpha(1,0)$ system to distinguish between a set of n binary patterns. The experimental design is as follows:

The stimuli which we have experimented with, were three sets of letters chosen from some computer perifery devices: IBM29 card puncher, VR14 video display and VT50 video display. The three sets of letters are shown on Figure 8. The motivation for chosing sets VR14 and VT50 are some recent psychological experiments concerning similarity [16],[34]. There is no

particular motivation for using set IBM29 besides the fact that this set was available. The letters were enumerated in natural, lexical order giving number 1 to the letter A and number 26 to Z. All of the letters are represented by 35-dimensional binary vectors. Those vectors are denoted by $X_1,\ldots,X_{26}$ and in such a form were presented to the experimental subject.

Figure 8

The subject was a program which simulates the behaviour of a $\alpha(1,0)$ learner, in particular nonlinear adaptive associative memory shown on Figure 5 which we call Linear Adaptive Array (LAA). Shown the binary vector, the program is supposed to recall the letter to which that vector is assigned. If the decision is ambiguous, the program produces question mark "?". In this experiment the inital condition of the system was chosen to be homogenious: all the memory elements was set to zero as well as the threshold values.

The trainer was represented by the program which represents the training strategy. Both OPENLOOP and PERC strategy were used as training strategies.

The relevant variables which we have observed during the experimental investigation were: the teaching sequence T, the length of the teaching sequence L(T), the length of the entire curriculum L(C) , where the curriculum C is defined as the sequence of all the letters appearing both in the teaching and in the examination trials.

Two series of experiments were performed. Here are the results:

Series 1: The influence of similarity

Experimenting with OPENLOOP we have chosen number of iteration K=100. The strategy was not successful for such a number. The letters which the learner failed to recognize were

from the set IBM29: C,F,J,L,P,U
from the set VR14: C,F,J,L,P
from the set VT50: C,F,J,L,O,P

For each of these letters, the learner answered with the sign "?". We have not experimented with larger values of K. It was obvious that these are the letters which are hardly distinguishable by the learner. The length of the training sequence is L(T)=2600 and the length of the curriculum is L(C)=2026, for all the sets of letters.

Experiments using the strategy PERC has shown that that strategy can be successively applied in this task. Here are the training sequences produced by this strategy for each set of letters. Each row represents an iteration of the procedure PERC described above. In the Appendix is given an example of the training protocol using this strategy.

IBM29: L(T*)=135 L(C*)=395

T* = ABCDEFGHIJKLMNOPQRSTUVWXYZ
CEFGHIJKLMNOPQRSTUVWXYZ
ABCDFGJLOPQRSUWZ
ACDEFHIJKLMNPRSTUVXY
BEFGHJKLMOPQRUWZ
ACJPRTVWXY
BDEFHIKLMNPQVZ
FOU
CEGJLS

VR14: L(T*)=151 L(C*)=411

T* = ABCDEFGHIJKLMNOPQRSTUVWXYZ
CDEFGHIJKLMNOPQRSTUVWXYZ
ABCFGJLOPQRSUWXY
ABCDEFGHIJKLMNOPQRSTUVZ
ABCDEFGHIJKLMNOPQRSTUWXZ
ABCDEFHJLMNPRVWXYZ
AHKMNPVWXY
BCEFGLOSU
J

VT50: L(T*)=207 L(C*)=545

T* = ABCDEFGHIJKLMNOPQRSTUVWXYZ
CDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHJKLMNOPQRSUVWXYZ
ABCFIJLOPQRSTUV
BCDEFGHJKLMNOPQRUVWXZ
ACDGIJOPRSTUVWY
BCDEFHJKLMNOPQSUXZ
BCDEFGIJLOQST
ABCDEFGHKLMNOPQRUVWXYZ
ACDGHIJLMNOQSUVWZ
BCEFGIJPRST
K

Some letters in the teaching sequence appear more frequently among others. Table I gives the distribution of the letters within the teaching sequences for each set of letters. It actually gives one of the solutions of the inequalities (15) for this case.

Table I

In the second phase of this experiment we have excluded from the alphabet the letters marked in the experiment with the strategy OPENLOOP. Now, modified IBM29 and VT50 sets have each 20 letters, and VT14 has 21 letters. This series of experiments has shown the predicted result: Both the strategies are effective and terminate after only one iteration, producing the teaching sequence of the same length, equal to the number of the letters

Table 1

The frequency of appearance of the letters within the teaching sequence

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
IBM29:	4	4	6	4	6	7	5	5	4	7	5	7	5	5	5	7	5	6	5	4	6	5	5	4	4	5
VR14:	6	6	7	5	6	7	6	6	4	7	5	7	6	6	6	7	5	6	6	4	6	5	6	6	5	5
VT50:	6	8	11	9	8	9	9	7	7	10	7	9	7	7	10	9	9	8	9	6	9	8	7	6	5	7

within the set. Moreover, strategy OPENLOOP produces shorter curriculum (20 versus 40 produced by strategy PERC). The explanation is simple: the curriculum produced by OPENLOOP does not contain examination trials.

Analysing the letters which make "all the troubles" we can see that they all meet the prediction of our theoretical investigation above: they all satisfy the condition of inclusion. The pair of letters which satisfy that condition are : (C,O), (F,E), (J,U), (L,E), (P,R), for all the three sets (U,O) for the set IBM29, and (O,D) for the set VT50.

Series 2: The influence of the transfer of training

Although more extensive studies of this phenomenon has been performed [7]-[11], here we give just two examples of the experiments. The strategy used is PERC. First we have provided training for the set of letters VR14. After that, training for the set IBM29 gave the result

L(T*)=48 L(C*)=178 T* = ADGIMNOQTUW
ACDGHJKNOQRSUWX
CDGHIJMNPSVWYZ
ABEFKLMX

whereas the training for the set VT50 gave

L(T*)=43 L(C*)=199 T*= ADGIMNOQTUVW
ACDGHJKOQRSV
BCDEFGILPSVZ
DQUVY
JV

We see evidence of the positive transfer of training. The curriculum is much shorter then the one generated in the case of "tabula rasa" initial conditions.

Series 3: Linearly dependent patterns

In this experiment was chosen a set of 40 patterns from the set IBM40: 10 digits and the special signs "+","-","=",and "/" were augmented to the set of 26 letters, shown on Figure 8. Since each pattern is represented by a 35-dimensional vector, the patterns are neccessarily linearly dependent. Here is the result of the experiment using strategy PERC:

L(T*) = 204 L(C*) = 604

T* = ABCDEFGHIJKLMNOPQRSTUVWXYZ-+=/1234567890
CEFGHIJKLMNOPQRSTUVWXYZ-+=/1234567890
ABCDFJLOPQRSUWZ-/23567890
ABCDEFHIJKLMNPRSTUVXY-+=123890
DEFGHJKLMNOPQRUWZ-/1234567890
ABCIJPRSTVWXY+36
EFHKLMNPRSVZ-127
FOU15
CEGJL

As noted obove, this type of experiments is of interst in associative memory studies. LAA has solved this particular problem in 10 iterations.

Series 4: A multisample problem

The task is LAA to recognize all the letters from the Figure 8 as a letters from an alphabet. Thus, three samples pro class are given. Such a task is actually of interest in pattern classification theory. We will not give entire teaching sequence for this task, since it is inconvinient to represent for example the letter "A" in three different way, representing three different sets. However, the task was solved in 10 iterations, with the training sequence of the length L(T*)=207, and length of the entire curriculum L(C*)=987.

Discussion on the experiments

In the experimental investigation with LAA we have been mainly concerned on the pair-association task and the influence of similarity upon the training process. We can see that in pair-association, the most interesting case appears when there are patterns which are included into some other patterns. As predicted in the theoretical analysis, assuming homogeneous initial conditions, all other cases are one-trial-learning cases.

The multisample problem although not main interst in this study was also considered in the experiments above. The fact that the training process in the Series 4 has converged, shows that the sets of samples were linearly separable.

SOME IMPLICATIONS TO THE THEORY OF ADAPTIVE CLASSIFIERS

Here we discuss two issues of interest in the theory of linear classifiers which are consequence of our analysis above: the usage of augmented fixed coordinate in the set of input patterns, and the geometrical interpretation of the pattern recognition training.

In the theoretical discussion above it is obtained result that the input patterns are maximaly similar, i.e. not distinguishable, by a linear classifier which keeps the thresholds fixed ,if they are positively colinear. In the theory of the linear machines [31],[3],[12],[20] the training is

performed using the patterns $\underline{x}'=[1,\underline{x}]=[x_o=1,x_1,x_2,..,x_n]$ which, if different, are never colinear by definition! Note that with augmented 1, the threshold value wio=Oi of the discriminant function $g_i(\underline{x})$ is changed in each teaching step. From the teaching space approach it can be shown, that if the training patterns are with augmented 1, the relations analogeous to (15) are

$$(a_{ii}+1)p_1 > (a_{21}+1)p_k + r_{ki} \qquad (16.1)$$

$$(a_{kk}+1)p_k > (a_{12}+1)p_i + r_{ik} \qquad (16.2)$$

for which using (14.5) can be shown that there exists solution for any $\underline{x}_i$ and $\underline{x}_k$, providing $\underline{x}_i \neq \underline{x}_k$. Thus, it turns out that the augmented dimension with fixed value 1 assumed in the perceptron convergence theorem is an assumption that is not irrelevant to the analysis of the convergence of the training of the linear classifiers. If augmented dimension is not assumed, the problem of positive colinearity can still be avoided assuming that the vectors have unith length, [28] in which positive colinearity means identity. The problem of colinearity is here emphasized rather than avoided, searching for the similarity concept built by the learning machines.

The next problem we are considering now is the geometrical interpretation of the recognition problem. Besides many of the distinctions in the theory of adaptive pattern recognition, such as supervised vs. unsupervised learning, parameter vs. nonparameter learning, distribution vs. distribution-free learning, geometrical vs. statistical separation, we will emphasize the distinction as boundary vs. prototype learning. According to that distinction, the geometrical interpretation of the pattern recognition problem given in the theory of linear classifiers is based on the boundary learning paradigm, where the parameters to be learned are the coefficients of some hyperplane. Then the pattern classification training problem in the feature space is to find a hyperplane which will separate the samples of a given class

from the samples of an anotner. On the other side, cluster analysis techniques and so-called unsupervising learning recognition procedures are based on the prototype learning paradigm, where for a given set of samples and given some measure of similarity the system decides classification basing on the similarity of the unknown pattern to the prototypes of a class. The parameters learned in that paradigm are some statistics of the given reference data sets. In the sequel we give geometrical interpretation of the behaviour of the linear machines during learning and decision making in terms of prototype learning paradigm.

Figure 9

Let us consider the case where n classes are given, each represented by only one sample, and a linear machine which uses $\alpha(1,0)$ learning law. Assume $\underline{x}_i$ and $\underline{x}_k$ are the samples which we observe (Figure 9). Let $\underline{w}_{io}$ and $\underline{w}_{ko}$ represent the initial values of the vectors $\underline{w}_i$ and $\underline{w}_k$ which will contain the information about the properties of the classes C_i and C_k respectively. With those initial conditions, in the space are shown the reachable states, i.e the values of the vectors $\underline{w}_i$ and $\underline{w}_k$ during the teaching procedure. Let us suppose that the training process is completed and the vectors $\underline{w}_i$ and $\underline{w}_k$ are represented by the points as shown on Figure 9. Now the vectors $\underline{w}_i$ and $\underline{w}_k$ are prototypes, or templates, of the classes C_i and C_k developed during the teaching process. In the case when multiple samples for a particular class are given, which is actualy of interest in the linear machines behavior, the common prototype vector of a particular class C_i will be weighted sum of the sample vectors, constructed during the training process:

$$\underline{w}_i = \sum_{j=1}^{J_i} p_{ij}\underline{x}_{ij} + \underline{w}_{io} \qquad (17)$$

where pij is number of the appearence of the j-th sample $\underline{x}_{ij}$ of the i-th class within the training sequence. J_i is the number of samples given for the class Ci.

Observe now that the behavior of a linear classifier is not changed if it computes the functions

$$f_i(\underline{x}) = P_{\underline{x}}(\underline{w}_i) + \theta_i \qquad (18)$$

instead of $g_i(\underline{x})$ defined in (1), where $P_{\underline{x}}(\underline{w}_i)$ is the orthogonal projection of $\underline{w}_i$ onto $\underline{x}$. That means that if during an examination trial some, possibly unknown, vector $\underline{x}$ appears in the feature space, then all the class prototypes are projected toward $\underline{x}$. If the system is equithreshold, the pattern will be assigned the i-th class iff

$$P_{\underline{x}}(\underline{w}_i) > P_{\underline{x}}(\underline{w}_k)$$

for all $k \neq i$. Thus, a geometrical interpretation can be given in terms of the projections of the prototype vectors on the unknown vector $\underline{x}$. This is an alternative to the separating surfaces point of view.

However, it cannot be said that this interpretation is more appropriate then the separating surface interpretation: in general case when thresholding is involved, it is not. It is just another point of view, pointing out the properties that linear classifiers have in common with other pattern recognition processes: some measure of similarity between the defined prototypes and the unknown object.

SOME IMPLICATIONS TO CLUSTER ANALYSIS

In this section is briefly discussed the possibility of the application of the similarity measure derived in this paper in the cluster analysis.

Cluster analysis is a prototype based classification technique. One of the best describing metaphors for this technique is the universe in which some attraction, or gravity, force is apriori defined, not neccessarily in Newtonian sense. Given some reference galaxies, or star clusters, as prototypes and an attraction force, the universe is divided into regions. When a new, unknown object appears in this universe, all the objects "vote" by means of their attraction force. The process is highly parallel and competitive. The object will be attracted to the cluster which has shown the strongest attraction force. The specification of the regions can be given apriori by the set of data points (stars) and the predefined attraction force, or designed iteratively by some outside system.

In cluster anlysis the usual term for the attraction force mentioned above is some measure of similarity, heuristicaly predefined between the data points in the feature space. The measure of similarity is usualy some decreasing function of a distance function. Various functions has been proposed as a measure of similarity for use in cluster analysis which will be not reviewed here. The extensive review is given in [27],[41]. However, we will mention two principaly different measures.

Given two patterns $\underline{x}_1$ and $\underline{x}_2$, the "most obvious" [12] measure of similarity is the euclidean metric d_e, which using our notation $a_{12}=\langle \underline{x}_1, \underline{x}_2 \rangle$ can be defined

$$d_e^2(\underline{x}_1,\underline{x}_2) = a_{11}+a_{22}-2a_{12} \qquad (19)$$

There are a number of variations of this metric distance used, such as Manhattan metric , weighted euclidean metric, Mahalanobis and Bathacharya distance.

In some applications a nonmetric measure is useful. The most often used nonmetric measure of similarity is the cosine of the angle between the patterns:

$$s_1(\underline{x}_1,\underline{x}_2) = a_{12}/\sqrt{a_{11}a_{22}} \qquad (20)$$

This measure sometimes is referred as a "classical" measure of similarity [22].

It is easy to see that the measure of similarity (12) derived in this paper to the measure $s_1(\underline{x}_1,\underline{x}_2)$ is related by

$$s_2 = \cos \text{ arc } \left(\begin{bmatrix} |\underline{x}_1| \\ |\underline{x}_2| s_1 \end{bmatrix}, \begin{bmatrix} |\underline{x}_1| s_1 \\ |\underline{x}_2| \end{bmatrix} \right) \qquad (21)$$

Without extensive analysis, it is evident that if used as a measure of similarity, the coefficient s_2 shows disadvantage of greater computational complexity. However, let us mention two interesting properties:

The measure of similarity s_2 is senzitive to the elongation of the amplitude of a vector, whereas s1 is not. Formaly, if $\underline{x}'_1 = k.\underline{x}_1$ for some k>0, then $s_1(\underline{x}'_1,\underline{x}_2)=s_1(\underline{x}_1,\underline{x}_2)$ whereas $s_2(\underline{x}'_1,\underline{x}_2)\neq s_2(\underline{x}_1,\underline{x}_2)$.

Interestingly enough, although the similarity measure s_2 is a nonmetric measure, the metric intuitive description can be given in the n-dimensional teaching space in terms of the angles as a distance measures. Consider the 3-dimensional pyramide in the similarity space as shown on figure 4. The three vertex angles satisfy the the property: the sum of any two angles is

greater than the third vertex angle. Thus, used as distance measures, those angles satisfy triangle inequality. That observation can be generalized to an n-dimensional convex polyhidron in the teaching space: considering three edges, one can construct a pyramid where vertex angles will satisfy the triangle inequality.

SOME IMPLICATIONS FOR ADAPTIVE CONTROL THEORY

The teaching space approach to adaptive pattern recognition used here is based on relations which contain the parameters of the training process rather then the parameters of the learning process, as has been discussed above. The learning process is viewed through the point of view of the teacher as a direct reference model control [29] where the behaviour of the learner is guided by a reference model stored somehow in the memory of the controller.

Now we discuss some implication of the teaching space approach related to the control theory. We are especially interested in issues concerning open-loop vs. closed loop control in pattern recognition training.

Let us consider the experimental paradigm of closed loop training used in our experiments. Here we will refer to this paradigm as to the paradigm I. This paradigm is illustrated on figure 10.1.

Figure 10

In this paradigm we have clearly distinguished two types of experimental trials: examination trials and teaching trials. In an examination trial, the learner is presented with a pattern to which it responds according to the present state of its knowledge base. In the same time the teacher generates and tests its expectation about the learner's output. After that, consistent with some training strategy, the teacher eventually performs a training trial, in which it simultaneously presents the stimulus pattern and the desired output, which unconditionaly updates the knowledge base of the learner. Let us note that this is a closed-loop control paradigm. Block [5] called it "reinforced learning".

Let us consider another experimental paradigm, which will be called paradigm II, and is shown on figure 10.2. Teacher in each trial generates the stimulus pattern along with the desired output. The learner receives the stimulus first, generates its output, and then compares it to the desired output. (The delay shown on figure 10.2. compensates the feedback delay neccesary to assure appropriate behaviour of the learner in this paradigm). On the basis of that comparison, the learner itself decides whether it will update its knowledge base or not. The teacher in this paradigm has no evidence about the actions of the learner. It generates the patterns according some, possibly random order. It was shown [31],[28] that this paradigm will assure the convergence of the learning process of the learner, providing that it implements itself the perceptron training strategy. Note that this is clearly open loop control paradigm. Block [5] calls it "forced" learning.

Thus, concerning the successfulness of the training, both paradigms I and II are equivalent. The pattern recognition training problems can be considered to be either closed loop or open loop, depending of how much structural complexity each of the player in this cooperative game has.

In paradigm I the learner is passive, causal [13], learning system: it builds the map of the teacher's knowledge base under complete control of the teacher. On the other hand, the teacher is an active, teleological [13] i.e. goal seeking system [24], which performs search through the space of its possible actions in order to perform the successful and efficient training. This kind of teacher is assumed and analysed in this paper.

In the paradigm II teacher is a passive system. It can be viewed as a generator of stimuli not being affected by the actions of the learner. On the other hand, the learner is an active learning system, which builds a map of the teacher knowledge base using some kind of error correction procedure. Such a procedure can be fixed increment error correction procedure, or "perceptron learning rule" [12]. Good discussion concerning this paradigm is given in [4].

Another issue is the problem of optimality of the training process. In order to assure successful training in the paradigm II is to be assumed the training sequence of infinite length in which all the patterns appear infinitely many times [31],[3],[12]. Although the training process converges after a finite number of trials, the trainer has no evidence of the state of the learner's knowledge. In the closed loop paradigm I, that information is available, and an algorithm can be designed which will assure stopping the training procedure once the learner has reached the desired state.

It is our feeling that the adaptive pattern recognition theory has not paid enough attention to the goal seeking character of the teacher in the pattern recognition process. We believe that such an analysis can be fruitfull and relevant to the general problem of control.

SOME IMPLICATIONS TO NEURAL MODELING

In several places above we have confirmed the importance of a similarity and the applicability of that concept in the explanation of the behavior of linear machines. Now recalling that the dot product elements mentioned as a basic elements of a linear machine are often considered as basic models of the neurons, we can state a hypothesis that the neurons perform similarity judgment rather then logical functions as proposed by McCulloch and Pitts. Two points are consistent with this hypothesis: 1) the fact that brains are the best pattern recognition devices, and 2) some evidence of the existence of the feature detectors refered by the works of Hubel end Wiesel. Assuming the above hypothesis as true, we can imagine certain regions of the brain as a pattern recognition machine consisting of the elements which are themselves pattern recognition devices, performing similarity judgment on the basis of feature matching process described above. We will not pursue this hypothesis further. It sounds probably speculative, but we believe that in the time McCulloch and Pitts stated their logical computation hypothesis it had sound speculative. The level of comlexity that we assign as a function of the neurons seems to be increasing, parallely with the present concern of the science. It is to be expected that the future assign the neurons even more complex functions than "pure" similarity judgment.

CONCLUSIONS

The teaching space approach toward the problem of trainable linear pattern classifiers taken here touched two problems concerning pattern recognition and learning which have not been widely investigated in connection with these machines: the problem of interpattern similarity and the goal seeking nature of the teacher in the pattern recognition training.

The analysis in the similarity space has made transparent the fact that the perceptron training procedure considered in the proof of the perceptron convergence theorem is performed using the vectors which are never colinear: otherwise the convergence is not assured. That is expressed in the terms of maximal similarity between the reference patterns, measured by a similarity coefficient derived from the learning process of the linear machines. That similarity coefficient has not been proposed so far in either areas dealing with problems of similarity. The explanation is simple: all the models derived so far are results of the heuristics applied directly on the observation of the objects in the feature space, whereas this measure is derived from the physiology and the behavior of an artificial classifier during its learning process. Including the similarity concept in the theory of linear pattern classifiers , where the terms of linear separability, discriminant functions and boundary surfaces are the main concepts, to some extent bridges the gap between these artificial devices and the natural pattern recognition systems, where the similarity is a basic concept. Moreover, the similarity concept built by linear machines, although basically a feature matching process, turns out not to be so trivial as was thought. It turns out that the similarity concept built into the artificial systems has several features, as it has similarity concept built by the human system [44].

The analysis in the teaching space shows that the teacher in a pattern recognition process can be viewed as a goal seeking system. It searches through the space of its possible actions, using some searching strategy, having the goal to approach the solution region in the teaching space. It has been pointed out that this view is not neccessarily unique: the teaching problem can be also viewed as a open loop control problem, assumung that the learner has higher level of structural complexity, having at least one internal loop which allows him to know what action has been taken in the previous time step. However, the optimality of the training procedure gives credit to the closed loop variant of the training process used in our experiments.

APPENDIX: DETAIL OF THE SIMULATION EXPERIMENTS

```
TRAINER:
  Strategy:              @PERC
LEARNER:                 LAA
  Learning constants
    reward               1
    punishment           0
  Initial conditions:
    Threshold values:    all zero
    Memory matrix:       cleared
STIMULI:
   Set                   IBM29
   Patterns:             A,B,C,D,E,F

   experiment has been started !!
   ------------------------------

        iteration   1                          curriculum
 EXAMINATION  shown A recalled ?                  a
    TEACHING   A                                  A
 EXAMINATION  shown B recalled A                  b
    TEACHING   B                                  B
 EXAMINATION  shown C recalled B                  c
    TEACHING   C                                  C
 EXAMINATION  shown D recalled B                  d
    TEACHING   D                                  D
 EXAMINATION  shown E recalled C                  e
    TEACHING   E                                  E
 EXAMINATION  shown F recalled E                  f
    TEACHING   F                                  F

        iteration   2
 EXAMINATION  shown A recalled A                  a
 EXAMINATION  shown B recalled B                  b
 EXAMINATION  shown C recalled C                  c
 EXAMINATION  shown D recalled D                  d
 EXAMINATION  shown E recalled E                  e
 EXAMINATION  shown F recalled ?                  f
    TEACHING   F                                  F

        iteration   3
 EXAMINATION  shown A recalled A                  a
 EXAMINATION  shown B recalled B                  b
 EXAMINATION  shown C recalled F                  c
    TEACHING   C                                  C
 EXAMINATION  shown D recalled D                  d
 EXAMINATION  shown E recalled F                  e
    TEACHING   E                                  E
 EXAMINATION  shown F recalled ?                  f
    TEACHING   F                                  F
```

```
        iteration   4
EXAMINATION  shown A recalled F        a
   TEACHING   A                        A
EXAMINATION  shown B recalled F        b
   TEACHING   B                        B
EXAMINATION  shown C recalled C        c
EXAMINATION  shown D recalled B        d
   TEACHING   D                        D
EXAMINATION  shown E recalled F        e
   TEACHING   E                        E
EXAMINATION  shown F recalled ?        f
   TEACHING   F                        F

        iteration   5
EXAMINATION  shown A recalled A        a
EXAMINATION  shown B recalled B        b
EXAMINATION  shown C recalled E        c
   TEACHING   C                        C
EXAMINATION  shown D recalled D        d
EXAMINATION  shown E recalled F        e
   TEACHING   E                        E
EXAMINATION  shown F recalled ?        f
   TEACHING   F                        F

        iteration   6
EXAMINATION  shown A recalled A        a
EXAMINATION  shown B recalled E        b
   TEACHING   B                        B
EXAMINATION  shown C recalled E        c
   TEACHING   C                        C
EXAMINATION  shown D recalled B        d
   TEACHING   D                        D
EXAMINATION  shown E recalled E        e
EXAMINATION  shown F recalled F        f

        iteration   7
EXAMINATION  shown A recalled A        a
EXAMINATION  shown B recalled B        b
EXAMINATION  shown C recalled C        c
EXAMINATION  shown D recalled D        d
EXAMINATION  shown E recalled E        e
EXAMINATION  shown F recalled F        f

end of the experiment
```

REFERENCES

[1]. Amari S-I. A mathematical approach to neural systems. In Metzler (Ed.) Systems Neuroscience, pp.67-117, Academic Press, 1977

[2]. Anderson J.A., Silverstein J.W., Ritz S.A., & Jones R.S. Distinctive features, categorical perceptrion, and probability learning: Some applications of neural model. Psychological Review, 84: 413-451, 1977

[3]. Andrews H.C. Introduction to Mathematical Techniques in Pattern Recognition Wiley-Interscience, 1972

[4].Barto A.G, & Sutton R.S. Goal seeking components for adaptive intelligence: An initial assesment. Tehnical report AFWAL-TR-81-1070, Wright Peterson Air Force Base, 1981

[5]. Block H.D. The perceptron: A model for brain functioning.I. Reviews of Modern Physics, 34:123-135, 1962

[6]. Block H.D.,Knight B.W, & Rosenblatt F. Analysis of a four-layer series-coupled perceptron. II. Reviews of Modern Physics, 34:135-142, 1962

[7]. Bozinovski S. The perceptron and a possibility of simulation of the training process. M.Sc. Thesis (In Serbocroatian). University of Zagreb, Zagreb, 1974

[8]. Bozinovski S. & Fulgosi A. The influence of similarity and the transfer of training upon the perceptron training.(In Serbocroatian) Proc. Symp.Informatica XX:3-12-1-5 1976

[9]. Bozinovski S., Santic A., & Fulgosi A. The normal training strategy in the pair association in the paradigm Teacher:human-Learner:machine.(In Serbocroatian) Proc. Conf. ETAN XXI IV:341-346, 1977

[10]. Bozinovski S. Some experiments in non-biological systems training. (In Macedonian) Proc. Conf. ETAN XXII IV:371-379, 1978

[11]. Bozinovski S. The influence of coincidence of the features of the patterns in adaptive pattern recognition. Ph.D. Thesis, (In Serbocroatian), University of Zagreb, 1981

[12]. Duda R.O & Hart P.E. Pattern Classification and Scene Analysis Willey-Interscience, 1973

[13]. Eckman D.P. & Mesarovic M.D. On some basic concepts of the general systems theory, Proc. 3rd International Congress on Cybernetics, pp.104-118, Namur, Belgium, 1961

[14]. Friedman L. A model of goal-directed behavior. In Artificial Intelligence pp. 66-82 Preprint of papers presented at Arificial Intelligence session, IEEE, January, 1963

[15].Gagne' R.M., Baker K.E., & Foster H. On the relation between similarity and transfer of training in the learning of discriminative motor tasks, The Psychological Review, 57:67:79, 1950

[16]. Gilmore G.C, Hersh H., Caramazza A., & Griffin J. Multidimensional letter similarity derived from recognition errors. Perception and Psychophyasics 25 :425-431, 1979

[17]. Girardeau F.L. Paired-associate learning: Similarity among stimulus terms. Psychosomic Science, 4(12):423-424, 1966

[18]. Glushkov V.M. The theory of instruction for a class of discrete perceptrons Zhurnal vychislitelnoi matematiki i matematiceskoi fiziki 2,pp.317-335, 1962

[19]. Hirai Y. A template matching model for pattern recognition: Self-organization of templates and template matching by a disinhibitory neural network. Biological Cybernetics 38:91-101, 1980

[20]. Hunt E.B. Artificial Intelligence, Academic Press, 1975

[21]. Hutchinson J.W. & Lockhead G.R. Similarity as distance: A structural principle for semantic memory. Journal of Experimental Psychology: Human Learning and Memory 3:660-678, 1977

[22]. Iijima T.& Genchi K. A theory of character recognition by pattern matching method. In Fu & Tou (Eds.) Learning Systems and Intelligent Robots Plenum Press, 1974

[23]. Ivakhnenko A.G. (Ed.) Perceptron - A Pattern recognition System (In Russian) Naukova Dumka, Kiev, 1975

[24]. Klopf A.H. Goal-seeking systems from goal seking components: Implications for AI. The Cognition and Brain Theory Newsleter, 3, 1979

[25]. Kohonen T. Associative Memory: A System-Theoretic Approach, Springer Verlag, 1977

[26]. Krumhansl C.L. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological review 85:445-463, 1978

[27]. Mardia K.V.,Kent J.T., & Bibby J.M. Multivariate Analysis. Academic Press, 1979

[28]. Minsky M. & Papert S. Perceptrons, MIT Press, 1969

[29]. Monopoli R.V. Model reference adaptive control with an augmented error signal IEEE Trans. Automatic Control, 19:474-484, 1974

[30]. Nakano K. Associatron - A model of associative memory. IEEE Trans. Systems, Man, and Cybernetics, 2:380-388, 1972

[31]. Nilsson N.J. Learning machines. McGraw-Hill, 1965

[32]. Nilsson N.J. Adaptive pattern recognition: A survey, In Oestreicher & Moore (Eds.) Cybernetic Problems in Bionics, pp.103-145,Gordon & Breach Science Publishers, New York, 1968

[33]. Ortony A. Beyond literal similarity. Psychological Review 86:161-180,1979

[34]. Podgorny P.& Garner W.R. Reaction time as a measure of inter- and intraobject similarity: Letters of alphabet. Perception & Psychophysics,26(1):37-52,1979

[35]. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65:386-408, 1958

[36]. Rosenblatt F. Perceptual generalization over transformation groups In Yovits & Cameron (Eds.) Self-Organizing Systems, pp.63-100, Pergamon Press, 1960

[37]. Rosenblatt F. Principles of Neurodinamics, Spartan Books, 1962

[38]. Sebestyen G.S. Decision Making Process in Pattern Recognition Macmillan, New York, 1962

[39]. Selfridge O.G. Pandemonium: A paradigm for learning. In Mechanisation of Thought Processes pp. 511-531, H.M.S.O. London, 1959

[40]. Shulman L.S. Psychological controversies in the teaching of science and mathematics. In Clarizio, Craig, & Mehrens (Eds.) Contemporary Issues in Educational Psychology pp.188-205, Allyn and Bacon, 1974

[41]. Spath H. Cluster Analysis Algorithms Ellis Norwood, 1980

[42]. Spinelli D.N. OCCAM: A computer model for a content addressable memory in the central nervous system. In Pribram & Broadbent Biology of memory pp.293-306, Academic Press, 1970

[43]. Steinbuch K. The learning matrix.(In German), Kybernetik, 1:36-45, 1961

[44]. Tversky A. Features of similarity. Psychological Review 84:327-352, 1977

[45]. Vanderplas J.M. Transfer of training and its relation to perceptual learning and recognition. Psychological Review, 65:375-385, 1958

[46]. Wigstrom H. A neural model with learning capability and its relation to mechanisms of association. Kybernetik, 12:204-215, 1973

FIGURE CAPTIONS

Figure 1. Adaptive classifier

Figure 2. Two-dimensional teaching space. Initial conditions of the learner are not homogenious. Marked points are the goal states.
2a): A case when $a_{12} > 0$. Angles α_1 and α_2 are positive and $\psi < 90°$.
2b): A case when $a_{12} < 0$; Angles α_1 and α_2 are negative and $\psi > 90°$.

Figure 3. Similarity space. Two vectors (a_{11}, a_{21}) and (a_{12}, a_{22}) are considered at a time. The angle between those vectors is a measure of dissimilarity between the patterns $\underline{x}_1$ and $\underline{x}_2$. The shaded area is the square of the inner product, a_{12}. Five cases of the relationship between the patterns are shown: Fig. 3a): anticolinear Fig 3b): $a_{12} < 0$, in general, Fig 3c): orthogonal Fig 3d): $a_{12} > 0$, in general, Fig 3e): positively colinear.

Figure 4. Three-dimensional teaching space. Initial conditions of the learner are homogenious. The goal points are inside the open ended pyramide with the vertex in the coordinate origin. The projections of the pyramide on the principal coordinate planes are solution regions for each pair of inequalities (15). The figure represents particular case of the example 1.

Figure 5. A nonlinear associative memory. Linear machine with equi-threshold values can be viewed as an associative memory which outputs are input to a maximum selector. This system is used in the experimental investigation.

Figure 6. Serch through a two-dimensional teaching space. Given two patterns, the angle β is defined. Four systems starting from different initial conditions search for the goal points using different strategies.

Figure 7. Openloop versus closed loop teaching strategy. Due to the initial conditions, the solution region is so translated that strategy OPENLOOP will never reach the goal point, since the patterns chosen satisfy the inclusion condition (search A). The strategy PERC will reach the goal point (serch B).

Figure 8. The sets of stimuli used in the experimental investigation.

Figure 9. Prototype concept applied to the theory of linear machines. $\underline{x}_i$ and $\underline{x}_k$ are reference samples for the class C_i and C_k respectively. Prototypes $\underline{w}_i$ and $\underline{w}_k$ are builded during the training period. If, during the examination period some, possibly unknown pattern $\underline{x}$ appears in the feature space, all the prototypes are projected toward $\underline{x}$.

Figure 10. Closed-loop (Fig 10a) versus open-loop (Fig 10b) training for pattern recognition.

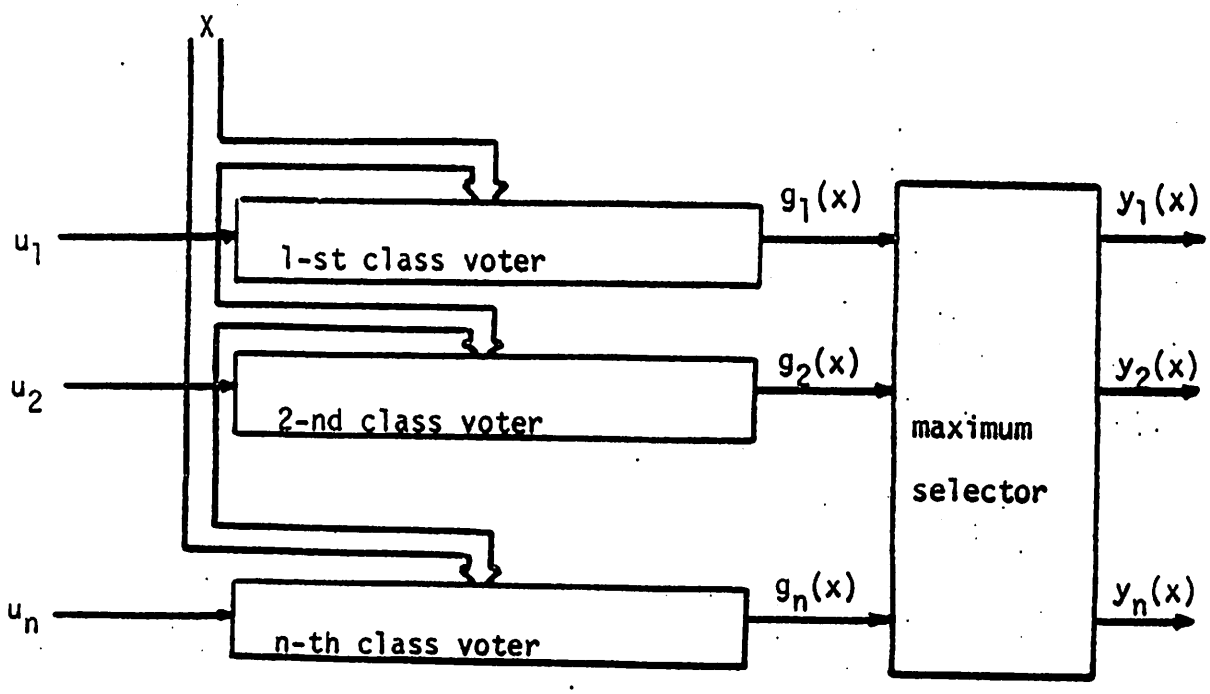
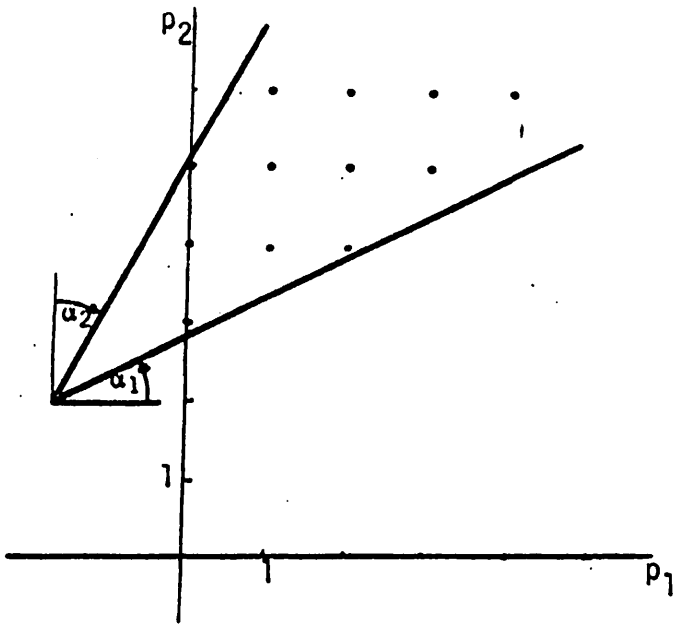
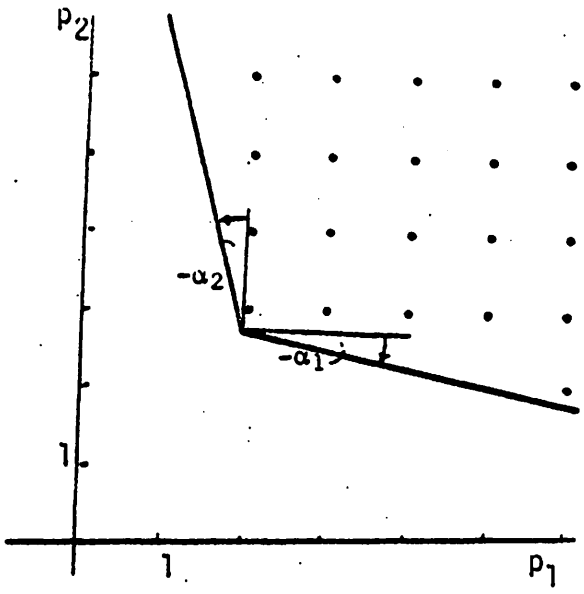

X
u_1
u_2
u_n
1-st class voter
2-nd class voter
n-th class voter
$g_1(x)$
$g_2(x)$
$g_n(x)$
maximum
selector
$y_1(x)$
$y_2(x)$
$y_n(x)$

Fig. 1

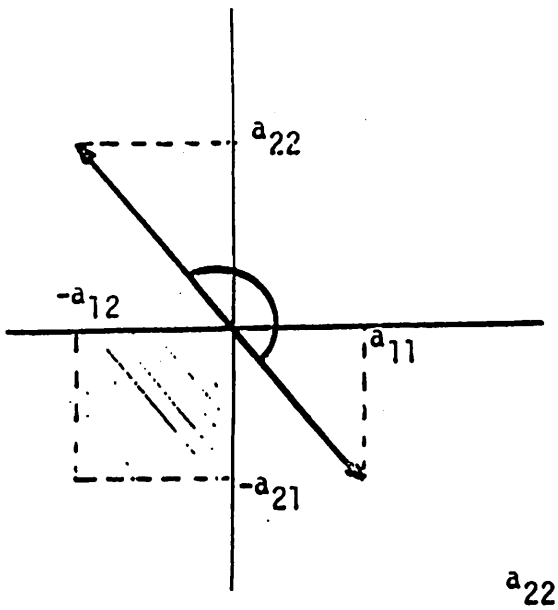


a

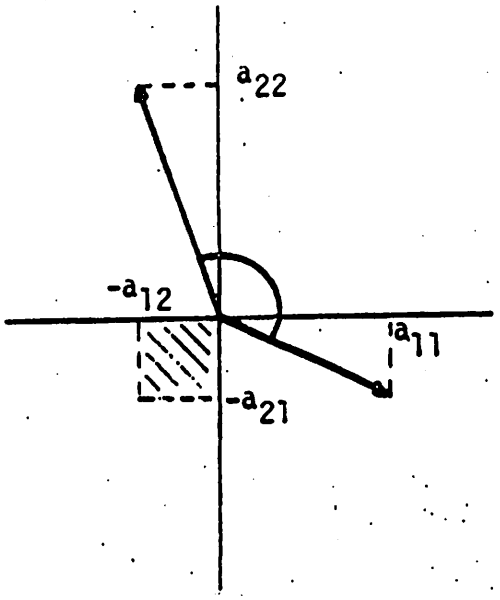


b

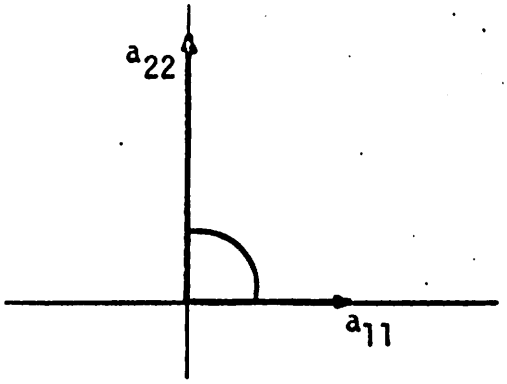
fig 2


a_{22}
$-a_{12}$
a_{11}
$-a_{21}$

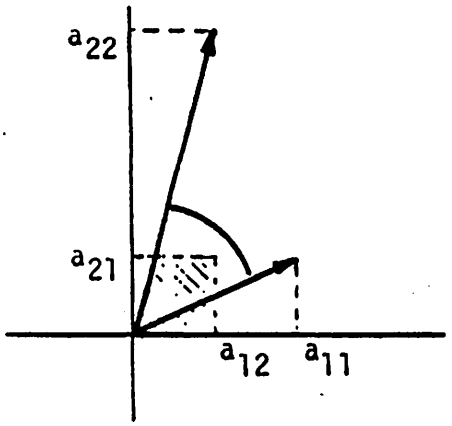
a


a_{22}
$-a_{12}$
a_{11}
$-a_{21}$

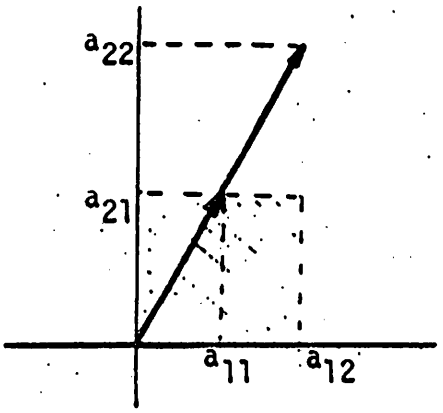
b


a_{22}
a_{11}

c


a_{22}
a_{21}
a_{12}
a_{11}

d


a_{22}
a_{21}
a_{11}
a_{12}

e

Fig 5

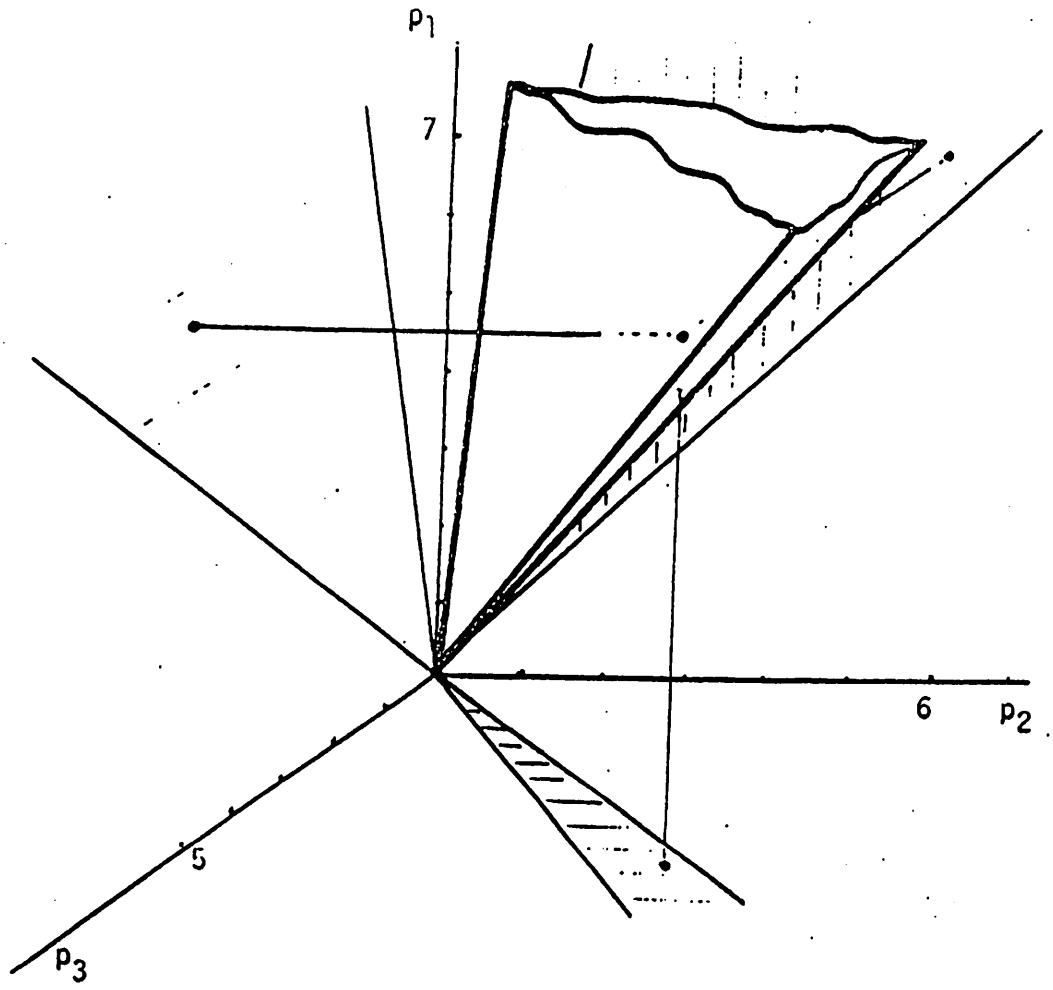


Fig 4

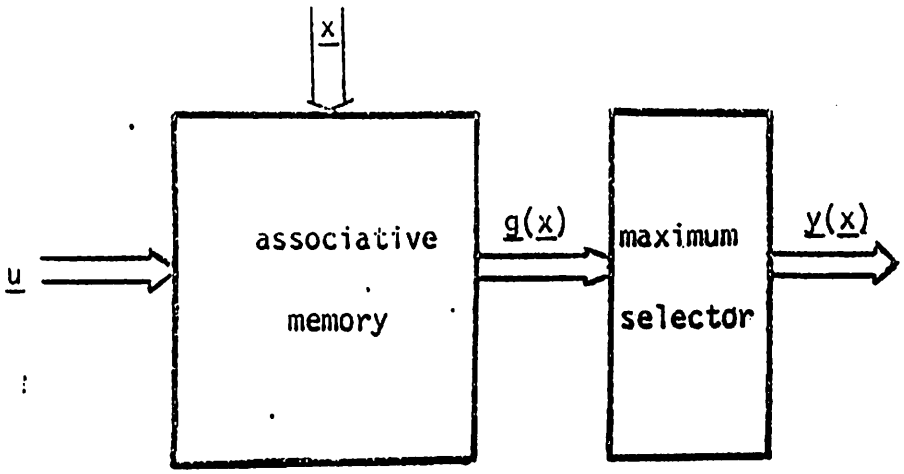


Fig 5

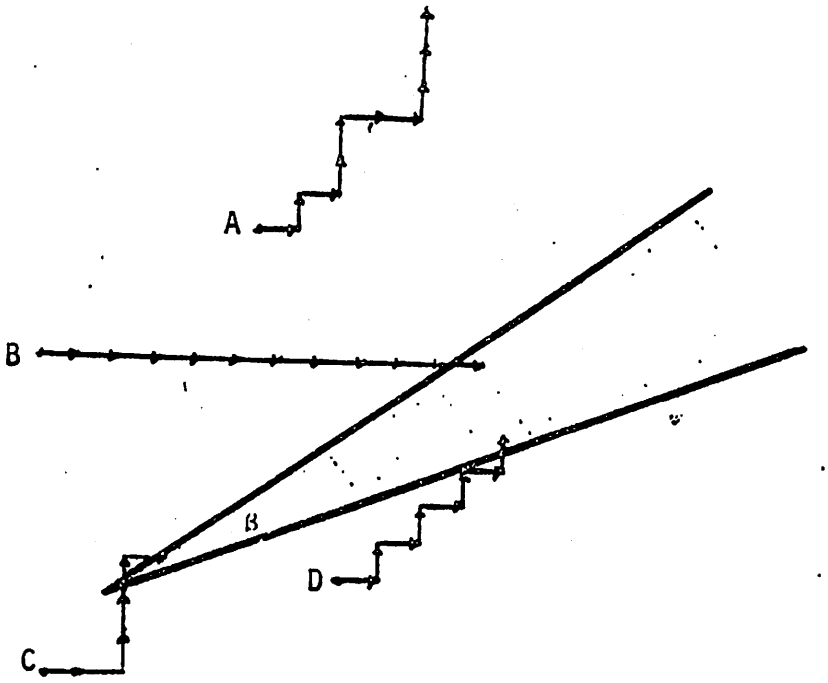
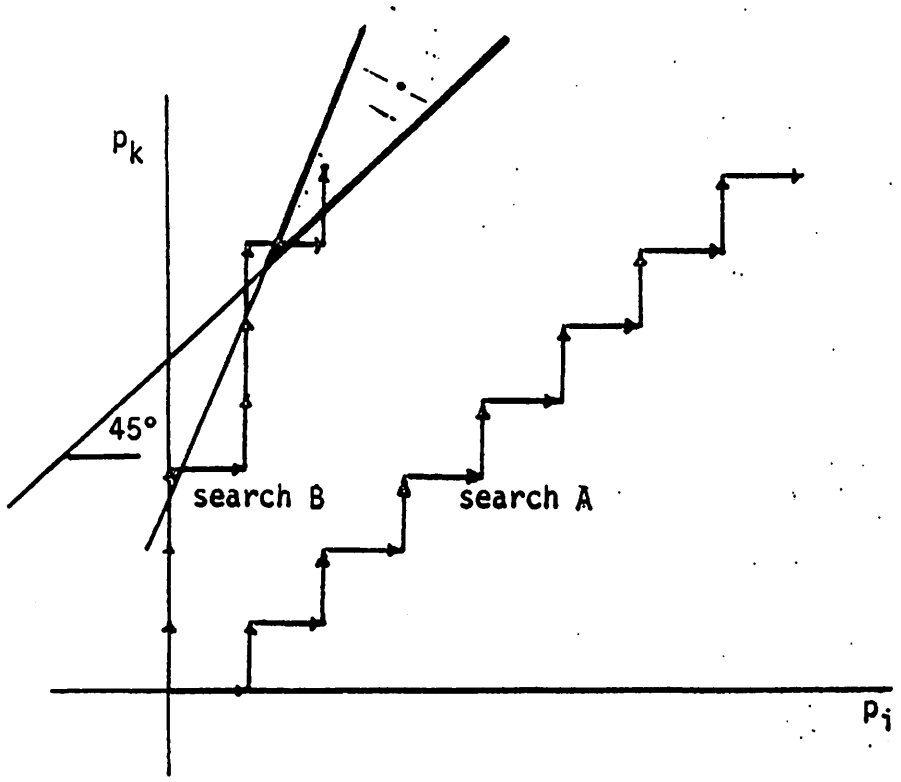
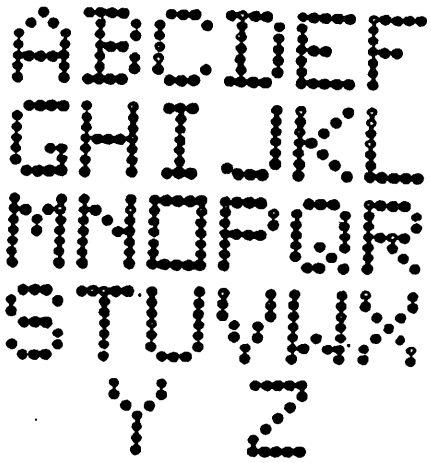
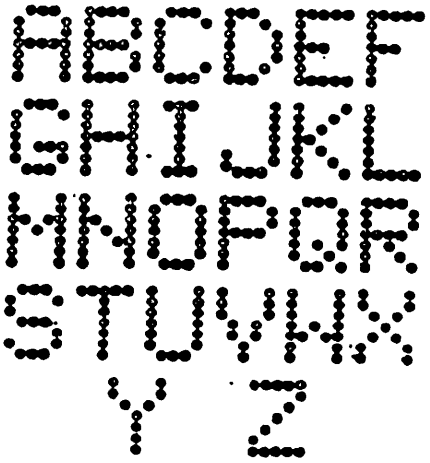


fig 6

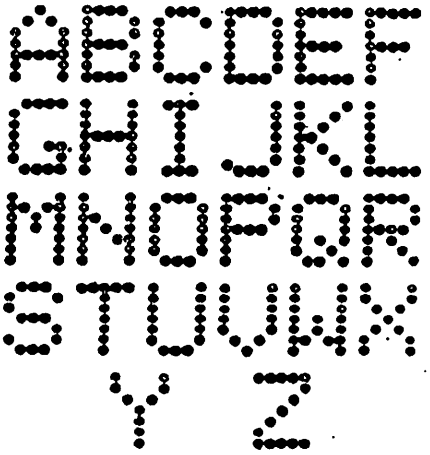

p_k
45°
search B
search A
p_i



IBM29



VR14



VT50

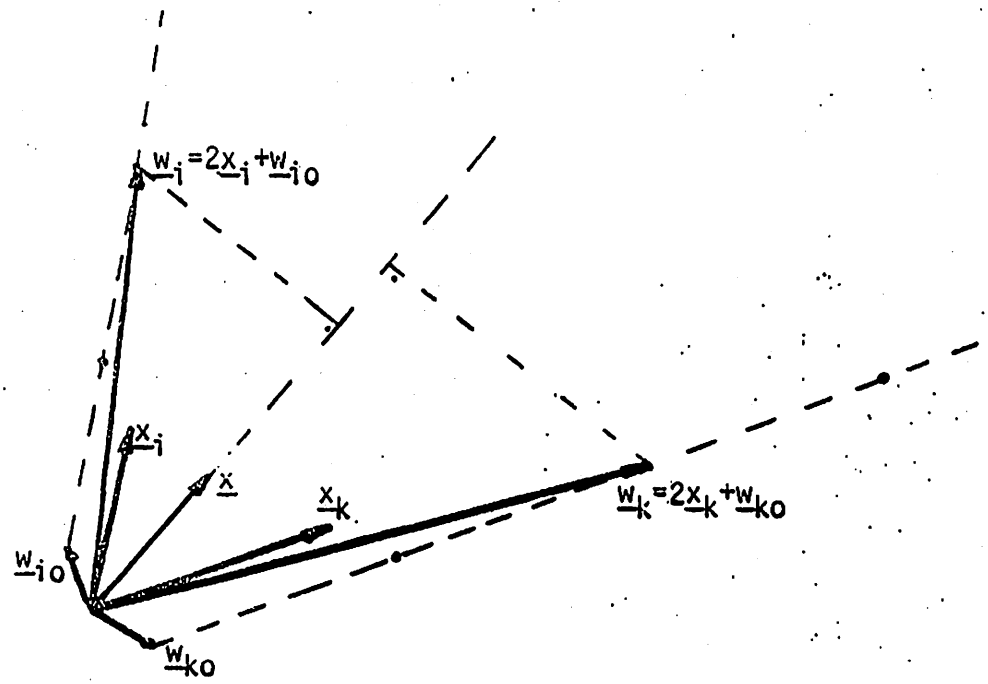
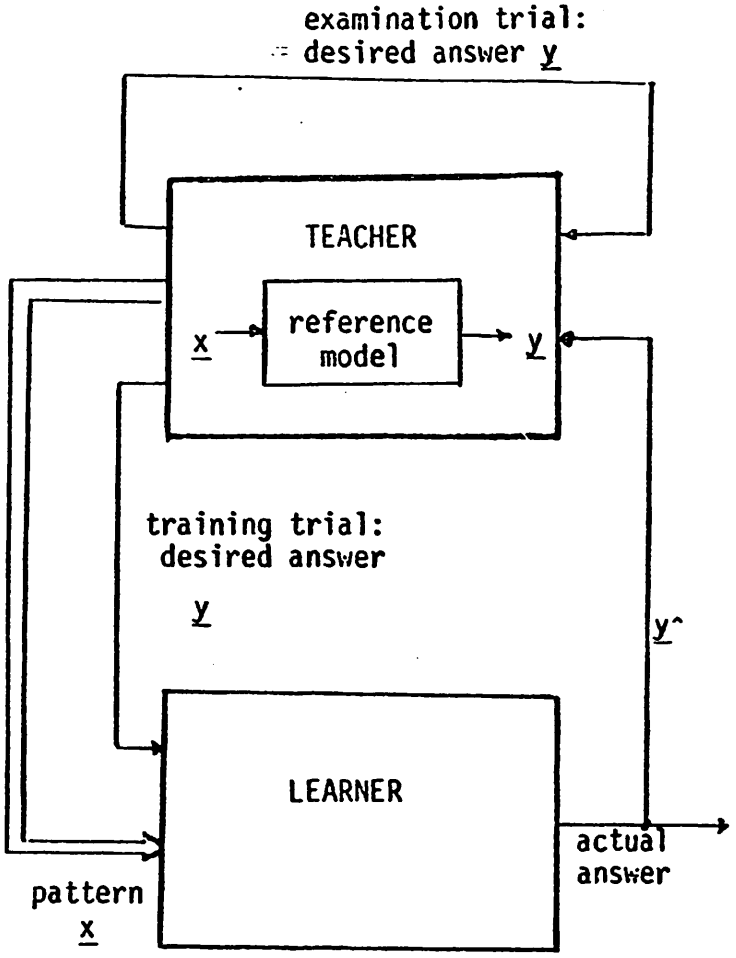
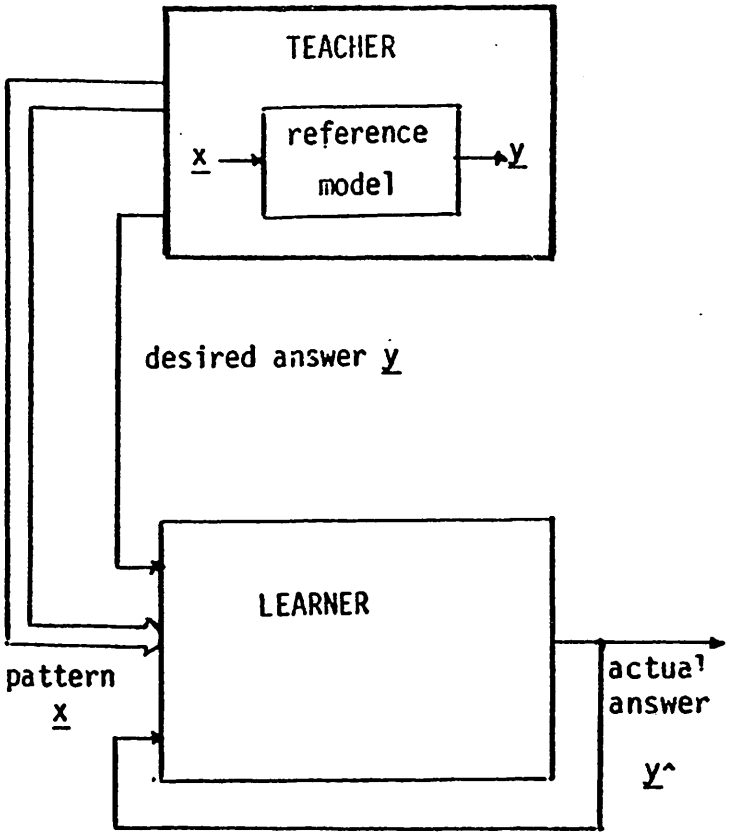

$\underline{w}_i = 2\underline{x}_i + \underline{w}_{io}$
$\underline{x}_i$
$\underline{x}$
$\underline{x}_k$
$\underline{w}_k = 2\underline{x}_k + \underline{w}_{ko}$
$\underline{w}_{io}$
$\underline{w}_{ko}$

Fig 3


examination trial:
desired answer y
TEACHER
x
reference model
y
training trial:
desired answer
y
y^
LEARNER
pattern
x
actual
answer

a


TEACHER
x
reference model
y
desired answer y
LEARNER
pattern
x
actual
answer
y^

b

The author is with the
Cybernetics Department
Electrical Engineering Faculty
University of Skopje
Karpos II, 91000 Skopje
Yugoslavia

Any comments on this paper
are greatly appreciated.