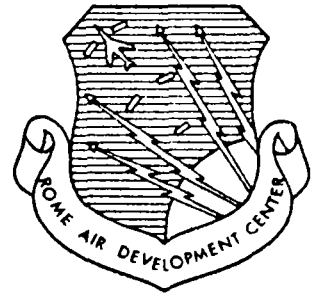


RADC-TDR-64-32

Nils J. Nilsson  
Computer Science Department  
Stanford University  
Stanford, CA. 94305-4110



CLASSIFICATION AND GENERALIZATION CAPABILITIES  
OF LINEAR THRESHOLD UNITS

TECHNICAL DOCUMENTARY REPORT NO. RADC-TDR-64-32

February 1964

Information Processing Branch  
Rome Air Development Center  
Research and Technology Division  
Air Force Systems Command  
Griffiss Air Force Base, New York

Project No. 5581, Task No. 558104

(Prepared under Contract AF30(602)-2943 by Thomas M. Cover,  
Stanford Research Institute, Menlo Park, California)

## DDC AVAILABILITY NOTICE

Qualified requesters may obtain copies from the Defense Documentation Center (TISIR), Cameron Station, Alexandria, Va., 22314. Orders will be expedited if placed through the librarian or other person designated to request documents from DDC.

DDC release to OTS is authorized.

## LEGAL NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

## DISPOSITION NOTICE

Do not return this copy. Retain or destroy.

## FOREWORD

The author wishes to express his thanks to N. Abramson, N. Nilsson, and A. Novikoff for many valuable discussions. He wishes to thank B. Brown, J. Koford, and B. Widrow for formulation of several related problems which led to the studies in this report. The research culminating in this report was begun with B. Efron, with whom a joint paper is in preparation.

Vertical text or artifacts along the right edge of the page, possibly bleed-through from the reverse side.

Key Words: Artificial Intelligence, Learning Machines, Adaptive Mechanisms, Pattern Recognition, Linear Threshold Units


### ABSTRACT

This report represents work in progress on properties of linear threshold functions. In a  $d$  dimensional binary space there exists  $n$  separate points ( $N = 2^d$ ). Furthermore there exists  $2^N$  possible combinations (dichotomy) of these points. Not all of these combinations can be separated by linear threshold functions. This paper concerns itself with determining which combination can or cannot be separated. Surfaces other than hyperplanes are also studied. These include surfaces obtained by multiple linear threshold devices and quadratic surfaces. Consideration is also given to training procedures in the separation of random patterns by linear threshold devices.

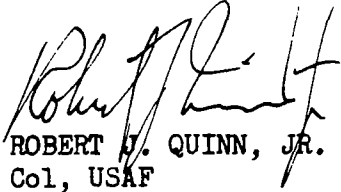
### PUBLICATION REVIEW

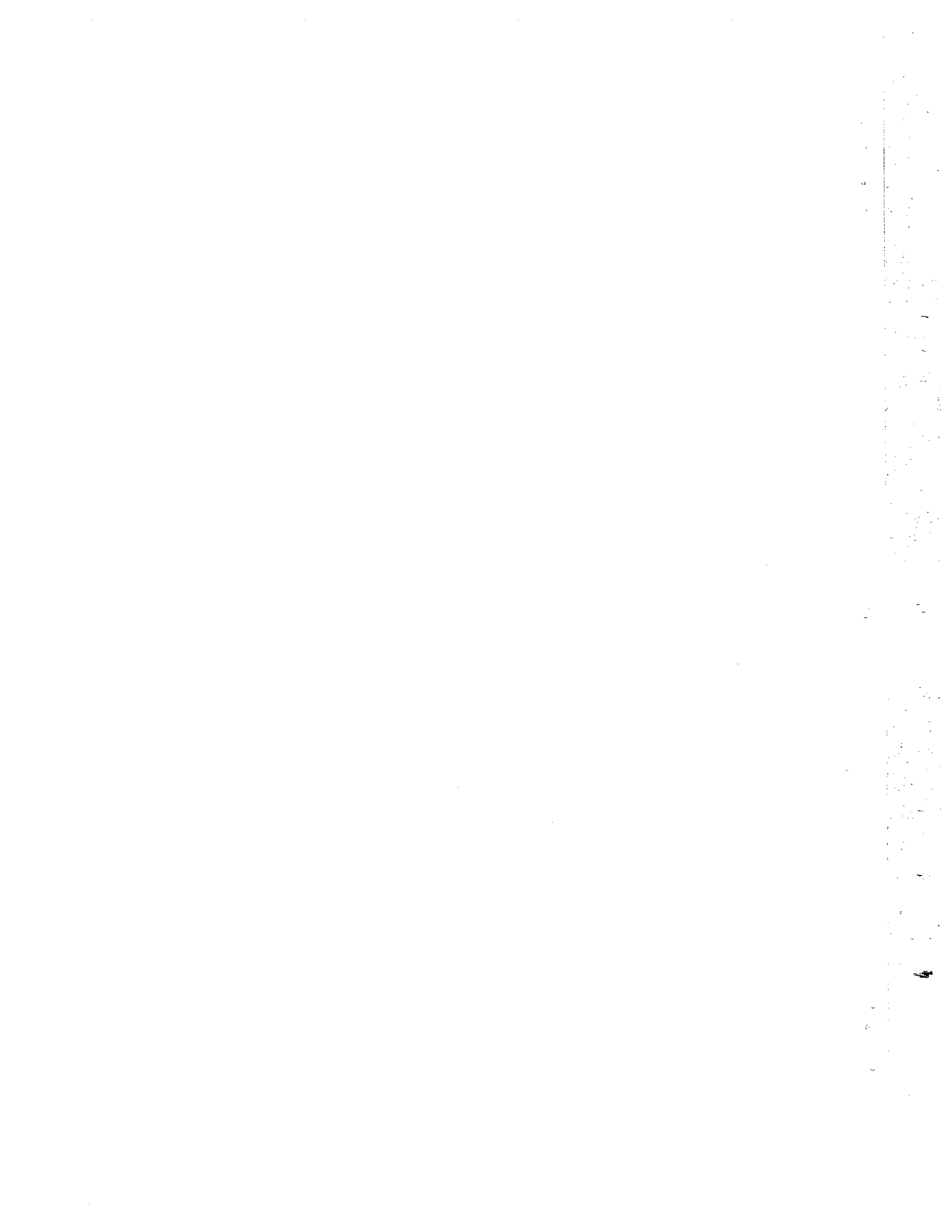
This report has been reviewed and is approved. For further technical information on this project, contact

Approved:

  
FRANK J. TOMAINI  
Chief, Info Processing Branch

Approved:

  
ROBERT J. QUINN, JR.  
Col, USAF  
Chief, Intel and Info Processing Div



## I INTRODUCTION AND SUMMARY

Consider a set of patterns which is represented by a set of vectors in a  $d$ -dimensional space. A homogeneous linear threshold function  $f$  is defined on this space:

$$f(x) = \begin{cases} 1, & w \cdot x > 0 \\ -1, & w \cdot x < 0 \\ 0, & w \cdot x = 0 \end{cases} \quad (1)$$

Any function of this form has a simple implementation indicated schematically in Fig. 1. Every homogeneous linear threshold function naturally dichotomizes the set of pattern vectors into two sets, the set of vectors  $x$  such that  $f(x) = 1$  and the set of vectors such that  $f(x) = -1$ . Geometrically, the two sets of pattern vectors are separated by the hyperplane

$$[x : f(x) = 0, \quad x \in R^d] \quad (2)$$

There are  $2^N$  functions taking values  $\pm 1$  on  $N$  points. Each function corresponds to a dichotomy of the  $N$  points. In general, there are fewer than  $2^N$  homogeneous linear threshold functions on  $N$  points. Section II is devoted to a short history of the problem of counting the number of homogeneous linear threshold functions.

There are many classes of separating surfaces other than hyperplanes which are tractable analytically and easily implemented by augmented

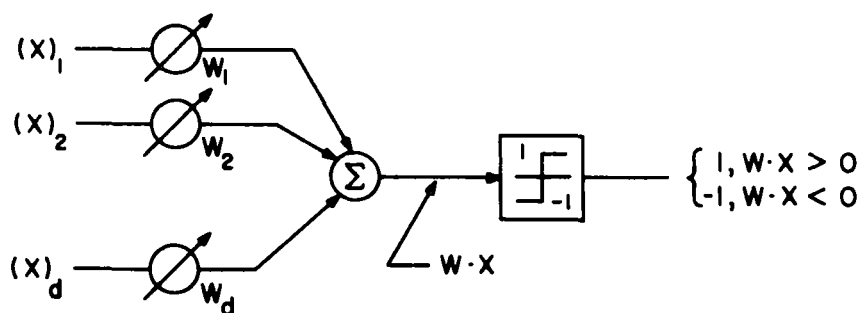


FIG. 1 HOMOGENEOUS LINEAR THRESHOLD UNIT AND IMPLEMENTATION OF SEPARATING HYPERPLANE

linear threshold devices. For example, in Section III, the results of Section II are generalized in order to count the number of ways in which hyperspheres, hypercones, and quadric surfaces can divide  $N$  points into two sets.

In Section IV, it is found that a linear threshold device has a natural separating capacity of two random patterns per adaptive weight, thus verifying experimental work by Koford<sup>1\*</sup> and Brown.<sup>2</sup> From the considerations in Section IV, it can be shown that a set of  $2d$  random inequalities in  $d$  unknowns has a solution with probability one-half.

One possible definition of the process of generalization is offered in Section V, in which it is shown that a large number of patterns must be used in training a linear threshold unit in order to ensure a high probability of unambiguous response to additional unknown patterns.

This report represents work in progress on properties of linear threshold functions. Future reports will concern:

- (1) A study of the inequality constraints on the weight vector  $w$  in Eq. (1).
- (2) Counting the number of dichotomies of  $N$  points in  $d$ -space that are separable with  $r$  or fewer errors.
- (3) Generalizing the constraint of general position in the theorems of Section II.
- (4) A study of capacities of networks of linear threshold devices.

---

\* References are listed at the end of the report.



## II SEPARATION THEOREMS

The following theorem is due to Wendel:<sup>3</sup>

*Theorem 1.* Let  $N$  points be randomly scattered on the surface of a unit sphere in  $d$ -space. Then the probability  $P_{N,d}$  that there exists some hemisphere containing all of the points is

$$P_{N,d} = \left(\frac{1}{2}\right)^{N-1} \sum_{i=0}^{d-1} \binom{N-1}{i} \quad (3)$$

The following theorem emphasizes the fact that the underlying problem is combinatorial rather than probabilistic:

*Theorem 2.* Let  $x_1, \dots, x_N$  be vectors in  $d$ -space such that no  $d$  of them are linearly dependent. Of the  $2^N$  possible sets  $\{\pm x_1, \dots, \pm x_N\}$  formed by assigning plus and minus signs, exactly  $C_{N,d}$  sets have the property that the entire set lies in some half space, where

$$C_{N,d} = 2 \sum_{i=0}^{d-1} \binom{N-1}{i} \quad (4)$$

Theorem 2 has been proved by numerous authors, but Winder,<sup>4</sup> Cameron,<sup>5</sup> Perkins, Whitmore, and Willis,<sup>6</sup> and Joseph<sup>7</sup> have emphasized the application of Theorem 2 to the problem of counting the number of linearly separable partitions of a set. All the authors listed above have used a variant of a proof, which appears in Schläfli,<sup>8</sup> of Theorem 3 or its dual statement Theorem 3':

*Theorem 3.*  $N$  hyperplanes in general position passing through the origin of  $d$ -space divide the space into  $C_{N,d}$  regions.

*Theorem 3'.* A  $d$ -dimensional subspace in general position in  $N$ -space intersects  $C_{N,d}$  orthants.

A set of hyperplanes is in general position in Theorem 3 if every  $d$  element subset of the set of normal vectors defined by the hyperplanes

is linearly independent. The corresponding dual definition of general position for Theorem 3' is less clearly stated. A  $d$ -dimensional subspace is in general position in  $R^N$  if the zero vector is the only element in the intersection of this subspace with any  $d$ -dimensional coordinate axis.

A proof of Theorem 3 will be sketched. See Schläfli,<sup>8</sup> Cameron,<sup>5</sup> Winder,<sup>4</sup> and Wendel<sup>3</sup> for similar treatments. Let  $C_{N,d}$  be the number of regions formed by  $N$   $(d-1)$ -dimensional subspaces in general position in  $d$ -space. Consider a  $(N+1)$ th  $(d-1)$ -dimensional subspace. It is intersected by each of the  $N$  subspaces in a  $(d-2)$ -dimensional subspace. The  $N$   $(d-2)$ -dimensional subspaces maintain general position. Hence they divide the new subspace into  $C_{N,d-1}$  regions. Thus the  $(N+1)$ th subspace intersects  $C_{N,d-1}$  of the  $C_{N,d}$  regions, forming  $C_{N,d-1}$  new regions. The total number of regions formed by  $N+1$  planes is then given by the recurrence relation

$$C_{N+1,d} = C_{N,d} + C_{N,d-1} \quad (5)$$

Using the obvious boundary conditions

$$C_{N,1} = 2$$

and

$$C_{1,d} = 2 \quad (6)$$

it is easily verified that  $C_{N,d}$  is given by Eq. (4).

Finally in Theorem 3'' we shall state Theorem 3 in an alternative equivalent form and present a new proof.

*Definition.* A dichotomy  $\{X^+, X^-\}$  of a set  $X = \{x_1, \dots, x_N \in R^d\}$  is linearly separable if and only if there exists  $w \in R^d$  such that

$$\begin{aligned} x \cdot w &> t, & x \in X^+ \\ x \cdot w &< t, & x \in X^- \end{aligned} \quad (7)$$

The dichotomy  $\{X^+, X^-\}$  is said to be *homogeneously linearly separable* if it is linearly separable with  $t = 0$ . A vector  $w$  satisfying Eq. (7) will be called a *solution* or *separating vector* and the corresponding

surface  $\{x : x \cdot w = t\}$  will be called the *separating hyperplane*. If  $t = 0$  we have a *separating hyperplane through the origin* and a *homogeneous solution vector*.

The following lemma is geometrically obvious. It will be used in the proof of Theorem 3" and will be applied in the section on generalization:

*Lemma.* Let  $\{X^+, X^-\}$  be a dichotomy of  $X = \{x_1, x_2, \dots, x_N \in R^d\}$  and  $x_{N+1}$  be a point in  $R^d$ . Then  $\{X^+ \cup \{x_{N+1}\}, X^-\}$  and  $\{X^+, X^- \cup \{x_{N+1}\}\}$  are both homogeneously linearly separable if and only if  $\{X^+, X^-\}$  is homogeneously linearly separable by a hyperplane through  $x_{N+1}$ .

*Proof:*

The dichotomy  $\{X^+ \cup \{x_{N+1}\}, X^-\}$  is homogeneously linearly separable if and only if there exists  $w$  such that

$$\begin{aligned} w \cdot x &> 0, & x \in X^+ \\ w \cdot x_{N+1} &> 0, \\ w \cdot x &< 0, & x \in X^- \end{aligned} \quad (8)$$

and  $\{X^+, X^- \cup \{x_{N+1}\}\}$  is homogeneously linearly separable if and only if there exists  $w$  such that

$$\begin{aligned} w \cdot x &> 0, & x \in X^+ \\ w \cdot x_{N+1} &< 0, \\ w \cdot x &< 0, & x \in X^- \end{aligned} \quad (9)$$

Using the connectedness of the open set  $\{w : w \cdot x > 0, x \in X^+; w \cdot x < 0, x \in X^-\}$  of separating vectors for  $\{X^+, X^-\}$  and the continuity of the inner product, we see that Eqs. (8) and (9) hold if and only if there exists a vector  $\hat{w} \in R^d$  separating  $\{X^+, X^-\}$  such that

$$\hat{w} \cdot x_{N+1} = 0 \quad (10)$$

Then the hyperplane  $\{v : v \cdot \hat{w} = 0\}$  separates  $\{X^+, X^-\}$  and contains the point  $x_{N+1}$ .

*Theorem 3''.* Let  $X$  be a set of  $N$  vectors in  $d$ -space, every  $d$  of which are linearly independent. Then  $X$  has  $C_{N,d}$  homogeneously linearly separable dichotomies, where  $C_{N,d}$  is given in Eq. (4).

*Proof:*

Let  $C_{N,d}$  be the number of homogeneously linearly separable dichotomies of the set  $X = \{x_1, x_2, \dots, x_N \in \mathbb{R}^d\}$ . Consider a new point  $x_{N+1}$  in general position with respect to  $X$ , and consider a dichotomy  $\{X^+, X^-\}$  of  $X$ . The new point can always be joined to at least one of the elements of the dichotomy in order to form a separable dichotomy of  $\{x_1, x_2, \dots, x_N, x_{N+1}\}$ . By the lemma,  $x_{N+1}$  can be joined to either element of the dichotomy if and only if there exists a separating vector  $w$  for  $\{X^+, X^-\}$  lying in the  $(d-1)$ -dimensional orthogonal subspace to  $x_{N+1}$ . There are precisely  $C_{N,d-1}$  homogeneously linearly separable dichotomies of the projections of  $X$  in this subspace. Hence

$$C_{N+1,d} = C_{N,d} + C_{N,d-1} \quad (11)$$

Repeated application of Eq. (11) to the terms on the right yields

$$C_{N,d} = \sum_{k=0}^{N-1} \binom{N-1}{k} C_{1,d-k} \quad (12)$$

from which the theorem follows immediately upon noting

$$C_{1,m} = \begin{cases} 2, & m \geq 1 \\ 0, & m < 1 \end{cases} \quad (13)$$

### III SEPARABILITY BY ARBITRARY SURFACES

A change in point of view will enable us to apply the results of the previous section to classes of separating surfaces which are geometrically different from hyperplanes, but analytically quite similar. Suppose we are given a family of surfaces  $\{\Phi\}$ , each of which naturally divides a given space into two regions and a collection of  $N$  points in this space, each of which is assigned to one of two classes,  $X^+$  or  $X^-$ . This dichotomy of the points is said to be separable relative to  $\{\Phi\}$  if there exists at least one surface  $\Phi$  such that all the  $X^+$  points are in one region and all the  $X^-$  points are in the other. The crucial property of the family of surfaces  $\{\Phi\}$ , in order that the results of the previous section apply, is that  $\{\Phi\}$  can be parameterized in such a way that  $\{\Phi\}$  is linear in its parameters. Hyperplanes, hyperspheres, and polynomial surfaces are special examples of such families.

Consider the set of  $N$  objects  $X = \{x_1, \dots, x_N\}$ . We shall refer to the elements of  $X$  as patterns for intuitive reasons. These patterns need not be considered as vectors in a vector space. On each pattern  $x \in X$  a set of real valued measurement functions  $\varphi_1, \varphi_2, \dots, \varphi_d$  comprises the vector of measurements

$$\varphi : X \rightarrow R^d \quad (14)$$

where  $\varphi(x) = [\varphi_1(x), \varphi_2(x), \dots, \varphi_d(x)]$ ,  $x \in X$ .

*Definition:* A dichotomy (binary partition)  $\{X^+, X^-\}$  of  $X$  is  $\varphi$ -separable if there exists a vector  $w$  such that

$$\begin{aligned} w \cdot \varphi(x) &> 0, & x \in X^+ \\ w \cdot \varphi(x) &< 0, & x \in X^- \end{aligned} \quad (15)$$

We shall count  $\{X^+, X^-\}$  and  $\{X^-, X^+\}$  as *distinct* dichotomies. We see that the separating surface in the measurement space is the hyperplane  $w \cdot \varphi = 0$ . The inverse image of this hyperplane is the surface  $\{x : w \cdot \varphi(x) = 0\}$  in the pattern space. The advantage in this general

formulation of the problem is that many interesting nonlinear surfaces in the pattern space can be mapped into hyperplanes in another space where the results of the previous section will apply.

*Definition:* Let the vector-valued measurement function  $\varphi$  be defined on the set of patterns

$$\varphi : X = \{x_1, \dots, x_N\} \rightarrow R^d \quad (16)$$

Then, a set of patterns  $X$  is in  $\varphi$ -general position if the following equivalent conditions hold:

(1) Every  $d$  element subset of the set of  $d$ -dimensional measurement vectors  $\{\varphi(x_1), \dots, \varphi(x_N)\}$  is linearly independent.

(1') Every  $d \times d$  submatrix of the  $N \times d$  matrix  $\Phi$

$$\Phi = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \dots & \varphi_d(x_1) \\ \varphi_1(x_2) & & & \\ \vdots & & & \\ \varphi_1(x_N) & \dots & & \varphi_d(x_N) \end{bmatrix} \quad (17)$$

has a non-zero determinant.

(1'') No  $d + 1$  patterns lie on the same  $\varphi$ -surface  $\{x : \varphi(x) \cdot w = 0\}$  in the pattern space.

Clearly Definition 1' is just an explicit algebraic statement of Definition 1. Note that general position is a strengthened rank condition on the matrix  $\Phi$  ( $\Phi$  has maximal rank  $d$  if at least one  $d \times d$  submatrix has nonzero determinant). Definition 1'' relates general position in the measurement space to general position in the pattern space.

*Theorem 4:* Let  $X = \{x_1, x_2, \dots, x_N\}$  be in  $\varphi$ -general position where  $\varphi(x) = [\varphi_1(x), \varphi_2(x), \dots, \varphi_d(x)]$ , then precisely  $C_{N,d}$  of the  $2^N$  dichotomies of  $X$  are  $\varphi$ -separable where

$$C_{N,d} = 2 \sum_{i=0}^{d-1} \binom{N-1}{i} . \quad (18)$$

If, in addition, the  $\varphi$ -surface  $\{x : \varphi(x) \cdot w = 0\}$  is constrained to contain the set of points  $Y = \{y_1, y_2, \dots, y_k\}$ , where the projection of  $\varphi(X)$  into the orthogonal subspace to  $\varphi(Y)$  is in general position, and  $\varphi(Y)$  is linearly independent, then there are  $C_{N,d-k}$   $\varphi$ -separable dichotomies of  $X$ .

*Proof:* (using Theorem 3")

Every  $d$ -element subset of the  $N$  vectors  $\varphi(x_1), \dots, \varphi(x_N)$  is linearly independent by hypothesis. Hence, by Theorem 3" there are precisely  $C_{N,d}$  homogeneously linearly separable dichotomies of  $\{\varphi(x_i), i = 1, 2, \dots, N\}$ . By definition these dichotomies correspond to the  $\varphi$ -separable dichotomies of  $X$ .

The condition that the  $\varphi$ -surface contains the set  $Y$  is that the weight vector  $w$  must lie in the  $(d - k)$ -dimensional subspace  $S$  where

$$S = \{w : w \cdot \varphi(y_i) = 0, \quad i = 1, 2, \dots, k\} .$$

Let  $\hat{\varphi}$  be the projection of  $\varphi$  onto  $S$ . Then, since

$$w \cdot \varphi = w \cdot \hat{\varphi} + w \cdot (\varphi - \hat{\varphi}) = w \cdot \hat{\varphi} \quad (19)$$

for all  $w$  in  $S$ , we see that a set of vectors  $\{\varphi\}$  is separable by a weight vector in  $S$  if and only if the set of their projections  $\{\hat{\varphi}\}$  is separable. Since the vectors  $\hat{\varphi}(x_1), \dots, \hat{\varphi}(x_N)$  are in  $\hat{\varphi}$ -general position in  $S$ , there are  $C_{N,d-k}$  homogeneously linearly separable dichotomies of  $\{\varphi(x_i); i = 1, 2, \dots, N\}$  by a vector  $w$  in  $S$ .

A natural generalization of linear separability is polynomial separability. For the ensuing discussion, consider the patterns to be vectors in an  $m$ -dimensional space. The measurement function  $\varphi$  then maps points in  $m$ -space into points in  $d$ -space.

Consider a natural class of mappings obtained by adjoining  $r$ -wise products of the pattern vector coordinates. The natural separating surfaces corresponding to such mappings are known as  $r$ th order rational varieties. A rational variety of order  $r$  obtained in a space of  $m$  dimensions is represented by a homogeneous equation in the coordinates  $(x)_i$  of the  $r$ th degree.

$$0 = \sum_{0 \leq i_1 \leq i_2 \leq \dots \leq i_r \leq m} a_{i_1 i_2 \dots i_r} (x)_{i_1} (x)_{i_2} \dots (x)_{i_r} \quad (20)$$

where  $(x)_i$  is the  $i$ th component of  $x$  in  $R^m$  and  $(x)_0 = 1$  in order to write the expression in homogeneous form. A simple counting argument gives the number of coefficients  $F_m^{(r)}$  in Eq. (20) as

$$F_m^{(r)} = \sum_{k=0}^r \binom{m+k-1}{k} \quad (21)$$

We remark that, since the surface represented by the coefficients is independent of a change of scale in the coefficients, there are really only  $F_m^{(r)} - 1$  independent coefficients in Eq. (20). The quantity  $F_m^{(r)} - 1$  is known in classical geometry as the number of degrees of freedom of the surface.

First order rational varieties are hyperplanes and second-order rational varieties are quadrics. Hyperspheres are quadrics with certain linear constraints on the coefficients.

In Theorem 4, the mapping  $\varphi: R^m \rightarrow R^{F_m^{(r)}}$  defined by

$$\varphi(x) = [1, (x)_1, \dots, (x)_m, (x)_1^2, \dots, (x)_i(x)_j, \dots, (x)_m^r] \quad (22)$$

yields the following result: A set of  $N$  points in  $m$ -space, such that no  $F_m^{(r)}$  points lie on the same  $r$ th-order rational variety, has precisely  $C_{N, F_m^{(r)}}$  dichotomies which are separable by an  $r$ th-order rational variety. If the variety is constrained to contain  $k$  points, the number of separable dichotomies is reduced to  $C_{N, F_m^{(r)}-k}$ .

Koford<sup>9</sup> has observed that augmenting the vector  $x \in R^d$  to yield a vector  $\varphi(x)$  as in Eq. (22) is especially easy to implement when the coefficients are binary. Bishop<sup>10</sup> has exhaustively found the number  $\varphi_m$  of quadrically separable truth functions of  $m$  arguments for low  $m$ .



From the foregoing it can be seen that  $\varphi_n$  is bounded above by

$$\varphi_n \leq C_{2^{n+1}, n+1} \binom{n+1}{2} \sim 2^{n^3/2 + O(n^2 \log n)} \quad (23)$$

In addition, Koford<sup>9</sup> notes that if the augmented vector  $\varphi(x)$  is used as an input to a linear threshold device (as in Fig. 2), then the standard training procedure will converge,<sup>11</sup> (by the Perceptron convergence theorem) in a finite number of steps to a separating  $\varphi$ -surface if one exists.

Table 1 lists several examples of families of separating surfaces. All patterns  $x$  should be considered as vectors in an  $m$ -dimensional space. The function  $\varphi(x) = (1, x)$  is a  $(m + 1)$ -dimensional vector.

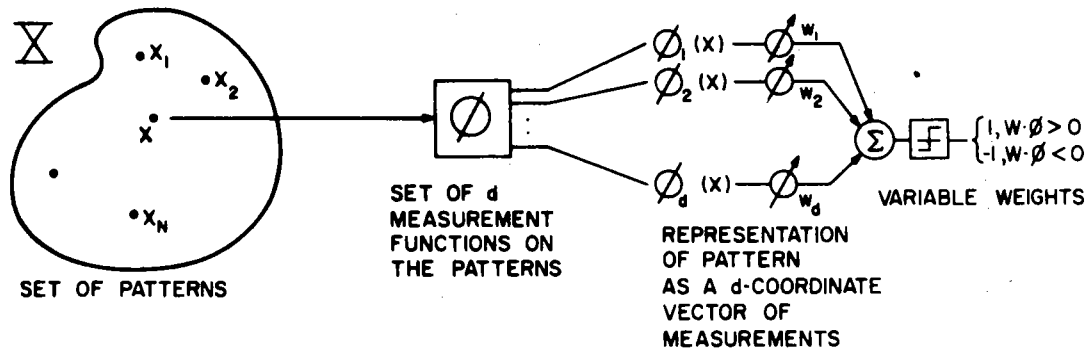


FIG. 2 MEASUREMENT TRANSFORMATION AND IMPLEMENTATION OF SEPARATING  $\phi$ -SURFACE

Table I

EXAMPLES OF SEPARATING SURFACES WITH THE CORRESPONDING NUMBER  
OF SEPARABLE DICHOTOMIES OF  $N$  POINTS IN  $m$  DIMENSIONS

MAPPING $\varphi$ DEFINED ON $R^m$	SEPARATING SURFACE IN PATTERN SPACE	NUMBER OF PARAMETERS OF $\varphi$ -SURFACE	GEOMETRICAL MEANING OF $\varphi$ -GENERAL POSITION	NUMBER OF $\varphi$ -SEPARABLE DICHOTOMIES OF $N$ POINTS	SEPARATING CAPACITY OF $\varphi$ -SURFACE
$\varphi(x) = x$	Hyperplane through origin	$m$	Every $m$ points linearly independent	$C_{N, m}$	$2m$
$\varphi(x) = (1, x)$	Hyperplane	$m + 1$	No $m + 1$ points on any hyperplane	$C_{N, m+1}$	$2m + 2$
$\varphi(x) = (1, x, \ x\ ^2)$	Hypersphere	$m + 2$	No $m + 2$ points on any hypersphere	$C_{N, m+2}$	$2m + 4$
$\varphi(x) = (x, \ x\ )$	Hypercone with vertex at origin	$m + 1$	No $m + 1$ points on hypercone	$C_{N, m+1}$	$2m + 2$
$\varphi(x)$ as in Eq. (22)	Rational $r$ th- order variety	$F_m^{(r)} = \sum_{k=0}^r \binom{m+k-1}{k}$	No $F_m^{(r)}$ points on same $r$ th-order rational variety	$C_{N, F_m^{(r)}}$	$2F_m^{(r)}$

#### IV RANDOM PATTERNS, CAPACITY

Suppose that the patterns  $X = \{x_1, x_2, \dots, x_N\}$  are chosen independently according to a probability measure  $\mu$  on the pattern space. How many dichotomies of  $X$  are  $\varphi$ -separable? Clearly, the results of the previous section will apply if  $X$  is in  $\varphi$ -general position. Now  $X$  is in  $\varphi$ -general position with probability one if and only if every  $\varphi$ -surface  $\{x : w \cdot \varphi(x) = 0\}$  has  $\mu$  measure zero.

Suppose that a dichotomy of  $X$  is chosen at random from the  $2^N$  equiprobable possible dichotomies of  $X$ . What is the probability  $P_{N,d}$  that this dichotomy is  $\varphi$ -separable? If  $X$  is in  $\varphi$ -general position with probability one, then with probability one there are  $C_{N,d}$   $\varphi$ -separable dichotomies. Thus

$$P_{N,d} = \left(\frac{1}{2}\right)^N C_{N,d} = \left(\frac{1}{2}\right)^{N-1} \sum_{i=0}^{d-1} \binom{N-1}{i} \quad (24)$$

Note that Theorem 1 now follows from the above with the identification  $\varphi(x) = x$ . The necessary and sufficient conditions on the measure  $\rho$  defined on the unit  $d$ -sphere, for Theorem 1 to hold are the following:

- (1) The  $\rho$  measure of any subspace is zero.
- (2)  $\rho$  is radially symmetric, i.e.,

$$\rho(B) = \rho(v : -v \in B) \quad (25)$$

Condition (2) is necessary in order that the probability (conditioned on  $X$ ) that a random dichotomy of  $X$  be separable is equal to the unconditional probability that a particular dichotomy of  $X$  (all  $N$  points in one hemisphere) be separable.

Let  $\{x_1, x_2, \dots\}$  be a sequence of random patterns as above and define the random variable  $N$  to be the largest integer such that  $\{x_1, x_2, \dots, x_N\}$  is  $\varphi$ -separable. From Eq. (24) we determine the probability density of  $N$

$$P_d(N = n) = \left(\frac{1}{2}\right)^n \binom{n-1}{d-1} \quad (26)$$

which is just the negative binomial distribution (shifted  $d$  units right) with parameters  $d$  and  $1/2$ ). Thus  $N$  corresponds to the waiting time for the  $d$ th failure in a series of tosses with a fair coin, and

$$E(N) = 2d \quad (27)$$

$$\text{Median}(N) = 2d \quad (28)$$

The asymptotic probability that  $N$  patterns are separable in  $d \doteq N/2 + (\alpha/2)\sqrt{N}$  dimensions is

$$P_{N, \frac{N}{2} + \frac{\alpha}{2}\sqrt{N}} \sim \Phi(\alpha)$$

where  $\Phi(\alpha)$  is the cumulative normal distribution

$$\Phi(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{x^2}{2}} dx$$

In addition, for  $\epsilon > 0$ ,

$$\lim_{d \rightarrow \infty} P_{2d(1+\epsilon), d} = 0$$

$$P_{2d, d} = \frac{1}{2}$$

$$\lim_{d \rightarrow \infty} P_{2d(1-\epsilon), d} = 1 \quad (29)$$

These results confirm the conjecture by Koford<sup>1</sup> that  $E[N] = 2d$ , and suggest that  $2d$  is a natural definition of the *separating capacity* of a family of surfaces that is linear in  $d$  parameters. Thus a linear threshold device can be said to have a separating capacity of 2 patterns per variable weight.

## V GENERALIZATION

It is generally believed that after a "large number" of training patterns the state of a linear threshold device is sufficiently constrained to yield an unambiguous response to a new pattern. We shall show in this section that this intuition is misleading in the case of random assignments. Physical considerations such as bunching according to category, and effects of training algorithms which tend to locate the separating hyperplane in the middle of the region between the two categories are not considered in this report.

Consider a set of  $N$  patterns  $X = \{x_1, x_2, \dots, x_N\}$  in  $d$ -space, which we shall call the training set. Let  $\{X^+, X^-\}$  be a linearly separable dichotomy of  $X$ . Note that we are speaking of linearly separable rather than homogeneously linearly separable dichotomies in this section. A new vector  $y$  is given. On what basis may  $y$  be assigned unambiguously to  $X^+$  or  $X^-$ ?

The classification of a vector  $y$  with respect to the partition  $\{X^+, X^-\}$  is said to be *ambiguous* if, among the class of hyperplanes separating  $X^+$  and  $X^-$ , there exists a hyperplane inducing the dichotomy  $\{X^+ \cup \{y\}, X^-\}$  and another hyperplane inducing the dichotomy  $\{X^+, X^- \cup \{y\}\}$ . We make the reasonable definition that  $y$  is ambiguous with respect to  $\{X^+, X^-\}$  if  $\{X^+, X^-\}$  is not linearly separable. In Fig. 3, for example,

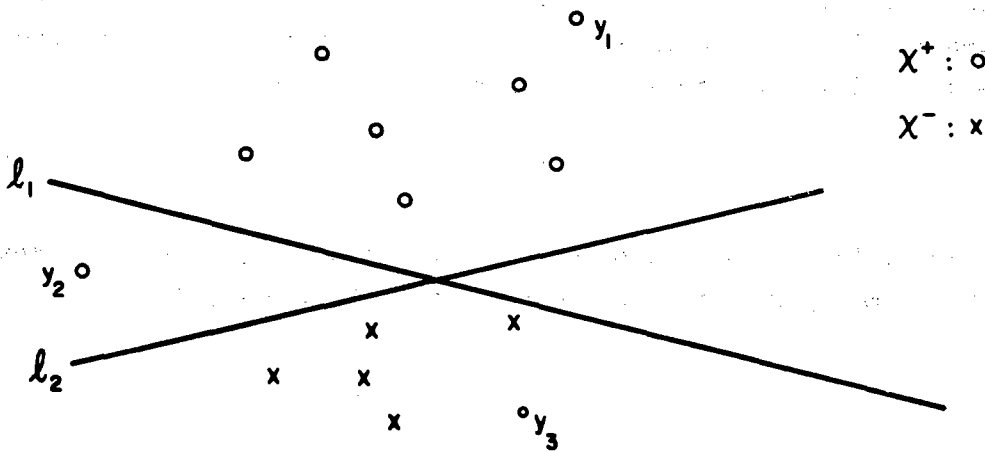


FIG. 3 GENERALIZATION

points  $y_1$  and  $y_3$  are unambiguously classifiable into sets  $X^+$  and  $X^-$  respectively, while point  $y_2$  has an ambiguous classification because lines  $\mathcal{L}_1$  and  $\mathcal{L}_2$  separate  $\{X^+, X^-\}$  but yield opposite classifications for  $y_2$ .

Note that the results of previous sections apply to generalization with respect to more general classes of surfaces if a proper mapping  $\varphi : X \rightarrow R^d$  is made.

*Proposition 1.* Let  $XU\{y\} = \{x_1, x_2, \dots, x_N, y\}$  be in general position in  $d$ -space. Let each of the linearly separable dichotomies of  $X$  have equal probability. Then the probability  $F_{N,d}$  that  $y$  is ambiguous with respect to a dichotomy of  $X$  is

$$F_{N,d} = \frac{C_{N,d}}{C_{N,d+1}} = \frac{\sum_{i=0}^{d-1} \binom{N-1}{i}}{\sum_{i=0}^d \binom{N-1}{i}} \quad (30)$$

Proposition 1 will follow from Proposition 2.

*Proposition 2.* Let  $XU\{y\} = \{x_1, x_2, \dots, x_N, y\}$  be in general position in  $d$ -space. Then  $y$  has an ambiguous classification with respect to  $C_{N,d}$  dichotomies of  $X$ .

*Proof:*

From the lemma of Section 2, the point  $y$  is ambiguous with respect to  $\{X^+, X^-\}$  if and only if there exists a hyperplane containing  $y$  which separates  $\{X^+, X^-\}$ . The proposition follows immediately from Theorem 4, where  $\varphi(x) = (1, x)$  and the separating vector  $w$  is constrained by

$$w \cdot \varphi(y) = 0 \quad (31)$$

Proposition 1 follows upon noting that there are  $C_{N,d+1}$  linearly separable dichotomies, by application of Theorem 4 with  $\varphi(x) = (1, x)$ .

Consider the behavior of  $F_{N,d}$

for  $N \gg d$ ,

$$F_{N,d} = \frac{\sum_{i=0}^{d-1} \binom{N-1}{i}}{\sum_{i=0}^d \binom{N-1}{i}} = \frac{\binom{N-1}{d-1}}{\binom{N-1}{d}} = \frac{d}{N-d} = \frac{d}{N} \quad (32)$$

for  $N = 2d$ ,

$$\lim_{d \rightarrow \infty} F_{2d,d} = 1 \quad (33)$$

and for  $N \leq d$ ,

$$F_{N,d} = \frac{2^N}{2^N} = 1 \quad (34)$$

For large dimension  $d$ , a plot of the probability  $F_{N,d}$  that a new pattern will be classified ambiguously with respect to a random dichotomy of the training set has the form shown in Fig. 4.

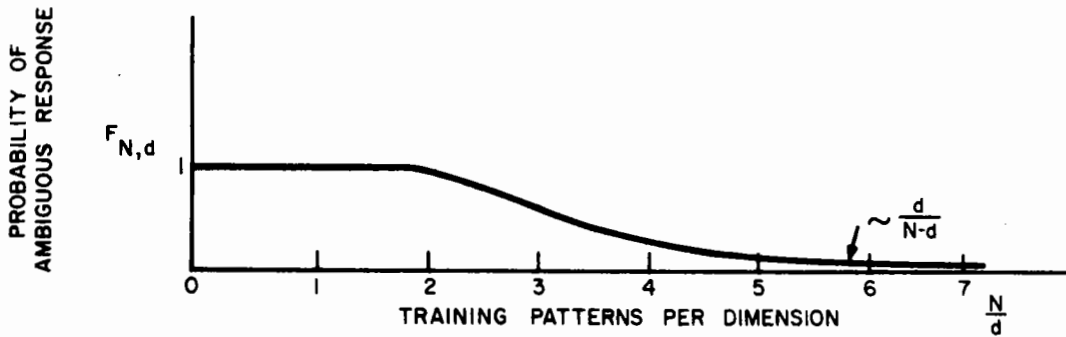


FIG. 4 PROBABILITY OF AMBIGUITY IN GENERALIZATION

Note the relatively large number of training patterns required for unambiguous generalization. Compare Fig. 4 to Fig. 5, where we see that the probability of ambiguous response remains high even after the probability of a consistent training set tends toward zero.

PROBABILITY THAT  $N$  RANDOM  
PATTERNS ARE SEPARABLE IN  $d$ -SPACE

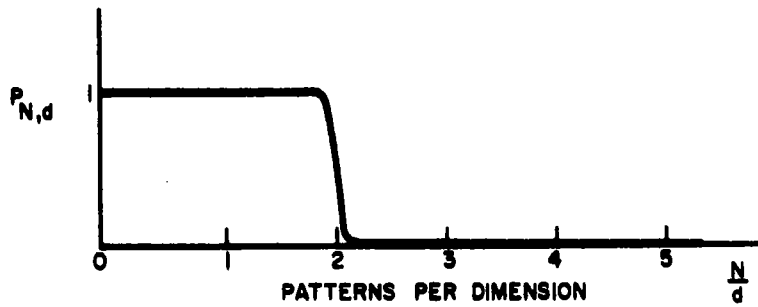


FIG. 5 TRAINING CAPACITY

In the event that the patterns themselves are randomly distributed, we remark that the comments made in Section IV concerning randomly distributed patterns and random dichotomies of the pattern set apply in full to this section. The crucial condition is that the pattern set be in general position with probability one.



## REFERENCES

1. B. Widrow, "Generalization and Information Storage in Networks of Adaline 'Neurons,'" *Self Organizing Systems*, p. 442 (Spartan Books, New York City, 1962).
2. R. Brown, "Logical Properties of Adaptive Networks," Stanford Electronics Laboratory Quarterly Research Review No. 4, III-6 to III-9 (1963).
3. J. G. Wendel, "A Problem in Geometric Probability," *Mathematica Scandinavica* **11**, pp. 109-111 (1962).
4. R. O. Winder, "Threshold Logic," Ph.D. Dissertation, Princeton University, Princeton, New Jersey (1962).
5. S. H. Cameron, "Proceedings of the Bionics Symposium 197-212," Wright Air Development Division, Tech. Report 60-600 (1960).
6. E. A. Whitmore and D. G. Willis, "Division of Space by Concurrent Hyperplanes," Unpublished internal report, Lockheed Missiles and Space Division, Sunnyvale.
7. R. D. Joseph, "The Number of Orthants in  $n$ -Space Intersected by a  $s$ -Dimensional Subspace," Tech. Memorandum 8, Project PARA, Cornell Aeronautical Laboratory, Buffalo, New York (1960).
8. L. Schläfli, *Gesammelte Mathematische Abhandlungen I* (Verlag Birkhauser, Basel, 1950).
9. J. Koford, "Adaptive Network Organization," Stanford Electronics Laboratory Quarterly Research Review 3, P. III-6 (1962).
10. A. B. Bishop, "Adaptive Pattern Recognition," 1963 WESCON, Session 1.5.
11. A. Novikoff, "On Convergence Proofs for Perceptrons," Symposium on Mathematical Theory of Automata, Polytechnic Institute of Brooklyn (April 1963, proceedings in press).

