Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition

THOMAS M. COVER

Abstract—This paper develops the separating capacities of families of nonlinear decision surfaces by a direct application of a theorem in classical combinatorial geometry. It is shown that a family of surfaces having d degrees of freedom has a natural separating capacity of $2d$ pattern vectors, thus extending and unifying results of Winder and others on the pattern-separating capacity of hyperplanes. Applying these ideas to the vertices of a binary n-cube yields bounds on the number of spherically, quadratically, and, in general, nonlinearly separable Boolean functions of n variables.

It is shown that the set of all surfaces which separate a dichotomy of an infinite, random, separable set of pattern vectors can be characterized, on the average, by a subset of only $2d$ extreme pattern vectors. In addition, the problem of generalizing the classifications on a labeled set of pattern points to the classification of a new point is defined, and it is found that the probability of ambiguous generalization is large unless the number of training patterns exceeds the capacity of the set of separating surfaces.

I. Definitions and History of Function-Counting Theorems

CONSIDER a set of patterns represented by a set of vectors in a d-dimensional Euclidean space E^d. A *homogeneous linear threshold function* $fw: E^d \rightarrow \{-1, 0, 1\}$ is defined in terms of a parameter or *weight vector* w for every vector x in this space:

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

where $w \cdot x$ is understood to mean the inner product of w and x.

Thus every homogeneous linear threshold function naturally divides E^d into two sets, the set of vectors x such that $fw(x) = 1$ and the set of vectors x such that $fw(x) = -1$. These two sets are separated by the hyperplane

$$\{x: f_w(x) = 0\} = \{x: x \cdot w = 0\} \tag{2}$$

which is the $(d-1)$-dimensional subspace orthogonal to the weight vector w. Let X be an arbitrary set of vectors in E^d. A dichotomy $\{X^+, X^-\}$ of X is *linearly separable* if and only if there exists a weight vector w in E^d and a scalar t such that

$$\begin{aligned} x \cdot w &> t, \quad \text{if } x \in X^+ \\ x \cdot w &< t, \quad \text{if } x \in X^-. \end{aligned} \tag{3}$$

The dichotomy $\{X^+, X^-\}$ is said to be *homogeneously linearly separable* if it is linearly separable with $t = 0$. A vector w satisfying

$$\begin{aligned} w \cdot x &> 0, \quad x \in X^+ \\ w \cdot x &< 0, \quad x \in X^- \end{aligned} \tag{4}$$

will be called a *solution vector*, and the corresponding orthogonal hyperplane $\{x: x \cdot w = 0\}$ will be called a *separating hyperplane* for the dichotomy $\{X^+, X^-\}$. In this, the homogeneous case, the separating hyperplane passes through the origin of the space and is, in fact, the $(d-1)$-dimensional orthogonal subspace to w. Finally, a set of N vectors is in *general position* in d-space if every subset of d or fewer vectors is linearly independent.

The foundations have been laid for the presentation of the fundamental function-counting theorem which counts the number of homogeneously linearly separable dichotomies of N points in d dimensions.

Theorem 1 (Function-Counting Theorem): There are $C(N, d)$ homogeneously linearly separable dichotomies of N points in general position in Euclidean d-space, where

$$C(N, d) = 2 \sum_{k=0}^{d-1} \binom{N-1}{k}. \tag{5}$$

The binomial coefficients comprising (N, d) defined for all real s and integer k by

$$\binom{s}{k} = \frac{s(s-1) \cdots (s-k+1)}{k!}. \tag{6}$$

This interesting theorem has been independently proved in different forms by many authors [1]–[6], but Winder [1], [2], Cameron [3], Joseph [4], and Whitmore and Willis [5] have emphasized the application of Theorem 1 to counting the number of linearly separable dichotomies of a set. In addition, Winder and

Manuscript received November 11, 1964; revised March 1, 1965. This work was performed at the Stanford Research Institute, Menlo Park, Calif., and at Stanford University, Stanford, Calif., and was partially supported by an ITT Federal Laboratories grant and an NSF fellowship. The material contained in this paper is a condensation of portions of the author's Ph.D. dissertation [23].

The author is with the Department of Electrical Engineering, Stanford University, Stanford, Calif.

Cameron independently applied Theorem 1 to the vertices of a binary n-cube in order to find an upper bound on the number of linearly separable truth functions of n variables. These authors [1]–[5] have all used a variant of a proof, which seems to have first appeared in Schläfli [6], of Theorem 2 or its dual statement Theorem 2′.

Theorem 2: N hyperplanes in general position passing through the origin of d-space divide the space into $C(N, d)$ regions.

Theorem 2′: A d-dimensional subspace in general position in N-space intersects $C(N, d)$ orthants.

Proofs of Theorem 2 appear in Schläfli [6], Winder [1], [2], Cameron [3], and Wendel [7]. Proofs in terms of the dual statement, Theorem 2′, can be found in Schläfli [6] and in Joseph [4]. It should be noted that most of these references are relatively obscure. A variation of the known proofs of Theorems 2 and 2′ will therefore be provided in the first portion of the proof of Theorem 3.

II. Separability by Arbitrary Surfaces

Many authors [8]–[12] have been concerned with separating sets of points with parametric families of surfaces. Cooper [8], [9] has been primarily concerned with the characterization of the natural class of decision surfaces for a given decision theoretic pattern-recognition problem, as well as with the complementary question of characterizing the natural class of problems for which a given family of decision surfaces is minimal complete. Bishop [13], Wong and Eisenberg [14], and Cooper [15] have been concerned with the problem of using rth-order polynomial surfaces to implement truth functions (separating a dichotomy of the vertices of a binary n-cube). Koford [16] and Aizerman et al. [10] have observed that standard training algorithms will converge to a separating surface if one exists. In this section, the number of dichotomies of a set of points which may be separated by a family of decision surfaces will be found. This number follows directly from the function-counting theorem when the family of separating surfaces and the set of points to be separated are carefully defined.

Consider a family of surfaces, 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 the family of surfaces if there exists a surface in the family that separates the points in X^+ from the points in X^-. Consider the set of N objects $X=\{x_1, \cdots, x_N\}$. The elements of X will be referred to as patterns for intuitive reasons. On each pattern $x\in X$, a set of real-valued *measurement functions* $\phi_1, \phi_2, \cdots, \phi_d$ comprises the vector of measurements

$$\phi: \quad X \rightarrow E^d \tag{7}$$

where $\phi(x)=[\phi_1(x), \phi_2(x), \cdots, \phi_d(x)]$, $x\in X$.

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

$$\begin{aligned} w\cdot\phi(x) &> 0, \qquad x \in X^+ \\ w\cdot\phi(x) &< 0, \qquad x \in X^-. \end{aligned} \tag{8}$$

Observe that the separating surface in the measurement space is the hyperplane $w\cdot\phi=0$. The inverse image of this hyperplane is the separating surface $\{x: w\cdot\phi(x)=0\}$ in the pattern space.

Definition: Let the vector-valued measurement function ϕ map a set of patterns $X=\{x_1, x_2, \cdots, x_N\}$ into E^d. The set X is said to be in *ϕ-general position* if Condition 1 holds.

Condition 1: Every k element subset of the set of d-dimensional measurement vectors $\{\phi(x_1), \phi(x_2), \cdots, \phi(x_N)\}$ is linearly independent for all $k\leq d$.

When $N\geq d$, Conditions 1, 1′, 1″, and 1‴ are equivalent.

Condition 1′: Every d element subset of the set of d-dimensional measurement vectors $\{\phi(x_1), \cdots, \phi(x_N)\}$ is linearly independent.

Condition 1″: Every $d\times d$ submatrix of the $N\times d$ matrix

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_d(x_1) \\ \phi_1(x_2) & & & \\ \vdots & & & \\ \phi_1(x_N) & & \cdots & \phi_d(x_N) \end{bmatrix} \tag{9}$$

has a nonzero determinant.

Condition 1‴: No $d+1$ patterns lie on any one ϕ-surface in the pattern space.

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

A lemma will now be established which plays a central role in the investigation in that it enables the extension of Theorem 1 to the case in which the family of decision surfaces is constrained to pass through a given set of points. This lemma also provides an alternative simple proof of the function-counting theorem and will be used in Section VI, Generalization and Learning.

Lemma 1: Let X^+ and X^- be subsets of E^d, and let y be a point other than the origin in E^d. Then the dichotomies $\{X^+\cup\{y\}, X^-\}$ and $\{X^+, X^-\cup\{y\}\}$ are both homogeneously linearly separable if and only if $\{X^+, X^-\}$ is homogeneously linearly separable by a $(d-1)$-dimensional subspace containing y.

Remarks: In geometrical terms, Lemma 1 says that a new point can be adjoined to either half of a separable dichotomy to form two new separable dichotomies if

and only if there exists a separating hyperplane through the new point which separates the old dichotomy. This is reasonable because, if such a hyperplane exists, small displacements of the hyperplane will allow arbitrary classification of the new point without affecting the separation of the old dichotomy. The proof makes these displacements explicit.

Proof: The set W of separating vectors for $\{X^+, X^-\}$ is given by $W=\{w: w\cdot x>0,\ x\in X^+;\ w\cdot x<0,\ x\in X^-\}$. The dichotomy $\{X^+\cup\{y\},\ X^-\}$ is homogeneously linearly separable if and only if there exists a w in W such that $w\cdot y>0$; and the dichotomy $\{X^+, X^-\cup\{y\}\}$ is homogeneously linearly separable if and only if there exists a w in W such that $w\cdot y<0$. If $\{X^+\cup\{y\}, X^-\}$ and $\{X^+, X^-\cup\{y\}\}$ are homogeneously linearly separable by w_1 and w_2, respectively, then $w^*=(-w_2\cdot y)w_1+(w_1\cdot y)w_2$ separates $\{X^+, X^-\}$ by the hyperplane $\{x: w^*\cdot x=0\}$ passing through y. Conversely, if $\{X^+, X^-\}$ is homogeneously linearly separable by a hyperplane containing y, then there exists a $w^*\in W$ such that $w^*\cdot y=0$. Since W is open, there exists an $\epsilon>0$ such that $w^*+\epsilon y$ and $w^*-\epsilon y$ are in W. Hence, $\{X^+\cup\{y\}, X^-\}$ and $\{X^+, X^-\cup\{y\}\}$ are homogeneously linearly separable by $w^*+\epsilon y$ and $w^*-\epsilon y$, respectively.

Theorem 3 generalizes the function-counting theorem to certain classes of nonlinear functions under constraints. In particular, it states that k independent constraints on the class of separating surfaces reduce the number of degrees of freedom of the class by k.

Theorem 3: If a ϕ-surface $\{x: w\cdot\phi(x)=0\}$ is constrained to contain the set of points $Y=\{y_1, y_2, \cdots, y_k\}$, where $\phi(y_1), \phi(y_2), \cdots, \phi(y_k)$ are linearly independent, and where the projection of $\phi(x_1), \phi(x_2), \cdots, \phi(x_N)$ onto the orthogonal subspace to the space spanned by $\phi(y_1), \phi(y_2), \cdots, \phi(y_k)$ is in general position, then there are $C(N, d-k)$ ϕ-separable dichotomies of X.

Proof: In the special case $k=0$ (no constraints) and $\phi(x)=x$ (linear separating surfaces), this theorem reduces to the statement of Theorem 1. We shall first prove this special case by induction on N and d. Let $C(N, d)$ be the number of homogeneously linearly separable dichotomies of $X=\{x_1, x_2, \cdots, x_N\}$. Consider a new point x_{N+1} such that $X\cup\{x_{N+1}\}$ is in general position, and consider the $C(N, d)$ homogeneously linearly separable dichotomies of X. If a dichotomy $\{X^+, X^-\}$ is separable, then either $\{X^+\cup\{x_{N+1}\}, X^-\}$ or $\{X^+, X^-\cup\{x_{N+1}\}\}$ must be separable. However, *both* dichotomies are separable, by Lemma 1, if and only if there exists a separating vector w for $\{X^+, X^-\}$ lying in the $(d-1)$-dimensional subspace orthogonal to x_{N+1}. A dichotomy of X is separable by such a w if and only if the projection of the set X onto the $(d-1)$-dimensional orthogonal subspace to x_{N+1} is separable. By the induction hypothesis there are $C(N, d-1)$ such separable dichotomies. Hence,

$$C(N+1, d) = C(N, d) + C(N, d-1). \tag{10}$$

Repeated application of (10) to the terms on the right yields

$$C(N, d) = \sum_{k=0}^{N-1}\binom{N-1}{k}C(1, d-k), \tag{11}$$

from which the theorem follows immediately on noting

$$C(1, m) = \begin{cases} 2, & m \geq 1 \\ 0, & m < 1. \end{cases} \tag{12}$$

Generalizing the proof to arbitrary ϕ and k, we first observe that the condition that a ϕ-surface contains the set $\{y_1, y_2, \cdots, y_k\}$ is that the parameter vector w which characterizes the surface must lie in the $(d-k)$-dimensional subspace L, where

$$L = \{w: w\cdot\phi(y_i) = 0,\ i = 1, 2, \cdots, k\}.$$

Let $\hat{\phi}$ be the orthogonal projection of ϕ onto L. Then, since $w\cdot\phi=w\cdot\hat{\phi}$ for all w in L, it is seen that a set of vectors $\{\phi\}$ is separable by a parameter vector in L if and only if the corresponding set of projections $\{\hat{\phi}\}$ is separable. Since, by the second assumption, $\hat{\phi}(x_1)$, $\hat{\phi}(x_2)$, $\cdots$, $\hat{\phi}(x_N)$ are in $\hat{\phi}$-general position in L, there are $C(N, d-k)$ homogeneously linearly separable dichotomies of $\{\phi(x_1), \phi(x_2), \cdots, \phi(x_N)\}$ by a vector w in L.

In defining a linear threshold function, it is difficult to decide whether to classify the points lying on the separating hyperplane into the $(+1)$ class or the (-1) class. This difficulty can be resolved by assigning these points to a third class and proving Theorem 4.

Theorem 4: Let $\{x_1, x_2, \cdots, x_N\}$ be a set of N vectors in general position in E^d. Let F be the class of functions $f: \{x_1, x_2, \cdots, x_N\}\rightarrow\{1, 0, -1\}$ defined for each w in E^d by

$$f_w(x_i) = \begin{cases} 1, & x_i\cdot w > 0 \\ 0, & x_i\cdot w = 0 \\ -1, & x_i\cdot w < 0, \quad i = 1, 2, \cdots, N. \end{cases} \tag{13}$$

Then there are $Q(N, d)$ functions in F, where

$$Q(N, d) = 2\sum_{k=0}^{d-1}\sum_{m=0}^{d-k-1}\binom{N}{k}\binom{N-k-1}{m}. \tag{14}$$

Proof: The number of functions f_w for which $f_w(x)=0$ on k given points of $\{x_1, \cdots, x_N\}$ corresponds to the number of different ways a homogeneous linear separating surface constrained to contain these points can dichotomize the remaining $N-k$ points. By Theorem 3 this number is just $C(N-k, d-k)$. Since there are $N!/k!(N-k)!$ ways to choose the k points for which $f_w=0$, the theorem follows upon summing on k.

Remarks: It can be verified that

$$Q(N, d) = 3^N, \qquad N \leq d \tag{15}$$

and

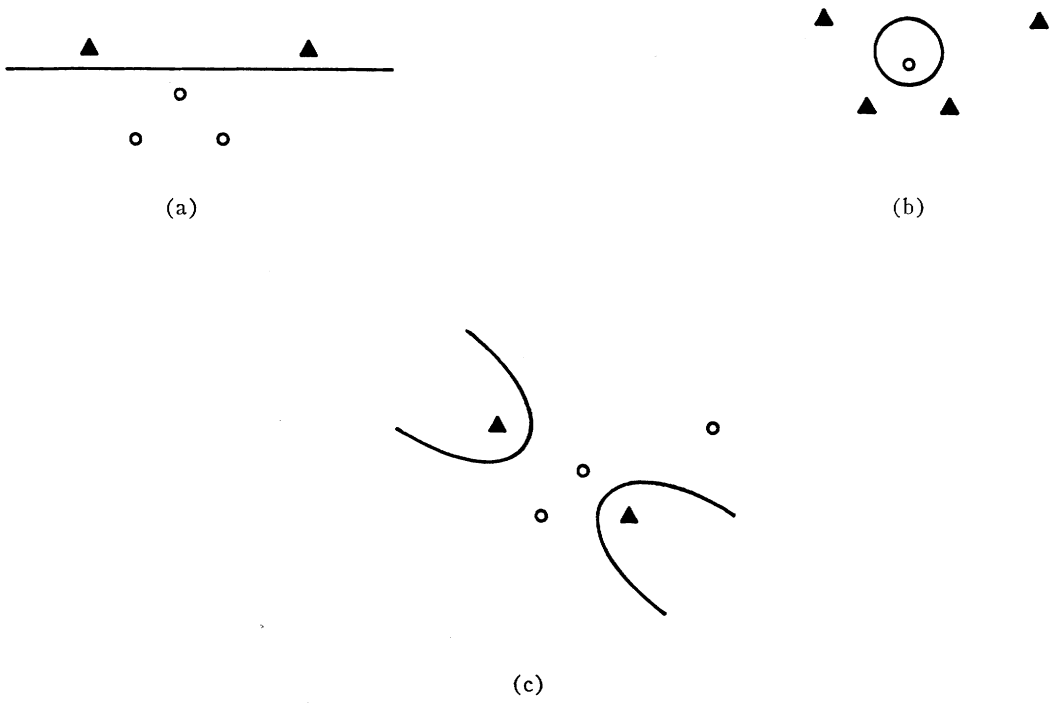

Fig. 1. Examples of ϕ-separable dichotomies of five points in two dimensions. (a) Linearly separable dichotomy. (b) Spherically separable dichotomy. (c) Quadrically separable dichotomy.

$$\lim_{N\to\infty} \frac{Q(N, d)}{C(N, d)} = 2^{d-1}.$$

III. Examples of Separating Surfaces

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 ϕ 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 rth-order rational varieties. A rational variety of order r in a space of m dimensions is represented by an rth-degree homogeneous equation in the coordinates $(x)_i$,

$$\sum_{0\le i_1\le i_2\le\cdots\le i_r\le m} a_{i_1 i_2\cdots i_r}(x)_{i_1}(x)_{i_2}\cdots(x)_{i_r} = 0, \quad (16)$$

where $(x)_i$ is the ith component of x in E^m and $(x)_0$ is set equal to 1 in order to write the expression in homogeneous form. Note that there are $(m-r)!/m!r!$ coefficients in (16). Examples of surfaces of this form are hyperplanes (first-order rational varieties), quadrics (second-order rational varieties), and hyperspheres (quadrics with certain linear constraints on the coefficients). Figure 1 illustrates three dichotomies of the same configuration of points. Of the 32 dichotomies of the five points in Fig. 1, precisely $C(5, 3) = 22$ are linearly separable, $C(5, 4) = 30$ are spherically separable, and $C(5, 5) = 32$ are quadrically separable. It is clearly true in general that linear separability implies spherical separability, which in turn implies quadric separability.

In order to establish the number of separable dichotomies of pattern vectors by general surfaces of these types, inspection of (16) and application of Theorem 3 with the mapping ϕ: $E^m \to E^{(m+r)!/m!r!}$ defined by

$$\phi(x) = (1, (x)_1, \cdots, (x)_m, (x_1)^2, \cdots, (x)_i(x)_j, \cdots, (x)_m{}^r) \quad (17)$$

yields the following result: A set of N points in m-space, such that no $(m+r)!m!r!$ points lie on the same rth-order rational variety, can be separated into precisely $C(N, (m+r)!m!r!)$ dichotomies by an rth-order rational variety. If the variety is constrained to contain k-independent points, the number of separable dichotomies is reduced by $C(N(m+r)!m!r!-k)$.

The original uses of the function-counting theorem were to establish upper bounds on the number of linearly separable truth functions of m variables. Since then, Bishop [13] has exhaustively found the number L_m of quadrically separable truth functions of m arguments for low m. From the foregoing it can be seen that L_m is less than or equal to the number of quadrically separable functions of 2^m points, with inequality instead of equality because the 2^m vertices of the binary m-cube are not in general position. Asymptotically we have the bound

$$L_m \le C\left(2^m, \binom{m+2}{2}\right) = 2^{m^3/2+0(m^2\log m)} \quad (18)$$

TABLE I

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

Mapping ϕ Defined on x in E^m	Separating Surface	Degrees of Freedom of ϕ-Surface	General Position	Number of ϕ-Separable Dichotomies of N Points	Separating Capacity
$\phi(x)=x$	hyperplane through origin	m	no m points on any subspace	$C(N, m)$	$2m$
$\phi(x)=(1, x)$	hyperplane	$m+1$	no $m+1$ points on any hyperplane	$C(N, m+1)$	$2(m+1)$
$\phi(x)=(1, x, \|\|x\|\|^2)$	hypersphere	$m+2$	no $m+2$ points on any hypersphere	$C(N, m+2)$	$2(m+2)$
$\phi(x)=(x, \|\|x\|\|)$	hypercone	$m+1$	no $m+1$ points on any hypercone	$C(N, m+1)$	$2(m+1)$
$\phi(x)=$all r-wise products of components of x	rational rth-order variety	$\binom{m+r}{r}$	no $\binom{m+r}{r}$ points on any rth-order surface	$C\left(N, \binom{m+r}{r}\right)$	$2\binom{m+r}{r}$

where $0(m^r \log m)$ is a remainder term which, for some K, is asymptotically bounded by $Km^r \log m$.

In general, for rth-order polynomial separating surfaces, the number $L_m(r)$ of separable truth functions of m variables is bounded above by

$$L_m(r) \leq C\left(2^m, \binom{m+r}{r}\right) = 2^{m^{r+1}/r!+0(m^r \log m)}. \quad (19)$$

Koford [16] has observed that augmenting the vector $x \in E^d$ to yield a vector $\phi(x)$, as in (17), is particularly easy to implement when the coefficients are binary. In addition, Koford notes—as do Aizerman [10] and Greenberg and Konheim [12]—that, if the augmented vector $\phi(x)$ is used as an input to a linear threshold device, then the standard fixed increment training procedure will converge [17] (by the Perceptron convergence theorem) in a finite number of steps to a separating ϕ-surface if one exists.

Table I lists several examples of families of separating surfaces. All patterns x should be considered as vectors in an m-dimensional space. The function $\phi(x)=(1, x)$ is an $(m+1)$-dimensional vector. The final column of Table I lists the separating capacities of the ϕ-surfaces, a measure of the expected number of random patterns which can be separated. The separating capacity will be made plausible as a useful idea in Section IV.

IV. Separability of Random Patterns

Two kinds of randomness are considered in the pattern dichotomization problem:

1) The patterns are fixed in position but are classified independently with equal probability into one of two categories.
2) The patterns themselves are randomly distributed in space, and the desired dichotomization may be random or fixed.

Under these conditions the separability of the set of pattern vectors becomes a random event depending on the dichotomy chosen and the configuration of the patterns. The probability of this random event and the maximum number of random patterns that can be separated by a given family of decision surfaces are to be determined.

Suppose that the patterns $x_1, x_2, \cdots, x_N$ are chosen independently according to a probability measure μ on the pattern space. The necessary and sufficient condition on u such that, with probability 1, $x_1, x_2, \cdots, x_N$ are in general position is d-space is that the probability be zero that any point will fall on any given $(d-1)$-dimensional subspace. In terms of ϕ-surfaces, a set of vectors chosen independently according to a probability measure μ is in ϕ-general position with probability 1 if and only if every ϕ-surface $\{x \in E^d : w \cdot \phi(x) = 0\}$ has μ measure zero.

Suppose that a dichotomy of $X = \{x_1, x_2, \cdots, x_N\}$ is chosen at random with equal probability from the 2^N equiprobable possible dichotomies of X. Let X be in ϕ-general position with probability 1, and let $P(N, d)$ be the probability that the random dichotomy is ϕ-separable, where the class of ϕ-surfaces has d degrees of freedom. Then with probability 1 there are $C(N, d)$ ϕ-separable dichotomies, and

$$P(N, d) = (\tfrac{1}{2})^N C(N, d) = (\tfrac{1}{2})^{N-1} \sum_{k=0}^{d-1} \binom{N-1}{k}, \quad (20)$$

which is just the cumulative binomial distribution corresponding to the probability that $N-1$ flips of a fair coin result in $d-1$ or fewer heads.

One of the first applications of the function-counting theorems to random pattern vectors was by Wendel [7], who found the probability that N random points lie in some hemisphere. Let the set of vectors $X = \{x_1, x_2, \cdots, x_N\}$ be in general position on the surface of a d-sphere with probability 1. In addition, let the joint distribution of $\{x_1, x_2, \cdots, x_N\}$ be unchanged by the reflection of any subset through the origin. Under these restrictions Wendel proves that the probability that a set of N vectors randomly distributed on the surface of a d-sphere is contained in some hemisphere is $P(N, d)$.

The proof of this result follows immediately from Schläfli's theorem and the reflection invariance of the

joint probability distribution of X. This invariance implies 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 category) be separable.

V. Separating Capacity of a Surface

It will be shown that the expected maximum number of randomly assigned vectors that are linearly separable in d dimensions is equal to $2d$. It is thus possible to conclude that a linear threshold device has an information storage *capacity*—relative to learning random dichotomies of a set of patterns—of two patterns per variable weight. This result was originally conjectured by Widrow and Koford and experimentally as reported by Widrow [18] for the case of pattern vectors chosen at random from the set of vertices of a binary d-cube. Brown [19] found experimentally that the conjecture held for patterns distributed at random in the unit d-sphere. This conjecture was supported theoretically by Winder [20], by Efron and Cover [21]–[23], and subsequently by Brown [24].

Let $\{x_1, x_2, \cdots\}$ be a sequence of random patterns as previously shown, and define the random variable N to be the largest integer such that $\{x_1, x_2, \cdots, x_N\}$ is ϕ-separable, where ϕ has d degrees of freedom. Then, from (20),

$$\begin{aligned} P_r\{N = n\} &= P(n, d) - P(n+1, d) \\ &= (\tfrac{1}{2})^n \binom{n-1}{d-1}, \quad n = 0, 1, 2, \cdots \end{aligned} \tag{21}$$

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

$$\begin{aligned} E(N) &= 2d \\ \text{Median } (N) &= 2d. \end{aligned} \tag{22}$$

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

$$P\left(N, \frac{N}{2} + \frac{\alpha}{2}\sqrt{N}\right) \sim \Phi(\alpha) \tag{23}$$

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

$$\Phi(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-x^2/2} dx. \tag{24}$$

In addition, for $\epsilon > 0$,

$$\begin{aligned} \lim_{d\to\infty} P(2d(1+\epsilon), d) &= 0 \\ P(2d, d) &= \tfrac{1}{2} \\ \lim_{d\to\infty} P(2d(1-\epsilon), d) &= 1 \end{aligned} \tag{25}$$

as was shown by Winder [20]. Thus the probability of separability shows a pronounced threshold effect when the number of patterns is equal to twice the number of dimensions. These results [21] confirm Koford's conjecture and suggest that $2d$ is a natural definition of the *separating capacity* of a family of decision surfaces having d degrees of freedom.

If the number of patterns is fixed at N and the dimensionality of the space in which the patterns lie is allowed to increase, it follows that the probability that the dimension d^* at which the set of patterns first becomes separable is given by

$$\begin{aligned} \Pr\{d^* = d\} &= P(N, d) - P(N, d-1) \\ &= (\tfrac{1}{2})^{N-1} \binom{N-1}{d-1}, \quad d = 1, 2, \cdots, N. \end{aligned} \tag{26}$$

Thus d^* is binomially distributed (shifted one unit right with parameters N and $\frac{1}{2}$), and

$$E(d^*) = \frac{N+1}{2}. \tag{27}$$

Equation (26) is partial justification for past statements that linear threshold devices are self-healing or can adjust around their defects [18] because the separating probability of a linear threshold device is relatively insensitive to the number of parameters d for $d > (N+1)/2$. The fact that $2d$ is indeed a critical number for a system of linear inequalities in d unknowns will be further established in Sections VI and VII.

VI. Generalization and Learning

Let X be a set of N pattern vectors in general position in d-space. This set of pattern vectors, together with a dichotomy of the set into two categories X^+ and X^-, will constitute a *training set*. On what basis can a new point be categorized into one of the two training categories? This is the problem of generalization.

Consider the problem of generalizing from the training set with respect to a given admissible family of decision surfaces (that family of surfaces that can be implemented by linear threshold devices). By some process, a decision surface from the admissible class will be selected which correctly separates the training set into the desired categories. Then the new pattern will be assigned to the category lying on the same side of the decision surface. Clearly, for some dichotomies of the set of training patterns, the assignment of category will not be unique. However, it is generally believed that, after a "large number" of training patterns, the state of a linear threshold device is sufficiently constrained to yield a unique response to a new pattern. It will be shown that the number of training patterns must exceed the statistical capacity of the linear threshold device before unique generalization becomes probable.

The classification of a pattern y with respect to the training set $\{X^+, X^-\}$ is said to be *ambiguous relative to a given class of ϕ-surfaces*, if there exists one ϕ-surface

in the class that induces the dichotomy $\{X^+ \cup \{y\}, X^-\}$ and another ϕ-surface in the class that induces the dichotomy $\{X^+, X^- \cup \{y\}\}$. That is, there exist two ϕ-surfaces, both correctly separating the training set, but yielding different classifications of the new pattern y. Thus, if w_1 and w_2 are the parameter weight vectors for the two ϕ-surfaces, then

$$w_1 \cdot \phi(x) > 0 \quad \text{and} \quad w_2 \cdot \phi(x) > 0 \qquad \text{for } x \in X^+$$
$$w_1 \cdot \phi(x) < 0 \quad \text{and} \quad w_2 \cdot \phi(x) < 0 \qquad \text{for } x \in X^-$$

and either

$$w_1 \cdot \phi(y) > 0 \quad \text{and} \quad w_2 \cdot \phi(y) < 0 \tag{28}$$

or

$$w_1 \cdot \phi(y) < 0 \quad \text{and} \quad w_2 \cdot \phi(y) > 0.$$

In addition, y is said to be ambiguous with respect to the training set if the training set is not separable.

In Fig. 2, for example, points y_1 and y_3 are unambiguous and point y_2 is ambiguous with respect to the training set $\{X^+, X^-\}$ relative to the class of all lines in the plane (not necessarily through the origin). Points y_1 and y_3 are uniquely classified into sets X^+ and X^-, respectively, by any line separating X^+ and X^-, while y_2 is classified into X^- by l_1, and into X^+ by l_2.

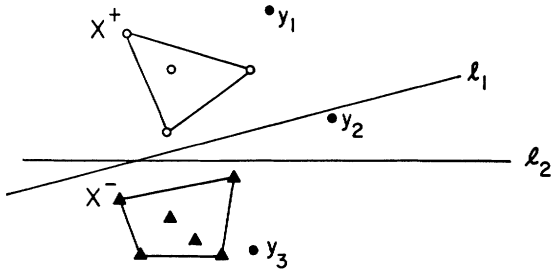

Fig. 2. Ambiguous generalization.

Theorem 6 establishes the probability that a new pattern is ambiguous with respect to a random dichotomy of the training set. This probability is independent of the configuration of the pattern vectors.

Theorem 6: Let $X \cup \{y\} = \{x_1, x_2, \cdots, x_N, y\}$ be in ϕ-general position in d-space, where $\phi = (\phi_1, \phi_2, \cdots, \phi_d)$. Then y is ambiguous with respect to $C(N, d-1)$ dichotomies of X relative to the class of all ϕ-surfaces. Hence, if each of the ϕ-separable dichotomies of X has equal probability, then the probability $A(N, d)$ that y is ambiguous with respect to a random ϕ-separable dichotomy of X is

$$A(N, d) = \frac{C(N, d-1)}{C(N, d)} = \frac{\sum_{k=0}^{d-2} \binom{N-1}{k}}{\sum_{k=0}^{d-1} \binom{N-1}{k}}. \tag{29}$$

Proof: From Lemma 1 of Section II, the point y is ambiguous with respect to $\{X^+, X^-\}$ if and only if there exists a ϕ-surface containing y which separates $\{X^+, X^-\}$. The proposition then follows from Theorem 3 on noting that the separating vector w obeys the linear constraint

$$w \cdot \phi(y) = 0. \tag{30}$$

Applying the proposition to the example in Fig. 2, where X has ten points, it can be seen that each of the y_i's is ambiguous with respect to $C(10, 2) = 20$ dichotomies of X relative to the class of all lines in the plane. Now $C(10, 3) = 92$ dichotomies of X are separable by the class of all lines in the plane. Thus, a new pattern is ambiguous with respect to a random, linearly separable dichotomy of X with probability

$$A(10, 3) = \frac{C(10, 2)}{C(10, 3)} = \frac{5}{23}. \tag{31}$$

The behavior of $A(N, d)$ is indicated by examination of its asymptotic form. Consider Badahur's expansion [25] of the cumulative binomial distribution, for $N \geq 2d$,

$$\sum_{i=0}^{d} \binom{N}{i} = \frac{1}{2}\binom{N}{d} F\left(N+1, 1; N-d+1; \frac{1}{2}\right) \tag{32}$$

where F is the hypergeometric function

$$F\left(N+1, 1; N-d+1; \frac{1}{2}\right) = 1 + \frac{N+1}{k+1}\frac{1}{2} + \frac{(N+1)(N+2)}{(k+1)(k+2)}\frac{1}{4} + \cdots \tag{33}$$

and

$$k = N - d. \tag{34}$$

Let $[x]$ denote the greatest integer less than x, and let $N = [\beta d]$, $\beta > 2$. Then the terms of the expansion of $F([\beta d]+1, 1; [\beta d]-d+1; \frac{1}{2})$ are uniformly bounded and positive, and the limit, as d increases, of the jth term is $(\beta/2(\beta-1))^j$. Thus, by the M-test of Weierstrass,

$$\lim_{\substack{N=[\beta d] \\ d \to \infty}} \frac{\sum_{i=0}^{d} \binom{N}{i}}{\binom{N}{d}} = \frac{1}{2} \frac{1}{1 - \beta/2(\beta-1)} = \frac{\beta-1}{\beta-2}, \qquad \beta > 2 \tag{35}$$

and

$$A^*(\beta) = \lim_{\substack{N=[\beta d] \\ d \to \infty}} A(N, d) = \begin{cases} 1, & 0 \leq \beta \leq 2 \\ \dfrac{1}{\beta - 1}, & \beta \geq 2 \end{cases}. \tag{36}$$

The graph of $A^*(\beta)$ is shown in Fig. 3. Note the relatively large number of training patterns required for unambiguous generalization. If it is recalled that the capacity of a linear threshold device is $\beta = 2$ patterns per variable weight, it will again be seen that the capac-

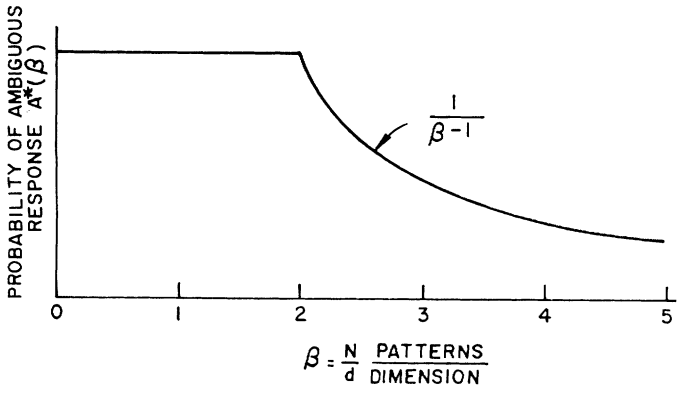

Fig. 3. Asymptotic probability of ambiguous generalization.

ity is a critical number in the description of the behavior of a linear threshold device.

If the patterns themselves are randomly distributed, the comments of Section IV concerning randomly distributed patterns and random dichotomies of the pattern set apply in full here. The crucial condition is that the pattern set be in general position with probability 1.

Thus, if a linear threshold device is trained on a set of N points chosen at random according to a uniform distribution on the surface of a unit sphere in d-space, and these points are classified independently with equal probability into one of two categories, then it is readily seen that the probability of error on a new pattern similarly chosen, conditioned on the separability of the entire set, is just $\frac{1}{2}A(N, d)$.

VII. Extreme Patterns

For any dichotomy of a set of points, there exists a minimal sufficient subset of extreme points such that any hyperplane correctly separating the subset must separate the entire set correctly. Thus for any dichotomy $\{X^+, X^-\}$ of a set of points in E^d, there exists a minimal set Z contained in $X^+ \cup X^-$ such that w satisfies

$$\begin{aligned} w \cdot x &> 0, \qquad x \in X^+ \\ w \cdot x &< 0, \qquad x \in X^- \end{aligned} \tag{37}$$

if, and only if, w satisfies the extremal constraints

$$\begin{aligned} w \cdot x &> 0, \qquad x \in X^+ \cap Z \\ w \cdot x &< 0, \qquad x \in X^- \cap Z. \end{aligned} \tag{38}$$

The set Z will be called the set of extreme points of the dichotomy. The vectors in Z form the boundary matrix investigated by Mays [26].

From this definition it follows that a point is an extreme point of the dichotomy $\{X^+, X^-\}$ if and only if it is ambiguous (in the sense of Section VI) with respect to $\{X^+, X^-\}$. Thus, for a set of N points in general position in E^d, each of the N points is ambiguous with respect to precisely $C(N-1, d-1)$ dichotomies of the remaining $N-1$ points. Hence, each of these $C(N-1, d-1)$ dichotomies is the restriction of two homogeneously linearly separable dichotomies of the original set of N points—the two dichotomies which differed only in the classification of the remaining point. Since there were $C(N, d)$ homogeneously linearly separable dichotomies of N points, it is clear that if one of the $C(N, d)$ separable dichotomies is selected at random according to an equiprobable distribution over the class, then a given point will be an extreme point with probability

$$2C(N-1, d-1)/C(N, d).$$

Then the expected number of extreme points $R(N, d)$ will be equal to the sum of the N probabilities that each point is an extreme point. Since these probabilities are equal,

$$E\{R(N, d)\} = \frac{2NC(N-1, d-1)}{C(N, d)}. \tag{39}$$

Utilizing 35, we may show that

$$\lim_{\substack{N=[\beta d] \\ d\to\infty}} E\left\{\frac{R(N, d)}{2d}\right\} = \begin{cases} \dfrac{\beta}{2}, & 0 \le \beta \le 2 \\ 1, & \beta \ge 2 \end{cases}. \tag{40}$$

See Cover [21] and [23] for related geometrical results.

Note that the limiting average number of extreme vectors is independent of β for $\beta \geq 2$. The capacity has played a role again. Also observe that, since $R(N, d)$ is a positive random variable, the probability is less than $1/t$ that R exceeds its mean value by a factor t.

The conclusion may be drawn that the average amount of necessary and sufficient information for the complete characterization of the set of separating surfaces for a *random, separable* dichotomy of N points grows slowly with N and asymptotically approaches $2d$ (twice the number of degrees of freedom of the class of separating surfaces). The implication for pattern-recognition devices is that the essential information in an infinite training set can be expected to be stored in a computer of finite storage capacity.

VIII. Conclusions

The original work of Winder, Joseph, Cameron, Schläfli, and others on counting the number of linearly separable dichotomies of a set of points has been developed as a unified whole and extended to counting 1) the number of nonlinearly separable dichotomies of a set of points, and 2) the separable dichotomies of random collections of points. Application of this work to the vertices of a binary n-cube has yielded upper bounds on the number of polynomially separable Boolean functions. It has been shown that the natural separating capacity of a family of separating surfaces is two pattern points for each degree of freedom, thus extending the work done by Koford and Winder. The separating capacity was seen to arise naturally as a critical parameter in defining the probability of unambiguous generalization, and in defining the number of extreme points characterizing the set of pattern points.

It was shown that, for a random set of linear inequalities in d unknowns, the expected number of extreme inequalities, which are necessary and sufficient to imply the entire set, tends to $2d$ as the number of consistent inequalities tends to infinity, thus bounding the expected necessary storage capacity for linear decision algorithms in separable problems. The results, even those dealing with randomly positioned points, have been combinatorial in nature, and have been essentially independent of the configuration of the set of points in the space.

Acknowledgment

The author wishes to express his gratitude to N. Abramson, B. Efron, N. Nilsson, and L. Zadeh for their comments and suggestions. He also wishes to thank B. Brown, J. Koford, and B. Widrow for the formulation of several related problems which led to the studies in this report, and B. Elspas of Stanford Research Institute for the limiting argument used in Section VI.

A paper on related questions is in preparation with B. Efron.

References

[1] Winder, R. O., Single stage threshold logic, *Switching Circuit Theory and Logical Design*, AIEE Special Publications S-134, Sep 1961, pp 321–332.
[2] ——, Threshold logic, Ph.D. dissertation, Princeton University, Princeton, N. J., 1962.
[3] Cameron, S. H., Tech Rept 60-600, *Proceedings of the Bionics Symposium*, Wright Air Development Division, Dayton, Ohio, 1960, pp 197–212.
[4] Joseph, R. D., The number of orthants in n-space intersected by an s-dimensional subspace, Tech Memo 8, Project PARA, Cornell Aeronautical Lab., Buffalo, N. Y., 1960.
[5] Whitmore, E. A., and D. G. Willis, Division of space by concurrent hyperplanes, unpublished Internal Rept, Lockheed Missiles & Space Co., Sunnyvale, Calif., 1960.
[6] Schläfli, L., *Gesammelte Mathematische Abhandlungen I.* Basel, Switzerland: Verlag Birkhäuser, 1950, pp 209–212.
[7] Wendel, J. G., A problem in geometric probability, *Mathematica Scandinavica*, vol 11, 1962, pp 109–111.
[8] Cooper, P. W., The hyperplane in pattern recognition, *Cybernetica*, no 4, 1962, pp 215–238.
[9] ——, The hypersphere in pattern recognition, *Information and Control*, vol 5, Dec 1962.
[10] Aizerman, M. A., E. M. Braverman, and L. I. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, *Automatika i Telemekhanika*, vol 25, Jun 1964; translation puplished Jan 1965, pp 821–837.
[11] Kaszerman, P., A nonlinear-summation threshold device, *IEEE Trans. on Electronic Computers (Correspondence)*, vol EC-12, Dec 1963, pp 914–915.
[12] Greenberg, H. J., and A. G. Konheim, Linear and nonlinear methods in pattern classification, *IBM J. Res. Develop.*, vol 8, Jul 1964, pp 299–307.
[13] Bishop, A. B., Adaptive pattern recognition, 1963 WESCON Rept of Session 1.5, unpublished.
[14] Wong, E., and E. Eisenberg, Iterative synthesis of threshold functions. To appear in *J. Mathematical Analysis and Applications.*
[15] Cooper, J. A., Orthogonal expansion applied to the design of threshold element networks, Rept SEL-63-123, TR 6204-1, Stanford Electronics Labs., Stanford University, Stanford, Calif., Dec 1963.
[16] Koford, J., Adaptive network organization, Rept SEL-63-009, *Stanford Electronics Laboratories Quarterly Research Review*, no 3, 1962, III-6.
[17] Novikoff, A., On convergence proofs for perceptrons, *Symposium on Mathematical Theory of Antomata.* Brooklyn, N. Y.: Polytechnic Press, 1963, pp. 615–622.
[18] Widrow, B., Generalization and information storage in network: of adaline "neurons," *Self Organizing Systems.* Washington: Spartan Books, 1962, pp 442, 459.
[19] Brown, R., Logical properties of adaptive networks, Rept SEP-62-109, *Stanford Electronics Laboratories Quarterly Research Review*, no 1, 1962, pp 87–88.
[20] Winder, R. O., Bounds on threshold gate realizability, *IEEE Trans. on Electronic Computers (Correspondence)*, vol EC-12, Oct 1963, pp 561–564.
[21] Efron, B., and T. Cover, Linear separability of random vectors, unpublished Internal Rept, Stanford Research Institute, Menlo Park, Calif., Mar 1963.
[22] Cover, T., Classification and generalization capabilities of linear threshold units, Documentary Rept RADC-TDR-64-32, Rome Air Development Center, Griffiss AFB, N. Y., Feb 1964.
[23] Cover, T. M., Geometrical and statistical properties of linear threshold devices, Rept SEL-64-052, TR 6107-1, Stanford Electronics Labs., Stanford University, Stanford, Calif., May 1964.
[24] Brown, R. J., Adaptive multiple-output threshold systems and their storage capacities, TR 6771-1, Stanford Electronics Labs., Stanford University, Stanford, Calif., Jun 1964.
[25] Bahadur, R. R., Some approximations to the binomial distribution function, *Annals of Mathematical Statistics*, vol 31, Mar 1960, pp 43–54.
[26] Mays, C. H., Adaptive threshold logic, Rept SEL-63-927, TR 1557-1, Stanford Electronics Labs., Stanford University, Stanford, Calif., Apr 1963.