AIM- 140

Marvin Minsky
Seymour Papert

LINEARLY UNRECOGNIZABLE PATTERNS

Reprinted from
Proceedings of Symposia in Applied Mathematics, Volume 19
MATHEMATICAL ASPECTS OF COMPUTER SCIENCE

Marvin Minsky
Seymour Papert

LINEARLY UNRECOGNIZABLE PATTERNS[1]

Introduction. The central theme of this study is the classification of certain geometrical properties according to the type of computation necessary to determine whether a given figure has them. Consider, for example, the following algorithm to determine whether a figure X is *convex*. For each pair of points (p,q) we define the function

$$\begin{aligned}\phi_{pq}(X) &= 1 \text{ if } (p \in X \text{ and } q \in X \text{ and midpoint } (p,q) \notin X)\\ &= 0 \text{ otherwise.}\end{aligned}$$

Then X is convex if and only if no $\phi_{pq}(X) = 1$ for any pair of points (p,q).

This shows, in a sense we shall presently define more precisely, that the "global" property of convexity can be determined by a simple computation from the "local" properties ϕ_{pq}.[2] Thus, if $\phi_{\text{convex}}(X)$ means that "X is convex," we have

$$\phi_{\text{convex}}(X) \Leftrightarrow \sum_{\text{all } p,q} \phi_{pq}(X) < 1.$$

Now we generalize this. We say that a *property* ψ *is of order* k if k is the smallest integer for which there exists a family Φ of predicates each of which depends only on a subset of k points of the figure X, and real numbers α_ϕ associated with each member ϕ of Φ such that

$$\psi(X) \Leftrightarrow \left(\sum_{\phi \in \Phi} \alpha_\phi \phi(X) > 0\right)$$

In this sense, we can assert that *the order of* ϕ_{convex} *is at most* 3. The determination of the orders of simple geometrical properties turns out to be far from trivial and presents many surprises. In fact, the greater part of the following analysis seems to be needed to prove that *connectedness*

[1] This work was supported in part by Project MAC, and M.I.T. research project sponsored by the Advanced Research Projects Agency, Department of Defense, under Office of Naval Research contracts Nonr-4102(01) and (02).

[2] We identify "property" (or "predicate") with the characteristic function of the set of objects that have the property.

is not of any finite order, i.e.,

There is no k for which connectedness is of order k.

This remains true if we relax the definitions, as we shall, to make the sums finite by considering the plane as a fine-grained infinite chess-board, considering its squares to be points, and allowing only figures which contain each square entirely or not at all—that is, we will consider a discrete model of geometry.

Apart from the purely mathematical interest of the results that come from it, we consider the concept of *finite order* worthy of study for a number of reasons connected with the theories of computation and of pattern recognition. We shall briefly outline some of these reasons.

Motivation of This Study.

(a) *Local vs. global geometric properties.* In problems of geometric pattern recognition, one is led to ask: to what extent can one use "local" properties—evidence obtained from looking at small portions of an object—as a basis for judgements about the "global" character of the object. For example, one can distinguish "line" drawings from other pictures on the basis of the existence of no interior points in the drawing—and this can be determined by a simple combination of evidence obtained from examining arbitrary small neighborhoods. On the other hand, one cannot obtain "local" evidence in favor of a drawing being "connected"—or so one might suspect—without having to combine such collections of evidence by a very complicated procedure.

Our first attempt to study this was based on the idea of *diameter-restricted predicates*, i.e., the restriction on the "local" properties is on the diameter of the set of points on which they depend rather than the number of points. The results of this study are summarized in §IX. However it soon became clear that the more interesting concept is order-restriction, and that the distinction we were seeking was not so much a question of geometry as a question about the theory of computation.

(b) *Serial vs. parallel computation.* What characterizes the extent to which an algorithm can have an essentially *serial*, as opposed to *parallel*, character? That is, to what degree can a computation be sped-up by doing several subcomputations at the same time? One would suspect, for example, that in many successive approximation computations there is little to be gained except, at great expense and redundancy, by parallel processing. We were led to suppose that the same is true for geometric connectedness recognition. One way to recognize that a set is disconnected (connected) is to find that there is a (no) curve dividing the set without intersecting it. One could therefore examine in parallel all possible separating curves, rather than serially trace through the paths within the set. But it would seem that the price of speeding-up the computation this way is superbly costly, and one looks for a way to get theoretical estimates of what is the exchange rate between the minimal serial and parallel amounts of computation. (The goal must be an exchange-cost curve.) One might hope that study of a particular problem, e.g., connectedness, would yield some insight into this general question of computation complexity for finite problems comparable with, say, that achieved in the theory of complexity of the recursive functions (Blum).

(c) *Theory of perceptrons and linear separability.* The pattern-recognition scheme known as the *perceptron* (Rosenblatt [6]) is known to be capable of learning to make any pattern discrimination which is within the scope of its potential ability—that is, if there is a set of parameter values that will suffice, it will find them. Thus, a good deal is known about this system's learning ability, and therefore one is particularly interested to know what is the scope of potential ability. Curiously enough, there seems to be nothing in the large perceptron literature on this question, and the present paper seems to be the first to link the linear-separation problem with the geometric-property problem.

The perceptron (and its derivatives) are of considerable interest mathematically because they are perhaps the simplest nontrivial parallel machines. One therefore ought to understand them thoroughly—as a sort of "linear case"—if one is to get any satisfactory theory of "higher-order" parallel computation schemes.

(d) *Mathematical aspects.* Linear separation computations have considerable mathematical significance in themselves. For example, if we ask for a maximum likelihood decision process based on Bayesian use of the results of statistically independent experiments, one obtains (Minsky and Selfridge [8]) a linear separation procedure. For another example, the generalization (in §1) of Boolean disjunctive normal form appears to yield surprising and fruitful results. Finally, the combination of group theory and linear inequalities seems to promise some new combinatorial techniques.

I. **Theory of linear Boolean separation functions.** In this section we shall confine ourselves to the analysis of the linear representation of predicates defined on an abstract set R, without any additional mathematical structure. The theorems proved here will be applied in later sections to sets with geometrical or topological structures. When necessary for truth R must be taken as finite.

Our theory deals with predicates defined on subsets of a given base space which we shall consistently denote by R. We use the following notational conventions:

(i) Let R be an arbitrary set and $\mathscr{F}$ a family of subsets of R. Using the letters $X, Y, Z, \cdots$ for subsets of R it is natural to associate with $\mathscr{F}$ a predicate $\phi_{\mathscr{F}}(X)$ which is **TRUE** if and only if $X \in \mathscr{F}$.

(ii) We shall use the letters ϕ and ψ to denote predicates defined on the set of subsets of R.

We shall use the notation $\psi(X)$ sometimes to mean the predicate whose value for a given X is **TRUE** or **FALSE**, sometimes to mean a binary set function whose value is 1 or 0. When we wish to employ the two senses in the same context we adapt the notation $\ulcorner \psi(X) \urcorner$ for the binary function whose value is 1 if $\psi(X)$ is **TRUE** and 0 if $\psi(X)$ is **FALSE**. We will usually use this only when there is a possibility of ambiguity, e.g., to distinguish between $\ulcorner 3 < 5 \urcorner = 1$, which is true, and $3 < \ulcorner 5 = 1 \urcorner$, which is false.

(iii) Occasionally it will be convenient in examples to use the traditional representation of $\psi(X)$ as a function of n "Boolean variables" where $n = |R|$. If the elements of R are $x_1, \cdots, x_n$, it is traditional to think of a subset X of R as an assignation of the values 1 or 0 to x_i according to whether the point x_i is in X or not, i.e., "x_i" is used ambiguously to stand for the ith point in the given enumeration of R, and for the set function $\ulcorner x_i \in X \urcorner$. This notation is particularly convenient when ψ is represented in the form of a standard Boolean function of two variables. Thus $x_i \vee x_j$ is a way of writing the set function

$$\phi(x) = \ulcorner x_i \in X \text{ or } x_j \in X \urcorner.$$

(iv) We need to express the idea that a function may depend only on a subset of the points of R. We denote by $S(\phi)$ the smallest subset S of R with the property that, for any subset X,

$$\phi(X) = \phi(X \cap S).$$

We call $S(\phi)$ the *support* of ϕ.

(v) Let Φ be a set of binary set-functions on R. We say that ψ is a *linear threshold function with respect to* Φ if to each member ϕ of Φ there corresponds a real number α_ϕ such that, for some real number θ:

$$\psi(X) = \ulcorner \sum_{\phi \in \Phi} \alpha_\phi \phi(X) > \theta \urcorner.$$

This is often written more briefly as

$$\psi = \ulcorner \sum \alpha_\phi \phi > \theta \urcorner.$$

We denote by $L(\Phi)$ the set of functions ψ expressible in this way.

(vi) We now introduce the central concept of *order*. The *order* of ψ is the smallest k for which there is a Φ satisfying

$$\psi \in L(\Phi),$$

$$\phi \in \Phi \Longrightarrow |S(\phi)| \leq k$$

where $|S(\phi)|$ is the cardinality of $S(\phi)$.

Functions of order 1 appear in the literature under the name of "linear threshold functions." It should be noted that the order of a constant function is zero, hence the number θ in the definition of $L(\Phi)$ can be replaced by 0 (or any other number) without changing the definition of order. Note also that the definition is unchanged if we use "$\geqq$," "$\leqq$," or "$<$" instead of "$>$" (assuming, when necessary, that R is finite).

(vii) ϕ is called a *mask* if there is a set A such that

$$\phi(X) = \lceil X \supset A \rceil.$$

We denote this function by ϕ_A.

In point-function notation a mask is a function of the form:

$$y_1 \wedge y_2 \wedge \cdots \wedge y_t$$

where $\{y_i\}$ is the subset A of R. In particular constant functions are masks.

Linear Representation.

Proposition. *All masks are of order* 1.

Proof. For each $x \in A$ define $\phi_x(X)$ as $\lceil x \in X \rceil$. Then

$$\phi_A = \lceil \sum_{x \in A} \phi_x \geqq |A| \rceil.$$

In particular the functions ϕ_x and ϕ_y are of order 1. Similarly the functions $x \vee y$, $x \wedge y$, $x \supset y$ are of order 1. But the "exclusive or," $x \oplus y$, and its complement, $x \equiv y$, are of order 2.

Example (i). $x_1 \vee x_2 \vee x_3$ is of order 1:

$$\lceil x_1 + x_2 + x_3 > 0 \rceil.$$

$x_1 \wedge x_2 \wedge x_3$ is also of order 1:

$$\lceil x_1 + x_2 + x_3 > 2 \rceil.$$

$x_1 \bar{x}_2 = \lceil x_1 + (1 - x_2) > 1 \rceil = \lceil x_1 - x_2 > 0 \rceil$ is of order 1.

$x_2 \vee \bar{x}_1 = \lceil x_2 + (1 - x_1) > 0 \rceil = \lceil x_2 - x_1 > -1 \rceil$, which is also $x_1 \supset x_2$, is of order 1.

Example (ii). $x_1 \equiv x_3$, which is

$$x_1 x_2 \vee \bar{x}_1 \bar{x}_2 = \lceil x_1 x_2 + (1 - x_1)(1 - x_2) > 0 \rceil$$
$$= \lceil 2x_1 x_2 - x_1 - x_2 > -1 \rceil$$

is of order 2. (Proof that it is not order 1 is in §II.)

Example (iii). Let M be an integer $0 < M < |R|$. Then the "counting function"

$$\psi^M(X) = \lceil |X| = M \rceil,$$

which recognizes when X contains *exactly* M points, is of order 2.

Proof. Consider the representation

$$\psi^M(X) = \lceil (2M-1)\sum_{\text{all } i} x_i + (-2)\sum_{i \neq j} x_i x_j \geq M^2 \rceil.$$

For any figure X there will be $|X|$ terms x_i with value 1, and $|X| \cdot (|X| - 1)/2$ terms $x_i x_j$ with value 1. Then the predicate is equal to

$$\psi^M(X) = \lceil (2M-1) \cdot |X| - |X| \cdot (|X| - 1) + 1 - M^2 > 0 \rceil$$

and the only (integer) value of $|X|$ for which this is true is $|X| = M$.

Note that the linear form for the counting function does not contain R explicitly. Hence it works as well for an infinite space R. Q.E.D.

Example (iv). The functions $\lceil |X| \geq M \rceil$ and $\lceil |X| \leq M \rceil$ are of order 1 because they are represented by $\lceil \sum x_i \geq M \rceil$ and $\lceil \sum x_i \leq M \rceil$.

Example (v). We can obtain an arbitrary function $f(|X|)$ of the area of a figure from the predicates used in (iv) above by writing

$$f(X) = f(0) + \sum_{k=1}^{R} (f(k) - f(k-1)) \cdot \lceil |X| > k \rceil.$$

The order of a function can be determined by examining its representation as a linear threshold with respect to sets of masks. To prove this we first show

Theorem (Positive normal form theorem). *Every ψ is a linear threshold function with respect to the set of all masks.*

Proof. The well-known disjunctive-normal-form theorem for Boolean functions tells us that any Boolean function $\psi(x_1, \cdots, x_n)$ can be written in the form (DNF)

$$\psi(X) = \bigvee_{i \in I} \psi_i(X)$$

where

$$\psi_i(X) = y_{i_1} y_{i_2} \cdots y_{i_n}$$

where for each i and j, $y_{ij} = x_j$ or $y_{ij} = \bar{x}_j$.

We can write this in linear form as:

$$\psi(X) = \lceil \sum_{i \in I} \psi_i > 0 \rceil$$

because, for any X, at most one term of the DNF is nonzero. Hence, we can replace logical "$\vee$" by arithmetic "$+$." Furthermore, since numerically $\overline{x}_i = 1 - x_i$, each ψ_i can be written in the form

$$\psi_i(X) = x_{i_1} \cdots x_{i_m}(1 - x_{i_{m+1}}) \cdots (1 - x_{i_n}),$$

supposing that the negative terms are at the right. Multiplying this out, we obtain an expression of the form

$$\psi_i(X) = \sum \beta_j Z_j$$

where Z_j is of the form

$$x_{i_1} \cdots x_{i_m} x_{h_1} \cdots x_{h_{m'}}, \text{ with } \{h_1, \cdots, h_{m'}\} \subset \{i_{m+1}, \cdots, i_n\}.$$

But such Z_j are masks, so that ψ_i is a linear combination of masks. It follows immediately that $\sum_{i \in I} \psi_i$ is itself a linear combination of masks

$$\psi = \sum \alpha_i Z_i$$

where each α_i is an integer and each Z_i a mask. Q.E.D.

Remark. The above construction shows not only that any Boolean function is "linear" in the set of masks in the "$\psi = \lceil \sum \alpha_i \phi_i > \theta \rceil$" sense, but is also linear in a stronger "$\psi = \sum \alpha_i \phi_i$" sense. It is interesting that this form is unique, and is therefore entitled to be called a "normal form." We call it a "positive normal form." To see the uniqueness, suppose that

$$\psi = \sum \alpha_i Z_i = \sum \beta_i Z_i$$

and consider the difference

$$\phi = (\sum \alpha_i Z_i - \sum \beta_i Z_i) = \sum (\alpha_i - \beta_i) Z_i = \sum \gamma_i Z_i.$$

Now $\phi(X)$ must be identically zero. To see this, consider first any set X of one element x_i. Then

$$\phi(X) = \phi(\{x_i\}) = \gamma_{\{x_i\}} x_i = \gamma_{\{x_i\}} \cdot 1 = 0$$

so $\gamma_{\{x_i\}} = 0$. Next, consider any two-element $X = \{x_i x_j\}$; then

$$\begin{aligned} \phi(\{x_i, x_j\}) &= \gamma_{\{x_i, x_j\}} x_i x_j + \gamma_{\{x_i\}} x_i + \gamma_{\{x_j\}} x_j \\ &= \gamma_{\{x_i, x_j\}} \cdot 1 = 0 \end{aligned}$$

so all two-element $\gamma_{\{x_i, x_j\}}$'s are zero. Similarly, by induction, all the γ's can be seen to vanish.

The proof of the positive normal form theorem implies also the

Theorem. *ψ is of order k iff k is the smallest number for which there exist a set Φ of masks satisfying*

$$\phi \in \Phi \Rightarrow |S(\phi)| \leq k$$

and

$$\psi \in L(\Phi).$$

Example (vi). A "Boolean form" has order no higher than the degree in its disjunctive normal form. Thus

$$\sum \alpha_{ijk} x_i x_j \overline{x}_k = \sum \alpha_{ijk} x_i x_j - \sum \alpha_{ijk} x_i x_j x_k$$

so that the negations can be removed without raising order. This particular order-3 form appears later in a perceptron that recognizes convex figures. Using this result we develop some more examples of the use of the concept of order.

Theorem. *If ψ_1 has order O_1 and ψ_2 has order O_2, then $\psi_1 \oplus \psi_2$ and $\psi_1 \equiv \psi_2$ have order* $\leq O_1 + O_2$.

Proof. The idea is to multiply together the positive mask representations $\lceil (\Sigma_1 \phi - \theta_1)(\Sigma_2 \phi - \theta_2) > 0 \rceil$ to get a positive form of order $\leq O_1 + O_2$. (Use "$>$" for $\equiv$ and "$<$" for $\oplus$.) This may not work in some cases where $\Sigma_1 = \theta_1$ or $\Sigma_2 = \theta_2$. In such cases, it is always possible to replace θ_1 and θ_2 by slightly different values, algebraically independent of the coefficients of Σ_1 and Σ_2, so that the predicates are unchanged but exact equality never holds.

Application.

Example (vii). Since $\psi^M(X) = \lceil \lceil |X| \geq M \rceil \equiv \lceil |X| \leq M \rceil \rceil$, we conclude that ψ^M has order ≤ 2, the result of Example (iii).

Question. What can be said about the orders of $\lceil \psi_1 \wedge \psi_2 \rceil$ and $\lceil \psi_1 \vee \psi_2 \rceil$? The answer to this question may be surprising, in view of the simple result of the previous theorem: it is shown in §V that for any order n, there exists a pair of predicates ψ_1 and ψ_2 both of order 1 for which $(\psi_1 \wedge \psi_2)$ and $(\psi_1 \vee \psi_2)$ have order $> n$. In fact suppose that $R = A \cup B \cup C$ where A, B, and C are large disjoint subsets of R. Then $\psi_1 = \lceil |X \cap A| > |X \cap C| \rceil$ and $\psi_2 = \lceil |X \cap B| > |X \cap C| \rceil$ each have order 1 because they are represented by

$$\lceil \sum_{x_i \in A} x_i - \sum_{x_i \in C} x_i > 0 \rceil \quad \text{and} \quad \lceil \sum_{x_i \in B} x_i - \sum_{x_i \in C} x_i > 0 \rceil$$

but, as shown in §V, $(\psi_1 \wedge \psi_2)$ and $(\psi_1 \vee \psi_2)$ have high orders.

II. **Group theory of linear inequalities.** In this section we consider linear threshold functions that are invariant under groups of permutations of the points of the base-space R. The purpose of this, realized finally in §V, is to establish a connection between the geometry of R and the question of when a geometric predicate can be a linear threshold function.

As an introduction to the methods introduced in this section we first consider a simple, almost trivial example. Suppose we wish to prove that the function $x_1 x_2 \vee \overline{x}_1 \overline{x}_2$ is not of order 1. To do so we might try to deduce a contradiction from the hypothesis that numbers α, β and θ can be found for which

$$\psi(x_1, x_2) = x_1 x_2 \vee \overline{x_1}\,\overline{x_2} = \lceil \alpha x_2 + \beta x_1 > \theta \rceil. \tag{1}$$

We could proceed directly by writing down the conditions on α and β:

$$\begin{aligned} x_1 = 0\,, \quad & x_2 = 0 \Rightarrow 0 > \theta, \\ x_1 = 1, \quad & x_2 = 0 \Rightarrow \alpha \leqq \theta, \\ x_1 = 0, \quad & x_2 = 1 \Rightarrow \beta \leqq \theta, \\ x_1 = 1, \quad & x_2 = 1 \Rightarrow \alpha + \beta > \theta. \end{aligned}$$

In this simple case it is easy enough to deduce the contradiction. But arguments of this sort are hard to generalize to more complex situations involving many variables. On the other hand the following argument, though it may be considered more complicated in itself, leads to elegant generalizations. First observe that the value of ψ is invariant under permutation of x_1 and x_2, that is,

$$\psi(x_1, x_2) = \psi(x_2, x_1).$$

Thus

$$\begin{aligned} \alpha x_1 + \beta x_2 > \theta, \\ \alpha x_2 + \beta x_1 > \theta; \end{aligned}$$

yields

$$((\alpha + \beta)/2)\, x_1 + ((\alpha + \beta)/2)\, x_2 > \theta$$

by adding the inequalities.

Similarly

$$\begin{aligned} \alpha x_1 + \beta x_2 \leqq \theta, \\ \alpha x_2 + \beta x_1 \leqq \theta \end{aligned}$$

yields

$$((\alpha + \beta)/2)\, x_1 + ((\alpha + \beta)/2)\, x_2 \leqq \theta.$$

It follows that if we write γ for $(\alpha + \beta)/2$, then

$$\psi(x_1, x_2) = \lceil \gamma x_1 + \gamma x_2 > \theta \rceil;$$

i.e., we can assume that the coefficients of x_1 and x_2 in the linear representation of ψ are equal. It follows that

$$\psi(X) = \lceil \gamma |X| > \theta \rceil \quad \text{or} \quad \lceil \gamma |X| - \theta > 0 \rceil$$

(if we assume that the space X has only the two points x_1 and x_2).

Now consider three values of X,

$$\begin{array}{lll} X_0 = \Lambda, & |X_0| = 0, & \gamma|X| - \theta \leq 0, \\ X_1 = \{x_2\}, & |X_1| = 1, & \gamma|X| - \theta > 0, \\ X_2 = \{x_1, x_2\}, & |X_2| = 2, & \gamma|X| - \theta \leq 0. \end{array}$$

Since X_0 and X_2 *satisfy* ψ, and X_1 does not, *the first-degree polynomial* $\gamma|X| - \theta$ *in* $|X|$ *would have to change direction twice, from positive to negative and back to positive as* $|X|$ *increases from* 0 *to* 2. This is clearly impossible. *Thus we learn something about* ψ *by averaging it over the permutations that leave it invariant.* The method is similar to that used in Haar measure theory. In fact, for order 1, it is the same method.

The generalization of this procedure involves consideration of groups of permutations on the set R and functions ψ invariant under these groups of permutations. In anticipation of application to geometrical problems, we recall the mathematical viewpoint from which every interesting geometrical property is an invariant of some natural transformation group.

Let G be a group of permutations of R; $g \in G$ and $X \subset R$, and define

$$\begin{aligned} X_g &=_{df} \{y \mid y = xg,\ x \in X\}, \\ \psi^g(X) &=_{df} \psi(X_g), \\ \psi =_G \phi &=_{df} (\exists g \in G)(\psi = \phi^g). \end{aligned}$$

Thus we define an equivalence relation of ϕ's with respect to a group G.

THE GROUP INVARIANCE THEOREM. *Let*

(i) *G be a finite group of permutations of R;*

(ii) *Φ be a set of predicates on R closed under G, i.e., $\phi \in \Phi$, $g \in G \Rightarrow \phi^g \in \Phi$;*

(iii) *ψ be in $L(\Phi)$ and invariant under G.*

Then there exists a linear representation of ψ,

$$\psi = \left\lceil \sum_{\phi \in \Phi} \beta_\phi \phi > 0 \right\rceil$$

for which the coefficients β_ϕ *depend only on the G-equivalence class of* ϕ, *i.e.*,

$$\phi =_G \phi' \Rightarrow \beta_\phi = \beta_{\phi'} .$$

PROOF. Divide Φ into equivalence classes by the relation $=_G$:

$$\Phi = \Phi_1 \cup \cdots \cup \Phi_k.$$

Now let $\psi = \lceil \sum_{\phi \in \Phi} \alpha_\phi \phi(X) > 0 \rceil$ be any linear representation of ψ and choose X such that $\psi(X)$ i.e., $\sum_{\phi \in \Phi} \alpha_\phi \phi(X) > 0$.

Since $\psi(X) = \psi(X_g)$, it follows that for each $g \in G$,

$$\sum_{\phi\in\Phi} \alpha_\phi \phi(X_g) = \sum_{\phi\in\Phi} \alpha_\phi \phi^g(X) > 0.$$

Since the sum of positive quantities is positive we can sum all such equations:

$$\sum_{g\in G} \sum_{\phi\in\Phi} \alpha_\phi \phi^g(X) > 0.$$

Since $\Phi = \bigcup_{i=1}^{k} \Phi_i$, the expression on the left can be written:

$$\sum_{g\in G} \sum_{i=1}^{k} \sum_{\phi\in\Phi_i} \alpha_\phi \phi^g = \sum_{i=1}^{k} \sum_{g\in G} \sum_{\phi\in\Phi_i} \alpha_\phi \phi^g.$$

Hence,

$$\sum_{i=1}^{k} \sum_{\phi\in\Phi_i} \left(\sum_{g\in G} \alpha_\phi \phi^g(X) \right) > 0.$$

Now observe that the set

$$\Phi_i g = \{\phi g \mid \phi\in\Phi_i\} = \{\phi \mid \phi\in\Phi_i\} = \Phi_i$$

because any g just permutes members of an equivalence class. Then also,

$$\Phi_i = \Phi_i g^{-1}.$$

Hence for any g

$$\sum_{\phi\in\Phi_i} \alpha_\phi \phi^g = \sum_{\phi\in\Phi_i g^{-1}} \alpha_\phi \phi^g = \sum_{\phi\in\Phi_i} \alpha_{\phi g} \phi.$$

So

$$\sum_{g} \sum_{\phi\subset\Phi_i} \alpha_\phi \phi^g = \sum_{g} \sum_{\phi\in\Phi_i} \alpha_{\phi g^{-1}} \phi = \sum_{\phi\in\Phi_i} \left(\sum_{g} \alpha_{\phi g^{-1}} \right) \phi.$$

Since as g runs over G, ϕ^g "covers" $\Phi_{E(\phi)}$, then $\sum_{g\in G} \alpha_{\phi g^{-1}}$ has the same value for all equivalent ϕ's, i.e., if $\phi\in\Phi_i$, $\sum_{g\in G}\alpha_{\phi g^{-1}}$ depends only on i. Therefore we can denote $\sum_{g\in G}\alpha_{\phi g^{-1}}$ by β_i obtaining:

$$\sum_{i=1}^{k} \sum_{\phi\in\Phi_i} \beta_i \phi(X) > 0$$

or

$$\sum \beta_{E(\phi)} \phi(X) > 0,$$

where $E(\phi)$ denotes "the equivalence class containing ϕ."

A similar argument shows that if $\sum \alpha_\phi \phi(X) < 0$, then $\sum \beta_{E(\phi)} \phi(X) < 0$. Thus $\psi = \lceil \sum \alpha_\phi \phi > 0 \rceil = \lceil \sum \beta_{E(\phi)} \phi > 0 \rceil$. We shall most often use this theorem in the following form:

Corollary 1. *Any function ψ, of order k has a linear representation*

$$\psi = \lceil \sum_{\phi} \alpha_\phi \phi > 0 \rceil$$

where Φ is the set of masks of degrees $\leq k$ and $\alpha_\phi = \alpha_{\phi'}$ wherever $S(\phi)$ can be transformed into $S(\phi')$ by an element of G.

PROOF. The corollary follows immediately from the theorem and the observation that, for masks, $\phi_A \equiv_G \phi_B$ if and only if $A = B_g$ for some $g \in G$.

COROLLARY 2. *Let $\Phi = \Phi_1 \cup \cdots \cup \Phi_m$ be the decomposition of Φ into equivalence classes by the relation $\equiv_G$. Then if ψ is in $L(\Phi)$ and Φ is closed under G, ψ can be written in the form*

$$\psi = \lceil \sum \alpha_i N_i(X) > 0 \rceil$$

where $N_i(X) = |\{\phi | \phi \in \Phi_i;\ \phi(X)\}|$, i.e., $N_i(X)$ is the number of ϕ's of the i-th type, equivalent under the group, that "fit" the argument X.

PROOF. ψ can be represented as

$$\begin{aligned}\psi &= \lceil \sum_{\phi \in \Phi} \alpha_\phi \phi > 0 \rceil \\ &= \lceil \sum_i \sum_{\phi \in \Phi_i} \alpha_\phi \phi > 0 \rceil \\ &= \lceil \sum_i \alpha_i \sum_{\phi \in \Phi_i} \phi > 0 \rceil = \lceil \sum_i \alpha_i N_i(X) > 0 \rceil .\end{aligned}$$

COROLLARY 3. (THE TRIVIALITY OF INVARIANT PREDICATES OF ORDER 1). *Let G be any transitive group of permutations on R (transitive means: for every pair $p, q \in R$ there is a $g \in G$ such that $pg = q$). Then the only first-order predicates invariant under G are of the forms:*

$$\psi(X) = \lceil |X| > m \rceil,$$

or

$$\psi(X) = \lceil |X| < m \rceil, \quad \textit{for some } m.$$

PROOF. Since the group is transitive all the one-point predicates $\phi_{\{p\}}$ are equivalent. Thus we can assume that

$$\psi(X) = \lceil \textstyle\sum_{p \in X} \alpha \phi_{\{p\}} > \theta \rceil \quad \text{(or with some other inequality sign)}$$

i.e., the coefficient α is independent of p. But $\sum_{p \in X} \alpha \phi_{\{p\}} > \theta$ can be transformed into $\sum_{p \in X} \phi_{\{p\}} > \theta / \alpha$ (for $\alpha > 0$; for $\alpha \leq 0$ a similar argument proves the corresponding assertion). But $\sum_{p \in X} \phi_{\{p\}} = |X|$. Thus order-1 invariant predicates can do nothing more than define a count on the cardinality or "area" of figures. In fact, an order-1 predicate is a measure, and the order-1 invariant predicate is the Haar measure.

III. **Applications of the group-invariance theorem.**

The Parity Function. In this section we develop in some detail the analysis of the particular predicate ψ_{PAR} defined by

$$\psi_{\text{PAR}}(X) = \lceil |X| \text{ is odd} \rceil .$$

Our interest in ψ_{PAR} is threefold: it is interesting in itself; it will be used for the analysis of other more important functions; and, especially, it illustrates our mathematical methods and the kind of question they enable us to discuss.

THEOREM. ψ_{PAR} *is of order* $|R|$.

That is, to compute ψ_{PAR} requires at least one predicate whose support covers the *whole space* R!

PROOF. Let G be the group of all permutations of R. Clearly ψ_{PAR} is invariant under G.

Now suppose that $\psi_{\mathrm{PAR}} = \lceil \sum \alpha_i \phi_i > 0 \rceil$ where the ϕ_i are masks with $|S(\phi_i)| \leq K$ and the α_i depend only on the equivalence classes defined by $\equiv_G$. Since masks with the same support are identical,

$$\phi_i \equiv_G \phi_j \Leftrightarrow |S(\phi_i)| = |S(\phi_j)|.$$

Thus

$$\psi_{\mathrm{PAR}} = \left\lceil \sum_{j=0}^{K} \left(\alpha_j \sum_{\phi \in \Phi_j} \phi \right) > 0 \right\rceil$$

where Φ_j is the set of masks whose supports contain exactly j elements. We now calculate for an arbitrary subset X of R,

$$C_j(X) = \sum_{\phi \in \Phi_j} \phi(X).$$

Since $\phi(X)$ is 1 if $S(\phi) \subset X$ and 0 otherwise, $C_j(X)$ is the number of subsets of X with j elements, i.e.,

$$C_j(X) = \binom{|X|}{j}$$

which is a polynomial of degree j in $|X|$.

It follows that

$$\psi_{\mathrm{PAR}} = \sum_{j=0}^{K} \alpha_j C_j(X)$$

is a polynomial of degree K in $|X|$, say $P(|X|)$.

Now consider a sequence

$$\Lambda = X_0 \subset X_1 \subset \cdots \subset X_{|R|} = R$$

of $|R| + 1$ nested subsets of R, and the sequence of values

$$P(|X_0|) = 0,\ P(|X_1|) = 1,\ P(|X_2|) = 0, \cdots, P(|X_{|R|}|).$$

This implies that $P(|X|)$ changes direction $|R|$ times as $|X|$ increases from 0 to $|R|$. But since P is a polynomial of degree K, it follows that $K = |R|$. Q.E.D.

From this we obtain the

THEOREM. *If $\psi_{\mathrm{PAR}} \in L(\Phi)$ and if Φ contains only masks, then Φ contains all the masks.*

PROOF. Suppose, if possible, that $\psi_{\mathrm{PAR}} \in L(\Phi)$, that Φ contains only masks, and the mask whose support is A does not belong to Φ.

Let $\psi_{\mathrm{PAR}} = \lceil \sum_{\phi \in \Phi} \alpha_\phi \phi > 0 \rceil$. Define, for any ψ, $\psi^A(X) = \psi(X \cap A)$. Clearly ψ^A_{PAR}, the parity function for subsets of A, is of order $|A|$ by the previous theorem.

Now consider ϕ^A for $\phi \in \Phi$. If $S(\phi) \subset A$, clearly $\phi^A = \phi$. If $S(\phi)$ is not a subset of A, ϕ^A is identically zero since

$$S(\phi) \not\subset A \Rightarrow S(\phi) \not\subset X \cap A \Rightarrow \phi(X \cap A) = 0 \Rightarrow \phi^A(X) = 0.$$

It follows that either $S(\phi^A)$ is a *proper* subset of A or ϕ^A is identically zero. Let Φ^A be the set of masks in Φ whose supports are subsets of A. Then $\psi^A_{\mathrm{PAR}} = \lceil \sum_{\phi \in \Phi^A} \alpha_\phi \phi > 0 \rceil$. But for all $\phi \in \Phi^A$, $|S(\phi)| < |A|$. It would follow that the order of ψ^A_{PAR} is less than $|A|$, which is a contradiction. Thus the hypotheses are impossible and the theorem follows. Q.E.D.

COROLLARY 1. *If $\psi_{\mathrm{PAR}} \in L(\Phi)$, then Φ must contain at least one ϕ for which*

$$|S(\phi)| = |R|.$$

The following theorem, also immediate from the above is of interest to students of threshold logic:

COROLLARY 2. *Let Φ be the set of all ψ^A_{PAR} for proper subsets A of R. Then* $\psi^R_{\mathrm{PAR}} \notin L(\Phi)$.

The following theorem gives a hint that certain functions that might be recognizable, in principle, by a very large perceptron, might not actually be realizable in practice because of huge coefficients.

Coefficients of the Parity Function. Suppose that we have a $\lceil \sum \alpha_i \phi_i > 0 \rceil$ that recognizes Parity $(|X|)$ with masks. Let us suppose that the recognition is *reliable*, e.g., that $\sum \alpha_i \phi_i > 2$ for odd parity, and $\sum \alpha_i \phi_i < 0$ for even parity. If we apply the full permutation group, we obtain the same reliable discrimination with a set of "average coefficients" α_i all equal for ϕ's of the same order. Then we obtain the inequalities

$$\left.\begin{array}{r} \alpha_1 > 2 \\ \alpha_2 + 2\alpha_1 < 0 \\ \alpha_3 + 3\alpha_2 + 3\alpha_1 > 2 \end{array}\right\} \quad \text{or} \quad \sum_{i=1}^{n} \binom{n}{i} \alpha_i \begin{array}{l} > 2, \text{ if } n \text{ is odd,} \\ < 0, \text{ if } n \text{ is even.} \end{array}$$

$$\cdots$$

Subtracting successive inequalities, define

$$D_n = \sum_1^{n+1} \binom{n+1}{i} \alpha_i - \sum_1^n \binom{n}{i} \alpha_i$$

$$= \alpha_{n+1} + \sum_1^n \left[\binom{n+1}{i} - \binom{n}{i} \right] \alpha_i = \alpha_{n+1} + \sum_1^n \binom{n}{i-1} \alpha_i$$

$$= \sum_0^n \binom{n}{i} \alpha_{i+1}$$

so that for all n,

$$(-1)^n D_n > 2 \quad \text{or} \quad [(-1)^n D_n - 2] > 0.$$

Using these inequalities, we will obtain a bound on the coefficients $\{\alpha_i\}$. We will sum the inequalities with certain positive weights; choose any $M > 0$, and consider

$$\sum_0^M \binom{M}{i} [(-1)^i D_i - 2] > 0.$$

Then

$$\sum_0^M \binom{M}{i} (-1)^i D_i > 2 \sum_0^M \binom{M}{i} = 2^{M+1}.$$

The left-hand side is

$$\sum_{i=0}^M \sum_{k=0}^i (-1)^i \alpha_{k+1} \binom{i}{k} \binom{M}{i} = \sum_{k=0} \sum_{i=k} (-1)^i \alpha_{k+1} \binom{i}{k} \binom{M}{i}$$

$$= \sum_{k=0}^M \sum_{i=K}^M (-1)^i \alpha_{k+1} \left(\frac{i!}{k!\,(i-k)!} \right) \left(\frac{M!}{i!\,(M-i)!} \right)$$

$$= \sum_{k=0}^M \sum_{i=k}^M (-1)^i \alpha_{k+1} \left(\frac{M!}{k!\,(M-k)!} \right) \left(\frac{(M-k)!}{(i-k)!\,(M-i)!} \right)$$

$$= \sum_{k=0}^M \alpha_{k+1} \binom{M}{k} (-1)^k \sum_{j=0}^{M-k} \left(\frac{(M-k)!}{j!\,(M-k-j)!} \right) (-1)^j$$

$$= \sum_{k=0}^M \alpha_{k+1} \binom{M}{k} (-1)^k (1-1)^{M-k}$$

$$= \alpha_{M+1} (-1)^M$$

so we have the

Theorem. *For each M,*

$$(-1)^M \alpha_{M+1} > 2^{M+1}.$$

These values hold for the average, so if the coefficients of each type are not equal, some must be even larger! This shows that it is impractical to use mask-like ϕ's to recognize parity-like functions: even if one could afford the huge number of ϕ's, one would have also to cope with huge ranges of their coefficients!

Remark. This has a practically fatal effect on the corresponding learning machines. At least $2^{|R|}$ instances of just the maximal pattern is required to "learn" the largest coefficient; actually the situation is far worse because of the unfavorable interactions with lower order coefficients. It follows, moreover that the information capacity necessary to store the set $\{\alpha_i\}$ of coefficients is greater than that needed to store the entire set of patterns recognized by ψ_{PAR}—that is, the even subsets of R. For, any uniform representation of the α_i's must allow $|R|$ bits for each, and since there are $2^{|R|}$ coefficients the total number of bits required is $|R| \cdot 2^{|R|}$. On the other hand there are $2^{|R|-1}$ even subsets of R, each representable by an $|R|$-bit sequence, so that $|R| \cdot 2^{|R|-1}$ bits would suffice to represent the subsets.

It should also be noted that ψ_{PAR} is not very exceptional in this regard because the positive normal form theorem tells us that all possible $2^{2^{|R|}}$ Boolean functions can be so encoded as linear threshold functions in the set of all masks. Then, on the average, specification of the coefficients of each requires $2^{|R|}$ bits.

Another predicate of great interest is associated with the geometric property of "connectedness:" Its application and interpretation is deferred to §V; the basic theorem is proved now.

The "One-in-a-box" Theorem.

Theorem. *Let $A_1, \cdots, A_m$ be disjoint subsets of R and define the predicate*

$$\psi(X) = \ulcorner (\forall\, i)(|X \cap A_i| > 0) \urcorner$$

i.e., there is at least one point of X in each A_i. Then if for all i, $|A_i| = 4m^2$, the order of ψ is $\geqq m$.

Corollary. *If $R = A_1 \cup A_2 \cup \cdots \cup A_m$, the order of ψ is at least the order of $(|R|/4)^{1/3}$.*

Proof. For each $i = 1, \cdots, m$ let G_i be the group of permutations of R which permutes the elements of A_i but do not affect the elements of the complement of A_i. Let G be the group generated by all elements of the G_i. Clearly ψ is invariant with respect to G. Let Φ be the set of masks of degree K or less. To determine the equivalence class of any $\phi \in \Phi$ consider the ordered set of occupancy numbers

$$\{|S(\phi) \cap A_i|\}.$$

Then $\phi_1 =_G \phi_2$ if, for each i, $|S(\phi_1) \cap A_i| = |S(\phi_2) \cap A_i|$. Let $\Phi_1, \Phi_2, \cdots, \Phi_M$ be the equivalence classes.

Now consider an arbitrary set X and an equivalence class Φ_j. We wish to calculate the number $N_j(X)$ of members of Φ_j satisfied by X, i.e.,

$$N_j(X) = |\{\phi | \phi \in \Phi_j \wedge S(\phi) \subset X\}|.$$

A simple combinatorial argument shows that

$$N_j(X) = \binom{|X \cap A_1|}{|S(\phi) \cap A_1|} \binom{|X \cap A_2|}{|S(\phi) \cap A_2|} \cdots \binom{|X \cap A_M|}{|S(\phi) \cap A_M|}$$

where

$$\binom{y}{n} = \frac{y(y-1) \cdots (y = n + 1)}{n!}$$

and ϕ is an arbitrary member of Φ_j. Since the numbers $|S(\phi) \cap A_i|$ depend only on the classes Φ_j and add up to not more than K, it follows that $N_j(X)$ can be written as a polynomial of degree K or less in the numbers $x_i = |X \cap A_i|$

$$N_j(X) = P_j(x_1, \cdots, x_n).$$

Now let $\psi = \lceil \sum \alpha_\phi \phi > 0 \rceil$ be a representation of ψ as a linear threshold function in the set of masks of degree less than or equal to K. By the argument which we have already used several times we can assume that α_ϕ depends only on the equivalence class of ϕ and write

$$\begin{aligned}\sum \alpha_\phi \phi(X) &= \sum_{j=1}^{M} \beta_j \sum_{\phi \in \Phi_j} \phi(X) = \sum_{j=1}^{M} \beta_j N_j(X) \\ &= \sum_{j=1}^{M} \beta_j P_j(x_1, \cdots, x_m)\end{aligned}$$

which, as a sum of polynomials of degree at most K, is itself such a polynomial. Thus we can conclude that there exists a polynomial of degree at most K,

$$Q(x_1, \cdots, x_m)$$

with the property that

$$\psi(X) = \lceil Q(x_1, \cdots, x_m) > 0 \text{ with } x_i = |X \cap A_i| \rceil$$

i.e., that for all i, $0 \leq x_i \leq 4m^2$

$$Q(x_1, \cdots, x_m) > 0 \Leftrightarrow (\forall i)(x_i > 0).$$

In $Q(x_1, \cdots, x_m)$ make the formal substitution,

$$x_i = (t - (2i - 1))^2.$$

Then $Q(x_1, \cdots, x_m)$ becomes a polynomial of degree at most $2K$ in t. Now let t take on the values $t = 0, 1, \cdots, 2m$. By property (ψ) Q must be positive for even t and negative or zero for odd t. By counting the number of changes of sign it is clear that $2K \geqq 2m$ i.e., $K \geqq m$. This completes the proof.

IV. **The and/or theorem.** We have already remarked that if $R = A \cup B \cup C$ the predicate

$$\psi_1(X) = \lceil |X \cap A| > |X \cap C| \rceil \quad \text{is of order 1,}$$

and stated without proof that

$$\psi(X) = \lceil\ |X \cap A| > |X \cap C| \wedge |X \cap B| > |X \cap C| \rceil$$

is not of bounded order as $|R|$ becomes large. We shall now prove this assertion. We can assume without any loss of generality that $|A| = |B| = |C|$ and our formal statement is that if $\psi_k(X)$ is the predicate of the stated form for $|R| = 3k$, then the order of $\psi_k \rightarrow \infty$ as $k \rightarrow \infty$. The proof is similar to that used for the parity theorem. We shall assume that the order of (ψ_k) is bounded by N for all k and derive the contradiction by showing that the associated polynomials would have to satisfy inconsistent conditions. The first step is to set up the associated polynomials for a fixed k. We do this by choosing the group which permutes within the sets A, B, C. The equivalence classes of masks are then characterized by three numbers, i.e., $|A \cap S(\phi)|$, $|B \cap S(\phi)|$ and $|C \cap S(\phi)|$. The number $N_\phi(X)$ of masks in this equivalence class satisfied by a given set X is

$$N_\phi(X) = \binom{|A \cap X|}{|A \cap S(\phi)|} \times \binom{|B \cap X|}{|B \cap S(\phi)|} \times \binom{|C \cap X|}{|C \cap S(\phi)|}.$$

If $|S(\phi)| \leqq N$ this is clearly a polynomial of degree at most N in the three numbers

$$x = |A \cap X|, \quad y = |B \cap X|, \quad z = |C \cap X|.$$

The group invariance theorem says that if

$$\psi_k = \lceil \sum_{\phi \in \Phi} \gamma_\phi \phi > 0 \rceil$$

when Φ is the set of masks with $|S(\phi)| \leqq N$, then

$$\psi_k(X) = \lceil \sum \alpha_i N_i(X) > 0 \rceil$$

where i runs over the set of equivalence classes of ϕ. But $\sum \alpha_i N_i(X)$ is a polynomial of degree at most N in x, y and z. Call it $P_k(x, y, z)$.

Now, by definition, for possible values of x, y, z (i.e., nonnegative integers $\leqq k$), $P_k(x, y, z) > 0$ if and only if $x > z$ **and** $y > z$. We shall show, through a series of lemmas, that this cannot be true for all k. The technical details of these lemmas are not essential for the subsequent sections.

Lemma 1. *Let $P_k(x,y,z)$ be an infinite sequence of polynomials of fixed degree n, with the property that for all positive integers x, y, z less than k,*

$$\text{(A)} \qquad \begin{aligned} x > z \ \textbf{and}\ y > z &\Rightarrow P_k(x,y,z) \geqq 0, \\ x \leqq z \ \textbf{or}\ y \leqq z &\Rightarrow P_k(x,y,z) \leqq 0. \end{aligned}$$

Then there exists a nonzero polynomial $P(x,y,z)$ of the same degree n with the property that the implications (A) hold for all positive integral values of x, y, z. This follows from the following compactness argument: Write

$$P_k(x,y,z) = \sum_{i=1}^{r} C_{k,i}\, m_i(x,y,z)$$

where $m_i(x,y,z)$ is an enumeration of the monomials in variables x, y and z. We can assume $\sum C_{k,i}^2 = 1$ since the hypotheses remain true if P_k is divided by $\sum C_{k,i}^2$. Now the bounded sequence $C_{1,1}, C_{2,1}, \cdots, C_{k,1}, \cdots$ must contain an infinite convergent subsequence S_1 of the integers for which

$$\{C_{k,1} \mid k \in S_1\} \text{ converges to a limit, say } C_1.$$

Now consider $\{C_{k,2} \mid k \in S_1\}$. There must be an infinite subsequence of S_1, say S_2, on which this converges to a limit, say C_2. Continuing in this way we find a subsequence $S = S_r$ of the integers and a set of numbers $C_1 \cdots C_r$ such that $\{C_{k,i} \mid k \in S\}$ converges to C_i for all $i \leqq r$. But then, for $k \in S$, $P_k(x,y,z)$ converges to the polynomial

$$P(x,y,z) = \sum_{i=1}^{r} C_i m_i(x,y,z) \quad \text{for all } x,y,z.$$

To see that $P(x,y,z)$ has the required properties, choose any positive integers x_0, y_0, z_0. For values of k smaller than the largest of these numbers, nothing can be said about $P_k(x_0,y_0,z_0)$. But for all sufficiently large k, $P_k(x_0,y_0,z_0)$ must be nonnegative if $x_0 > z_0$ and $y_0 > z_0$ (and nonpositive if $x_0 < z_0$ or $y_0 < z_0$). It follows immediately that $P(x_0,y_0,z_0)$ is nonnegative (nonpositive) under the same conditions. To see that P is nonzero note that $\sum c_i^2 = 1$.

Lemma 2. *If a polynomial $f(\alpha,\beta)$ satisfies the following conditions for all integral values of α and β, then it is identically zero*:

$$\text{(B)} \qquad \alpha > 0 \ \textbf{and}\ \beta > 0 \Rightarrow f(\alpha,\beta) \geqq 0,$$

$$\text{(C)} \qquad \alpha \leqq 0 \ \textbf{or}\ \beta \leqq 0 \Rightarrow f(\alpha,\beta) \leqq 0.$$

Proof. Suppose that a polynomial of degree N, $f(\alpha,\beta)$, satisfies the conditions (B) and (C) and is not identically zero. Without loss of generality we can suppose that

$$f(\alpha,\beta) = \alpha^N g(\beta) + r(\alpha,\beta)$$

where $g(\beta)$ is not identically zero and $r(\alpha,\beta)$ has degree less than N in α.

For any β for which $g(\beta) \neq 0$, there is an $\alpha_0 > 0$ such that

$$|\alpha_0^N g(\beta)| > |r(\alpha_0, \beta)|.$$

Thus $f(\alpha_0, \beta)$ has the same sign as $\alpha_0^N g(\beta)$, i.e., as $g(\beta)$ since α_0^N is positive. It follows from (B) and (C) that

$$\text{(D)} \qquad \begin{aligned} \beta > 0 &\Longrightarrow g(\beta) \leq 0, \\ \beta < 0 &\Longrightarrow g(\beta) \leq 0. \end{aligned}$$

(The conditions (D) hold for all β: if $g(\beta) \neq 0$ by preceding argument; if $g(\beta) = 0$, tautologously.) We now derive a contradiction by considering separately two cases:

(a) *N even*. Since $g(\beta)$ is not identically zero, there is some $\beta_0 > 0$ for which $g(\beta_0) \neq 0$. By (D), $g(\beta_0) > 0$. Thus $\alpha^N g(\beta_0) > 0$ so that for $|\alpha|$ sufficiently large

$$\alpha^N g(\beta_0) + r(\alpha, \beta_0) > 0$$

i.e., $f(\alpha, \beta_0) > 0$. But we are free to choose a negative value of α, i.e., we can find α_0, β_0 such that

$$\alpha_0 < 0 \quad \text{and} \quad f(\alpha_0, \beta_0) > 0$$

which contradicts (C).

(b) *N odd*. Choose $\beta_0 < 0$ for which $g(\beta_0) \neq 0$; then $g(\beta_0) < 0$, by (D). Choose negative α_0 as before. Then $\alpha_0^N g(\beta_0) > 0$ and $f(\alpha_0, \beta_0) > 0$, again contradicting (C). Q.E.D.

LEMMA 3. *No nonzero polynomial $P(x, y, z)$ can satisfy the following conditions for all positive integral values of x, y, z:*

$$\begin{aligned} x > z \quad \textbf{and} \quad y > z &\Longrightarrow P(x, y, z) \geqq 0, \\ x \leqq z \quad \textbf{or} \quad y \leqq z &\Longrightarrow P(x, y, z) \leqq 0. \end{aligned}$$

PROOF. Suppose that $P(x, y, z)$ has these properties. Define $Q(\alpha, \beta, z) = P(z + \alpha,\ z + \beta,\ z)$. Let M be the highest power of z in Q so that

$$Q(\alpha, \beta, z) = z^M f(\alpha, \beta) + R(\alpha, \beta, z)$$

where R is of degree less than M in z.

Now choose any α_0 and β_0 for which $f(\alpha_0, \beta_0) \neq 0$. For sufficiently large z, say z_0

(a) $z_0 + \alpha_0 > 0$ and $z_0 + \beta_0 > 0$,

(b) $|z_0^M f(\alpha_0, \beta_0)| > |R(\alpha_0, \beta_0, z_0)|$.

It follows that

$$\begin{aligned} f(\alpha_0, \beta_0) \geqq 0 &\Longleftrightarrow Q(\alpha_0, \beta_0, z_0) \geqq 0, \\ &\Longleftrightarrow P(z_0 + \alpha_0, z_0 + \beta_0, z_0) \geqq 0. \end{aligned}$$

Thus

$$\begin{aligned}\alpha_0 > 0 \quad \text{and} \quad \beta_0 > 0 &\Rightarrow z_0 + \alpha_0 > z_0, \text{ and } \quad z_0 + \beta_0 > z_0 \\ &\Rightarrow P(z_0 + \alpha_0, z_0 + \beta_0, z_0) \geqq 0 \\ &\Rightarrow f(\alpha_0, \beta_0) \geqq 0\end{aligned}$$

and similarly $\alpha_0 < 0$ or $\beta_0 < 0 \Rightarrow f(\alpha_0, \beta_0) \leqq 0$. But this is true for all α_0, β_0. Thus by the previous lemma, $f(\alpha, \beta) = 0$. It follows that $P(x, y, z)$ is of degree zero in z, which is only possible if it is identically zero. Q.E.D.

This concludes the proof of the AND-OR theorem. It is clear that the reason the theorem is true has to do with the algebraic geometry of the "occupancy" polynomials. If it were not for the constraints concerning integer values of the variables, the theorem would be an immediate consequence of Bezout's theorem.

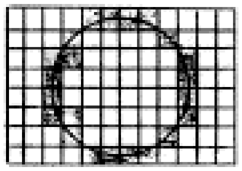
V. **The "order-limited" perceptron.**

The Order of Some Geometrical Predicates. Now we consider the problem of computing the *order* of a number of interesting geometrical predicates. As a first step, we have to provide the underlying space R with the topological and metric properties necessary for defining geometrical figures; this was not necessary in the case of predicates like Parity and others related to counting, for these are not really geometric in character.

The simplest procedure that is rigorous enough yet not too mathematically fussy seems to be to divide the Euclidean plane, E^2, into squares as an infinite chess board. The set R is then taken as *the set of squares*. A figure X of E^2 is then identified with that set of elements of R—i.e., that collection of squares—that contain at least one point of X. Thus to any subset X of E^2 corresponds the subset $\hat{X}$ of R defined by

$$\hat{X} = \{ \hat{x} \in R \mid \hat{x} \cap X \neq \Lambda \}.$$

Now, although X and $\hat{X}$ are logically distinct no serious confusion can arise if we identify them, and we shall do so from now on. Thus we refer to certain subsets of R as "circles," "triangles," etc., meaning that they can be obtained from real circles and triangles by the map $X \rightarrow \hat{X}$. Of course, this means that near the "limits of resolution" one begins to obtain apparent errors of classification because of the finite "mesh" of R. Thus a small circle



will not look very round.

When it is necessary to distinguish between F and F we will say that two figures X, X' of E^2 are in the same R-tolerance class if $\hat{X} = \hat{X}'$. In this we follow the general mathematical approach proposed by E. C. Zeeman for treating this kind of problem. To avoid inessential questions of how the group-invariance theorem applies to infinite groups, assume below when necessary that R has the toroidal topology.

We begin by listing some geometric predicates of rather small order.

(a) $k = 1$. When we say "geometric property" we mean something that is at least invariant under translation, usually also invariant under rotation, and often invariant under dilatation. The first two invariances combine to define the "congruence" group of transformations and all three the "similarity" group. For $k = 1$, just the translation group suffices for the Group Invariance Theorem to tell us that all coefficients are equal, hence the only patterns that can be of order 1 are those defined by a single cut in the cardinality or area of the set:

$$\psi = \lceil |X| > A \rceil \quad \text{or} \quad \psi = \lceil |X| < A \rceil.$$

Note: If translation invariance is *not* required, then order-1 can compute other properties, i.e., concerning *moments* about *particular* points or axes. However these are not "geometric."

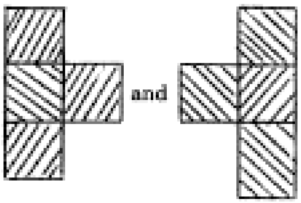
(b) $k = 2$. For $k = 2$ things are more complicated. As shown in §I it is possible to make a double cut in the area of the set, hence we can do the counting trick, and recognize those figures whose areas are

$$\psi = \lceil A_1 < |F| < A_2 \rceil.$$

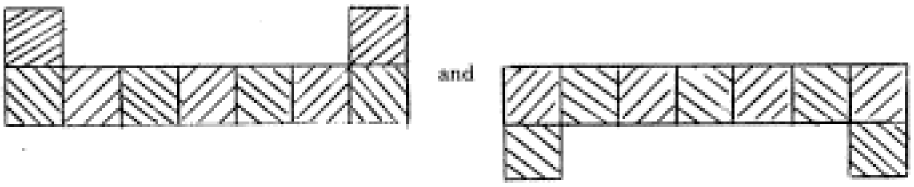
(In fact, in general we can always find a function of order k that recognizes the sets satisfied by any k inequalities concerning their cardinality.) Now consider only the group of translations and masks of order 2. Then two masks x_1x_2 and $x_1'x_2'$ are equivalent if and only if the difference *vectors*

$$x_1 - x_2 \quad \text{and} \quad x_1' - x_2'$$

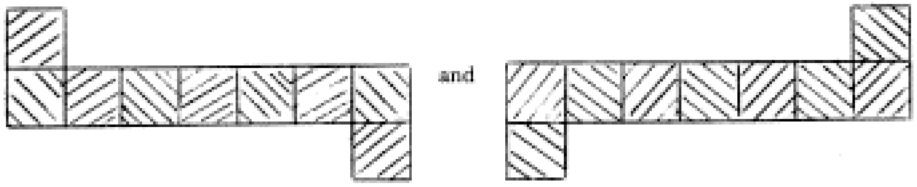
are equal. Then, with respect to the translation group, a figure is completely characterized (up to $k = 2$) by its "difference-vector spectrum," defined as the sequence of the numbers of pairs of points separated by each possible directed distance. The two figures:



have the same difference-vector spectra, hence no order-2 predicate can make a classification which is both translation invariant and separates these two figures. Similarly,

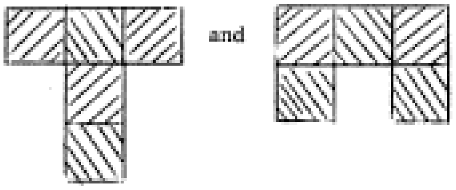


are indistinguishable, while



have different difference-vector spectra.

If we add the requirement of invariance under rotation, the last pair above becomes indistinguishable, for the spectra now relevant classify together all differences of the same length, whatever their orientation.[3] An interesting pair of figures rotationally distinct, but still indistinguishable, for $k = 2$, is the pair



which have the same direction-independent distance-between-point-pair statistics. There is an interesting theoretical direction here, but we will not stop to look into it. Many interesting proposals for pattern recognition machines are related to the theory of these geometric spectra. The classic paper of Bledsoe and Browning [1] is related to this, as is the work on "integral geometry" of Novikoff [4].

(c) $k = 3$. As k increases, the class of realizable discriminations grows,

[3] Note that we did *not* allow reflections, yet these reflectionally opposite figures are now confused! One should be cautious about using "intuition" here. The theory of rotational invariance requires careful attention to the effect of the discrete retinal approximation, but can presumably be made consistent by application of Zeeman's methods; for the dilatation "group," there are serious difficulties.

and our detailed understanding wanes. It is interesting to discover that the predicate

$$\psi(X) = \lceil X \text{ is a single, solid, convex figure} \rceil$$

is of order $\leqq 3$, as noted in the Introduction, because

$$\psi_{\text{CONVEX}}(X) = \lceil \sum_{\text{all } a,b} \lceil a \in X \text{ and } b \in X \text{ and midpoint } (a,b) \notin X \rceil < 1 \rceil$$

is of order 3. Presumably this predicate cannot be realized with order 2. It is not difficult to show that the set of solid rectangles (with axis parallel to the mesh of R) can be recognized by a predicate of order 3. This is true also for the set of hollow rectangles (with borders one square thick). It is much more difficult to show, but true, that the set of hollow *squares* has order three! Intuitively one might suppose that at least order 4 is required to insure equality of side lengths.

Another example of a predicate that can be realized with $k = 3$, for any n, is

$$\lceil \text{the points of } X \text{ are collinear, and broken into not more than } n \text{ segments} \rceil.$$

(d) $k = 4$. Using the fact that any three points determine a circle, we can make a perceptron with masks of order $k = 4$ for the following predicates:

$$\psi(X) = \lceil X \text{ is the perimeter of a complete circle} \rceil.$$

PROOF.[4] Define, for all concyclic quadruples of points in R; a, b, c, d

$$\phi_{abcd}(X) = \lceil a \in X \text{ and } b \in X \text{ and } c \in X \text{ and } d \notin X \rceil$$

and then realize ψ as

$$\psi = \lceil \sum_{a,b,c,d} \phi_{abcd} < 1 \rceil.$$

Many other curious and interesting predicates can be shown by similar arguments to have small orders. One should be careful not to conclude that this means that there are practical consequences of this, unless one is prepared to face the fact that

(a) large numbers of ϕ's are required, of the order of R^{k-1} for the examples given above.

(b) the threshold conditions are sharp, so that engineering considerations may cause difficulties in realizing the linear summation, especially if there is any problem of noise. Even with simple square-root noise, for $k = 3$ or larger, the noise grows faster than the retinal size.

[4] An alternative method is to integrate the curvature of line elements. This leads to interesting questions about the precision of global functions that can be approximated by summation of local elements with a given precision. Curvature requires order 4, in a sense. The predicate defined here admits a few uninteresting exceptions.

(c) a very slight change in the pattern-definition destroys the recognizability.

Furthermore, in most cases there will be more efficient machines, for the same amount of hardware, to realize these rather simply-defined patterns. Low-order recognition has often the character of a "trick," and one cannot generalize freely. The AND-OR order theorem tells us that some simple relations between simple properties of figures can be prohibitively hard to recognize.

VI. **Connectivity: A geometric property with unbounded order.** We define *connectedness* as follows:

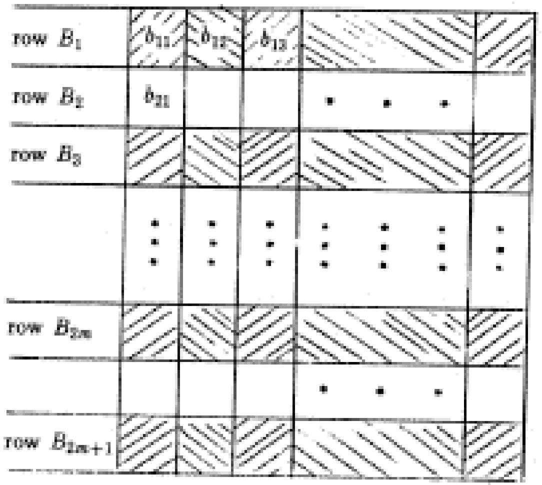
Two points of R are *adjacent* if they are squares (in the map $F \rightarrow \bar{F}$) with a common edge. A figure is connected if, given any two points P_1, P_2 of the figure, we can find a path through adjacent squares from P_1 to P_2.

Theorem. *The predicate*

$$\psi(X) = \ulcorner X \textit{ is connected} \urcorner$$

has arbitrarily large orders as $|R|$ *grows in size.*

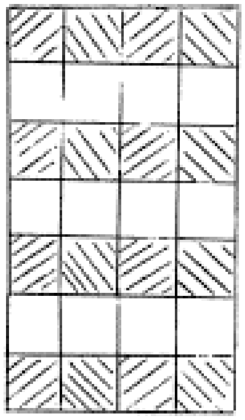
Proof. Suppose that (X) could have order $< m$. Consider an array of $(2m+1) \times 4m^2$ adjacent squares of R arranged in $2m+1$ rows of $4m^2$ squares each. Let G_0 be the set of points shaded in the diagram below;



i.e., the array points whose row indices are odd, and let G_1 be the remaining squares of the array. Let $\mathscr{F}$ be the family of figures obtained from the figure G_0 by adding subsets of G_1. It is clear that if $F \in \mathscr{F}$ it is of the form $G_0 \vee F_1$, where $F_1 \subset G_1$. Now F will be connected if and only if its F_1 contains at least one square from each even row; that is, if the set F_1 satisfies the "one-in-a-box" condition (see end of §3). The theorem then follows from the One-in-a-Box Theorem.

To see the details of how the One-in-a-Box Theorem is applied, if it is not already clear, consider the figures of family $\mathscr{F}$ as a subset of all possible figures on R. Clearly, if we had an order-k predicate that could recognize connectivity on R, we could have one that worked on $\mathscr{F}$; namely the same predicate with constant zero inputs to all variables not in the small array. And since all points of the odd rows have always value 1 for figures in $\mathscr{F}$, this in turn means that we could have an order-k predicate to decide the one-in-a-box property on set G_1; namely the same predicate further restricted to having constant one inputs to the points in G_0. Thus each Boolean function of the original predicate is replaced by the function obtained by fixing some of its variables to zero and one; this operation can never increase the order of a function. But since this last predicate cannot exist, neither can the first.

An Example. Consider the special case for $k = 2$, and the equivalent one-in-a-box problem for a G_1-space of the form



in which $m = 3$ and there are just 4 squares in each row. Now consider a ψ of degree 2; we will show that it cannot characterize the connectedness of pictures of this kind. Suppose that $\psi = \lceil \sum \alpha_i \phi_i > \theta \rceil$ and consider the equivalent form, symmetrized under the full group of permutations that interchange the rows *and* permute within rows.[5] Then there are just three equivalence-classes of masks of degree ≤ 2, namely:

single points: $\phi_i^1 = x_i$,

point-pairs: $\phi_{ij}^{11} = x_i x_j$ (x_i and x_j in same row),

point-pairs: $\phi_{ij}^{12} = x_i x_j$ (x_i and x_j in different rows);

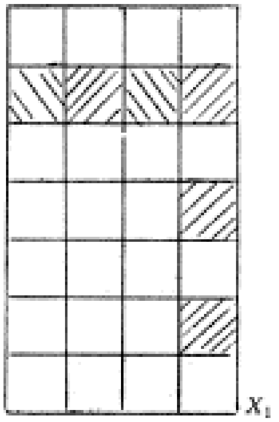
[5] Note that this is not the same group used in proving the general theorem.

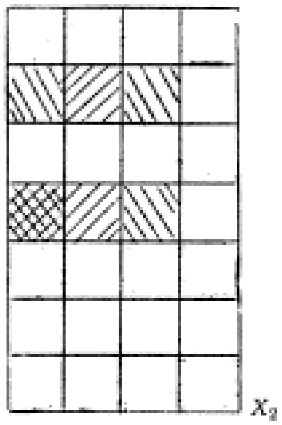
hence any order-2 predicate must have the form

$$(1) \qquad \psi = \alpha_1 N^1(X) + \alpha_{11} N^{11}(X) + \alpha_{12} N^{12}(X) > \theta$$

where N^1, N^{11}, and N^{12} are the numbers of point sets of the respective types in the figure X.

Now consider the two figures:





In each case one counts:

$$N^1 = 6, \qquad N^{11} = 6, \qquad N^{12} = 9;$$

hence the form (1) has the same value for both figures. But X_1 is connected while X_2 is not! Note that here $m = 3$ so that we obtain a contradiction with $|A_i| = 4$, while the general proof required $|A_i| = 4m^2 = 36$. It is known also that if $k = 6$, we can get a similar result with $|A_i| = 16$.

The case of $k = 2$, $m = 3$, $|A_i| = 3$ *is* of order 2, since one can in fact express the connectivity predicate for that space as

$$\psi = \ulcorner N^1(X) + N^{12}(X) - 2N^{11}(X) > 4 \urcorner.$$

Cut-wise Connectivity. It should be observed that the proof of the previous theorem applies only to a property of connectivity in its classical sense but to the stronger predicate defined by:

A figure X is "cutwise disconnected" if there is a *straight line* L such that:

F does not intersect L and does not lie entirely to one side of L.

The general connectivity definition would have "curve" for L instead of "straight line," and one would expect that this would require a higher order for its realization.

Relations Between Perceptrons. The study of the order of predicates is often facilitated by the reduction of a given predicate to another simpler one. Although we do not have a satisfactory theory of any class of reductions, or even a clear enough insight into the nature of the relations which might play a role analogous to "homomorphism," "quotient" and so on in more developed areas of mathematics, the following examples are useful in particular applications and indicate an interesting area for future research.

(a) Let us say that a perceptron system, P, is defined by the basic set R and a set Φ of predicates on subsets of R. A second perceptron system, P', is a subperceptron system of P if the basic set R' is a subset of R and if its set of predicates Φ' is that obtained by relativising the members of Φ to R', i.e., all predicates $\phi' \in \Phi'$ satisfy

$$X \subset R' \Rightarrow \phi'(X) = \phi(X) \quad \text{for some} \quad \phi \in \Phi$$

and all predicates ϕ' satisfying this condition are in Φ'. Clearly the order of any predicate of the form ψ' for P' is at most that of ψ for P.

(b) Isomorphism must be given the following natural sense: Let P be defined by R and Φ and P' by R' and Φ'. Then an isomorphism, f, is an isomorphic map f: $R \rightarrow R'$ of the sets R with the property that for each $\phi \in \Phi$ there is exactly one $\phi' \in \Phi$ satisfying $\phi(x) = \phi'(f(x))$ (where $f(x) = \{p \in R' \mid \exists q \in R; f(a) = p\}$).

(c) P' is obtained from P by a *collapsing operation* f, if f is a map from points of R' to disjoint sets of R, i.e.,

$$p \in R' \Rightarrow f(p) \subset R,$$
$$p \neq q \Rightarrow f(p) \cap f(a) = \Lambda.$$

A predicate ψ' on R' is obtained from a predicate ψ on R by the collapsing map f if $\psi'(X') = \psi(f(X'))$, for $x' \subset R'$.

THEOREM (COLLAPSING THEOREM). *If f is a collapsing map from R to R' and ψ' is obtained from a predicate ψ by f, then the order of ψ' is not greater than that of ψ.*

PROOF. Let $\psi = \lceil \sum_\Phi \alpha_\phi \phi > 0 \rceil$ where Φ is the set of masks of degree less than k on R. Now for any $X' \subset R'$,

$$\begin{aligned} \psi'(X') &= \psi(f(X')) \\ &= \lceil \sum_{\Phi} \alpha_\phi \phi(f(x')) > 0 \rceil. \end{aligned} \tag{1}$$

We next observe that (1) remains true if Φ is replaced by the set $\dot{\Phi}$ of masks ϕ for which $s(\phi) \subset f(R')$, for if

$$s(\phi) \not\subset f(R'), \qquad \phi(f(X')) = 0 \quad \text{for all } X' \subset R'.$$

Now for $\phi \in \bar{\Phi}$ we have

$$s(\phi) \subset \bigcup \{ f(p) \mid p \in R' \},$$

in fact

$$s(\phi) \subset \bigcup \{ f(p) \mid f(p) \cap s(\phi) \neq \Lambda \}.$$

Thus,

$$\begin{aligned} X' &\supset \{ p \mid f(p) \cap s(\phi) \neq \Lambda \} \\ &\Rightarrow f(X') \supset \bigcup \{ f(p) \mid f(p) \cap s(\phi) \neq \Lambda \} \supset s(\phi) \end{aligned}$$

i.e., $X' \supset \{ p \mid f(p) \cap s(\phi) \neq \Lambda \} \Rightarrow f(X') \supset s(\phi) \Rightarrow \phi(f(X'))$. On the other hand, if $\phi(f(X'))$, i.e., $f(X') \supset s(\phi)$, it follows that

$$f(p) \cap s(\phi) \neq \Lambda \Rightarrow p \in X'$$

since $f(p) \cap f(q) = \Lambda$ for $p \neq q$. Thus $\phi(f(x')) = \ulcorner X' \subset \{ p \mid f(p) \cap s(\phi) \neq \Lambda \} \urcorner$. In other words $\phi(f(x'))$ is a mask on R' with support

$$\{ p \mid f(p) \cap s(\phi) \neq \Lambda \}.$$

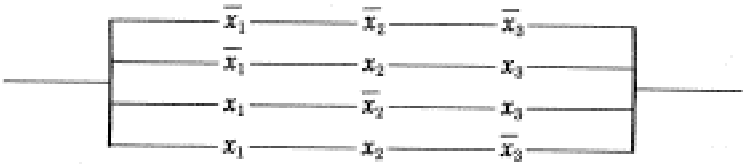
But since the sets of the form $f(p)$ are disjoint, for different p, it follows that

$$|\{ p \mid f(p) \cap s(\phi) \neq \Lambda \}| \leq |s(\phi)| < k.$$

Going back to Equation (1) we see, then, that ψ' is represented as a linear function of masks of degree less than k. Q.E.D.

Huffman's Construction for ψ_{con}. We shall illustrate the application of the preceding concept by giving an alternative proof that ψ_{con} has no finite order, based on a construction suggested to us by D. Huffman.

The intuitive idea is to construct a switching network which will be connected if an even number of its n switches are in the "on" position. Thus the connectedness problem is reduced to the parity problem. The network is shown in the diagram for $n = 3$.


$\overline{x}_1$ $\overline{x}_2$ $\overline{x}_3$
$\overline{x}_1$ x_2 x_3
x_1 $\overline{x}_2$ x_3
x_1 x_2 $\overline{x}_3$

The interpretation of the symbols x_i and $\overline{x}_i$ is as follows: when x_i is in the "on" position contact is made whenever x_i appears, and broken whenever $\overline{x}_i$ appears; when x_i is in the "off" position contact is made where $\overline{x}_i$ appears and broken where x_i appears. It is easy to see that the whole net is connected in the electrical and topological sense if the number of switches in the "on" position is 0 or 2. The generalization to n is obvious:

(a) List the terms in the classical normal form for ψ_{PAR} considered as a point function, which in the case *n even* can be written:

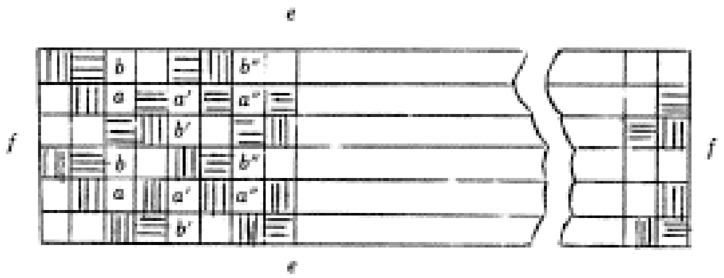
$$\psi_{\mathrm{PAR}}(x_1 \cdots x_n) = \overline{x}_1\overline{x}_2 \cdots \overline{x}_n \vee x_1x_2\overline{x}_3\ \overline{x}_4\ \cdots \overline{x}_n \vee \cdots \vee x_1x_2 \cdots x_n.$$

(b) Translate this Boolean expression into a switching net by interpreting conjunction as series coupling and disjunction as parallel coupling.

(c) Construct a perceptron which "looks at" the position of the switches.

The reductive argument, in intuitive form, is as follows: the Huffman switching net can be regarded as defining a class $\mathscr{F}$ of geometric figures which are connected or not depending on the parity of a certain set, the set switches in "on" position. We thus see how a perceptron for ψ_{con} on one set, R, can be used as a perceptron for ψ_{PAR} on a second set R'. As a perceptron for ψ_{PAR}, it must be of order at least $|R'|$. Thus the order of ψ_{con} must be of order $|R'|$. We shall use the collapsing theorem to formalize this argument. But before doing so we note that a certain price has been paid for its intuitive simplicity: the set R is much bigger than the set R', in fact $|R|$ must be of the order of magnitude of $2^{|R'|}$, so that the best result to be obtained from the construction is that the order of ψ_{con} must increase with $|R|$ like $\log|R|$. This gives a weaker bound, $\log|R|$ compared with $|R|^{1/3}$, if we wish to estimate the order.

Connectivity on a Toroidal Space $|R|$. Our earliest attempts to prove that $\psi_{\mathrm{connected}}$ has unbounded order led to the following curious result: The predicate $\psi_{\mathrm{connected}}$ on an $2n \times 6$ *toroidally* connected space $|R|$ has order $\geq n$. The proof is by construction: consider the space



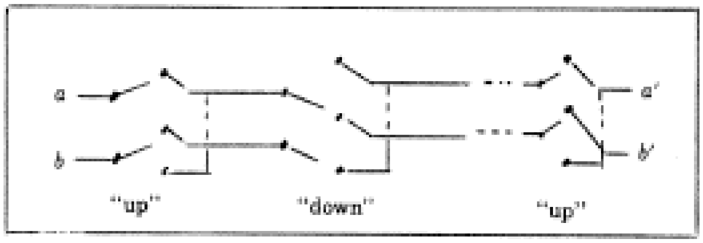
in which the edges e, e and f, f are identified. Consider the family of subsets of R that satisfy the conditions:

(i) All the shaded points belong to each $X \in \mathscr{F}$.

(ii) For each $X \in \mathscr{F}$ and each i, either both points marked $a^{(i)}$ or both points $b^{(i)}$ are in $\mathscr{F}$, but no other combinations are allowed.

Then it can be seen, for each $X \in \mathscr{F}$, that X is either one connected figure or X divides into two separate connected figures. Which case actually occurs depends only on the parity of $|\{i|a^{(i)} \in X\}|$. Then using the Collapsing Theorem and the order $(\psi_{\mathrm{PAR}}) = |R|$ theorem, we find that ψ_{con} has order $|R|/12$.

The idea of this proof came from the attempt to reduce *connectivity* to *parity* directly by representing the switching diagram:



If an even number of switches are in the "down" position then a is connected to a' and b to b'. If the number of down switches is odd, a is connected to b' and a' to b. This diagram can be drawn in the plane by bringing the vertical connections around the end; then one finds that the predicate $\ulcorner a$ is connected to $a' \urcorner$ has for order some constant multiple of $|R|$. If we put the toroidal topology on R, the order becomes $\geqq$ constant times $|R|$; this is also true for a 3-dimensional nontoroidal R. Because of these results, we conclude that the order $\sim |R|^{1/3}$ obtained for $\psi_{\mathrm{connected}}$ is too low.

ADDED IN PROOF: We have since shown that the order is at least $\sim |R|^{1/2}$ in the plane.

Some Other Geometrical Predicates. A number of other important geometric predicates that almost certainly have unbounded orders are:

1. Symmetry: $\ulcorner X$ is a symmetric about some line in the plane.$\urcorner$[6]
2. "Twins": $\ulcorner X$ consists of two disjoint congruent subfigures.$\urcorner$
3. Concentricity: $\ulcorner X$ contains an interior hole.$\urcorner$

Curiously enough, the predicate

$$\ulcorner X \text{ has a single connected comparent} \urcorner \vee \ulcorner X \text{ contains a hole} \urcorner$$

has order 2. This can be shown by a construction using the Euler relation, (Holes = 1 + Edges − Vertices − Faces), even though each separately has unbounded order.

VII. **Connectivity and serial computation.** It seems intuitively clear that the reason that the abstract quality of connectivity cannot be captured by a machine of finite order is that it has an inherently serial character; one cannot conclude that a figure is connected by any simple order-

[6] But, if a *particular* axis line is chosen in advance, then only order 2 is required!

independent combination of simple tests. The same is true for the much simpler property of *parity*. In the case of parity, there is a stark contrast between our "worst possible" result for finite-order machines (§III) and the following "best possible" result for the *serial* computation of parity. Let $x_1, x_2, \cdots, x_n$ be any enumeration of the points of R and consider the following algorithm for determining the parity of $|X|$:

```
START:  set $i$ to 0
EVEN:   add 1 to $i$
        If $i = |R|$ then STOP; parity is EVEN
        If $x_i = 0$, go to EVEN; otherwise go to ODD:
ODD:    add 1 to $i$
        If $i = |R|$ then STOP; parity is ODD
        If $x_i = 0$, go to ODD; otherwise go to EVEN:
```

where "go to α" means continue the algorithm at the instruction whose name is α.

Now this program is "minimal" in two respects: first in the number of computation-steps per point, but more significant, in the fact that the program requires no temporary storage-place for partial information accumulated during the computation, other than that required for the enumeration variable i. (In a sense, the process requires one binary-digit of current information, but this can be absorbed [as above] into the algorithm-structure.)

This suggests that it might be illuminating to ask for connectivity: how much storage is required by the best serial algorithm? The answer, as shown below, is that it requires no more than about 2 times that for storing the enumeration variable alone! To study this problem it seems that the Turing machine framework is the simplest and most natural, because of its simple uniform way of handling information storage.

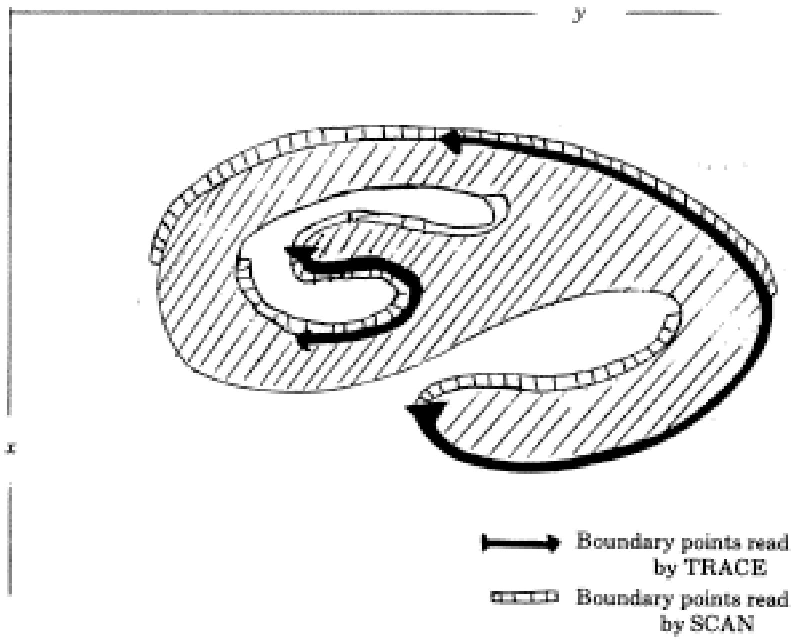
A Serial Algorithm for Connectivity. Connectivity of a geometric figure X is characterized by the fact that between any path (p, q) of points of X there is a path that lies entirely in X. An equivalent definition, using the enumeration $x_1, \cdots, x_{|R|}$ of the points of R is: X is connected if and only if for each point x_i after the first point in X, there is a path to some x_j in X for which $i > j$. (Proof: by recursion, then, each point of X is connected to the *first* point in X.) Using this definition of connectivity we can describe a beautiful algorithm to test whether X is connected. We will consider only figures that are "reasonably regular"—to be precise, we suppose that X is bounded by a number of oriented, simple, closed curves so that for each point x_i on a boundary there is defined a unique "next point" x_{i^*} on that boundary. We choose x_{i^*} to be the boundary point to the left of x_i when facing the complement of X. We will also assume that points x_i and x_{i+1} that are *consecutive* in the enumeration are *adjacent* in R. Finally, we will assume that X does not touch the edges of the space R.

```
START:    Set $i$ to 0 and go to SEARCH
SEARCH:   Add 1 to $i$. If $i = |R|$, Stop and print "$X$ is $NULL$."
          If $x_i \in X$ then go to SCAN, otherwise go to SEARCH.
SCAN:     Add 1 to $i$. If $i = |R|$, Stop and print "$X$ is connected."
          If $x_{i-1} \in X$ or $x_i \notin X$ go to SCAN, otherwise
          Set $j$ to $i$ and go to TRACE.
TRACE:    Set $j$ to $j^*$
          If $j = i$, Stop and print "$X$ is disconnected."
          If $j > i$, go to TRACE.
          If $j < i$, go to SCAN.
```

Notice that at any point in the computation, it is necessary to keep track of the indexes of just the two points x_i and x_j.

Analysis. **SEARCH** simply finds the first point of X in the enumeration of R. Once such a point of X is found, **SCAN** searches through all of R, eventually testing every point of X. The current point, x_i, of **SCAN** is tested as follows: If x_i is not in X, then no test is necessary and **SCAN** goes on to x_{i+1}. If the previous point x_{i-1} was in X (and, by induction, is presumed to have passed the test) then x_i, if in X, is connected to x_{i-1} by adjacency. Finally, if $x_i \in X$ and $x_{i-1} \notin X$, then x_i is on a boundary curve B. **TRACE** circumnavigates this boundary curve. Now if B is a boundary curve it is either (i) an exterior boundary of a previously encountered component of X, in which case some point of B must have been encountered before or (ii) B is an interior boundary curve, in which case a point of B must have been encountered before reaching x_{i-1} which is *inside* B or (iii) B is the exterior boundary curve of a never-before-encountered component of X, the only case in which **TRACE** will return to x_i without meeting an x_j for which $j < i$. Thus **SCAN** will run up to $i = |R|$ if and only if X has a single nonempty connected component.

Note that we can *count* the number of components of X by introducing K, initially zero, and adding 1 to K each time **TRACE** reaches the $i = j$ exit. Note also that the algorithm is quite efficient; the only points examined more than once are some of the boundary points, and none of them is examined more than three times (see figure below).



The Turing Machine Version of the Connectivity Algorithm. It is convenient to assume that R is a $2^n \times 2^n$ square array. Let $x_1, \cdots, x_{2^{2n}}$ be an enumeration of the points of R in the order

$$\begin{array}{cccc}
1 & 2^n+1 & \cdots & (2^n-1)2^n+1 \\
2 & 2^n+2 & \cdots & (2^n-1)2^n+2 \\
\cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdot \\
2^n & 2\cdot 2^n & \cdots & 2^n\cdot 2^n.
\end{array}$$

This choice of dimension and enumeration makes available a simple way to represent the situation to a Turing machine. The Turing machine must be able to specify a point x_i of R, find whether $x_i \in X$, and in case x_i is a boundary point of X, find the index j^* of the "left neighbor" of x_i. The Turing Machine tape will have the form

I_y	$\cdot\cdot\, n \,\cdot\cdot$	I_x	$\cdot\cdot\, n \,\cdot\cdot$	J_y	$\cdot\cdot\, n \,\cdot\cdot$	J_x	$\cdot\cdot\, n \,\cdot\cdot$	K

where "$\cdot\cdot\, n \,\cdot\cdot$" denotes an interval of n blank squares. Then the intervals to the right of I_x and I_y can hold the x and y coordinates of a point of R.

We will suppose that the Turing machine is coupled with the outside world, i.e., the figure X, through an "oracle" that works as follows: certain internal states of the machine have the property that when entered, the resulting next state depends on whether the coordinates in the I (or J) intervals designate a point in X. It can be verified, though the details are tedious, that all the operations described in the algorithm can be performed by a fixed Turing machine that uses no tape squares other than those in "$\cdot \cdot n \cdot \cdot$" intervals. For example, "$i = |R|$" if and only if there are all zeros in the "$\cdot \cdot n \cdot \cdot$"s following I_x and I_y. "Add 1 to i" is equivalent to: "start at J_y and move left, changing 1's to 0's until a 0 is encountered and changed to 1 or until I_y is met. The only nontrivial operation is computing j^* given j. But this requires only examining the neighbors of x_j, and that is done by adding ± 1 to the J_x and J_y coordinates, and consulting the oracle.

Since the Turing machine can keep track of which "$\cdot \cdot n \cdot \cdot$" interval it is in, we really need only one symbol for punctuation, so the Turing machine can be a 3-symbol machine. By using a block encoding, one can use a 2-symbol machine, and, omitting details, we obtain the result:

THEOREM. *For any ϵ there is a 2-symbol Turing machine that can verify the connectivity of a figure X on any rectangular array R, using less than* $(2 + \epsilon) \log_2 |R|$ *squares of tape.*

For *convexity* there is a similar procedure that makes three tests:

i. X is not disconnected by any vertical line that does not intersect X.

ii. The intersection of X with any vertical line is a connected segment.

iii. The outer boundary of X does not change the sign of its curvature.

A detailed construction shows that each test requires only one index point, so that

THEOREM. *For any ϵ there is a 2-symbol Turing machine that can verify the convexity of a figure X on any rectangular array R, using less than* $(1 + \epsilon) \log_2 |R|$ *squares of tape.*

This last result is certainly minimal since $\log_2 R$ squares are needed just to indicate a point of R, and all points must be examined. We are quite sure that the connectivity algorithm is minimal, also, in its use of tape, but we have no proof. In fact, we do not know any method, in general, to show that an algorithm is minimal in storage, except when information-theoretic arguments can be used. Incidentally, it is not hard to show that $\ulcorner |X|$ is prime$\urcorner$ requires no more than $(2 + \epsilon) \log_2 |R|$ squares (and presumably needs more than $(2 - \epsilon) \log_2 |R|$).

We do not definitely know any geometric predicates that require higher orders of storage, but we suspect that in an appropriate sense, the topological equivalence of two figures (e.g., two components of X) requires something more like $|R|$ than like $\log|R|$ squares. There are, of course, recursive function-theoretic predicates that require arbitrarily high, indeed non-computable, orders of storage, but none of these is known to have straightforward geometric interpretations.

VIII. **Multi-layer perceptrons.** We have found a number of limitations of perceptrons, as defined above, and we have suggested that these may point toward as yet unknown theorems about parallel machines in general. On the other hand one suspects that some, at least, of the results above are not so general, and might not survive minor relaxations of the definitions. One direction of generalization that seems important is that of relaxing the constraint that the ϕ's be simply weighted and added. We have not found any particularly enlightening generalization on the lowest level—e.g., of replacing addition by an arbitrary commutative operation. An easier direction is to consider compositions of perceptrons. The remainder of this section explores some kinds of composite perceptrons. Unfortunately we do not understand them very well, so this section is more concerned with problem-posing than with problem-solving.

Gamba Machines. Consider functions of the form

$$\ulcorner \sum_j \beta_j \ulcorner \sum_i \alpha_{ij} x_i > \theta_j \urcorner > \theta \urcorner .$$

This form was proposed and realized in a series of machines built by A. Gamba [2]. It is essentially an order-1 composition of order-1 perceptrons, and is of interest to us for a number of reasons:

(i) The parity problem is solved neatly by

$$\ulcorner \sum_{j=0}^{|R|} (-1)^j \ulcorner \sum_{\text{all } i} x_i > j \urcorner > 0 \urcorner .$$

Thus only $|R|$ functions are needed, each itself of order 1 in the $\{x_i\}$. In fact, any predicate $f(|X|)$ that depends only on the area $|X|$ can be realized, as

$$f(|X|) = f(0) + \sum_{j=0}^{|R|} \ulcorner \sum x_i > j \urcorner \cdot (f(j+1) - f(j)).$$

The problem that led to our formulation of the AND-OR theorem also is solved neatly:

$$\ulcorner \ulcorner \sum_{x \in A} x - \sum_{x \in C} x > 0 \urcorner + \ulcorner \sum_{x \in B} x - \sum_{x \in C} x > 0 \urcorner \geq 2 \urcorner$$

is 1 if and only if $|X \cap A| > |X \cap C|$ and $|X \cap B| > |X \cap C|$. This might suggest that this class of machines might transcend the other kinds of limitations we have found for machines of finite-order.[7]

We are quite certain that this impression is misleading; that the deeper geometric properties are still outside the reach of this kind of "2-layer" perceptron. The inclusion of AND and OR is due to the 2-layer construction; any Boolean function is obtainable, in such a manner, through its normal form, but for most functions there will still be too many terms for practical interest. The $\lceil |X \cap A| > |X \cap C| \rceil$ type of predicates are within reach because they are simple area functions and hence fit precisely the inner $\sum \alpha_{ij}$ first-order predicate forms. In fact, *any* class $\mathscr{F}$ of figures can be recognized by a Gamba-machine because

$$\lceil \sum_{X \in \mathscr{F}} \lceil \sum_{x \in X} x_i \geq |X| \rceil > 0 \rceil$$

realizes it. But, this general form requires a special "Gamba-mask" for each $X \in \mathscr{F}$. Although the above examples show that in special cases more economical representations are possible, this is not true in general (as one can see by considering the number 2^{2^n} of possible functions). In particular we conjecture, for example, that for the Connectivity Predicate, the machine would require a number of masks of an order approaching the number of simple-closed-curves in R. Even for convexity, we doubt that that predicate can be realized with significantly fewer than the number of ϕ's needed for the order-3 1-layer machine.

(ii) In spite of its apparent simplicity, analysis of the geometric predicate problem for Gamba machines appears to require methods quite different from those we have used. First, because of the arbitrary order-1 predicate permitted in the inner sum, the notion of *order* does not seem to apply, and theorems must concern restrictions on the numbers of terms. Second, we have not found a way to carry the group-averaging methods into the inner α_{ij} coefficients, so that we cannot use the techniques that come from the group-invariance theorem. It is difficult to see how to analyse other multi-layer and composite perceptrons until this simple case is better understood. How much weaker are the machines with $\alpha_{ij} > 0$ or those with all $\theta = 0$? We have no characterization of what they can do.

(iii) The Gamba-machine is of considerable practical interest because of the possibility of realizing the inner, and even the outer, sums by inexpensive, highly parallel optical methods. Using coherent light and properly

[7] Note that the Gamba-machine can have order as large as $|R|$, in the $\{x_i\}$. If the inner predicate threshold were removed then, because

$$\sum_j \theta_j \sum_i \alpha_{ij} x_i = \sum_i \left(\sum_j \theta_j \alpha_{ij} \right) x_i,$$

one would have merely an order-1 function in the $\{x_i\}$.

prepared photographic transparencies, one can realize each inner sum (even with complex coefficients!) with a picture p_j whose density at point x_i is α_{ij}. By shrewd optics, one can even do this (with fixed p_j) for all translations of the source pattern X. Because of these technological possibilities it is important to have a better theory; we expect that the result will be favorable to problems like recognition of printed characters, but still very poor for the more abstract properties like detection of connectivity, symmetry, topological equivalence, and the like.

IX. **The diameter-limited perceptron.** In this section we discuss the power and limitations of the "diameter-limited" perceptrons: those in which each ϕ can see only a circumscribed portion of the retina R.

We consider a machine that sums the weighted evidence about a picture obtained by experiments ϕ_i each of which report on the state of affairs within a circumscribed region r_i of *diameter less than or equal to some length* D. That is, Diameter $(S(\phi)) < D$. We will suppose that D is uniform over the ϕ's of the machine (each actual region that affects a ϕ_i can be smaller, but not larger). We suppose also that in a practical sense D is small compared with the full dimensions of the space R. That is, D should be small enough that none of the ϕ's can see the whole of an interesting figure (or else we would not have an effective limited-diameter situation, and there would be no interesting theory) but D should be large enough that a ϕ_i has a chance to detect an interesting "local feature" of the figure.

We will consider first some things that a diameter-limited perceptron can recognize, and then some of the things it cannot.

(a) *Blank picture, or black picture.* A diameter-limited perceptron can tell when a picture is entirely black, or entirely white: suppose that the set of ϕ_i's is chosen to *cover* the retina in regions, that may overlap, and that we define ϕ_i to be zero when all the points it can see are white, otherwise its value is 1. Then $\sum \phi_i > 0$ if the picture has one or more black points, and not if the picture is blank. Similarly, we could define the ϕ_i's to be 1 when they see any white point, 0 otherwise, thus distinguishing the all-black picture from all others.

For later examples, it is important here to notice why these patterns can be recognized: it is not that any ϕ-unit can really say that there is strong evidence that the figure is all-white (although it has a slight correlation with this); but any ϕ can definitely say that it has conclusive evidence that the picture is *not* all white. Some interesting patterns have this character; that one can *reject* all pictures not in the class because each must have, somewhere or other, a local feature that is definitive and can be detected by what happens within a region of diameter D.

(b) *Area cuts.* We can distinguish, for any number S, the class of figures whose area is greater than S. To do this we define a ϕ_i for each point to be 1 if that point is black, 0 otherwise. Then $\sum x_i > S$ is a recognizer for the class in question. (One can do slightly better; if the ϕ's look at regions of area A, then one can recognize this pattern by using only of the order of $(R \log A)/A$ units.)

(c) *Nonintersecting lines*. One can say that a pattern is composed of nonintersecting lines if, in each small region, the pattern is composed of separate line-segments, or blank. Then, if we make each ϕ have value zero when this condition is met, unity when it is not, then $\sum \phi_i > 0$ will reject all figures not in the class.

(d) *Triangles*. We can make a diameter-limited perceptron recognize the figures consisting of exactly one triangle (either solid or outline) by the following trick: We use two kinds of ϕ's: the first has weight $+1$ if its field contains a vertex (two line segments meeting at an angle), otherwise its value is zero. The second kind, ϕ_i^*, has value zero if its field is blank, or contains a line segment, solid black area, or a vertex, but has value $+1$ if the field contains anything else, including the end of a line segment. Provide enough of these ϕ's so that the entire retina is covered, in nonoverlapping fashion, by both types. Finally assign weight 1 to the first type and a very large positive weight W to those of the second type. Then

$$\sum \phi_i - W \sum \phi_i^* < 4$$

will be a specific recognizer for triangles. (But also the null-picture is accepted.) Similarly, by requiring that the first kind of unit recognize right-angle vertices, the machine can be made to recognize the class of rectangles (setting the threshold to be < 5).

Note that this does not generalize to a very wide class of geometric recognition abilities. The triangle and rectangle cases are rather peculiar; the triangle because it is the simplest figure that has true vertices. The rectangle can be recognized because it has four equal angles; the system cannot be specialized to recognize, for example, exactly the squares. It is interesting, in view of the limitations we will establish shortly, to see why these patterns can be recognized by the diameter-limited machine; a rectangle is the only figure that has four *or fewer* right angles and no free line ends, etc.

(e) *Absolute template-matching*. Suppose that one wants the machine to recognize exactly a certain figure X_0 and no other. Then the diameter-limited machine can be made to do this by partitioning the retina into regions, and in each region a ϕ-function has a value 0 if that part of the retina is exactly matched to the corresponding part of X_0, otherwise the value is 1. Then

$$\sum \phi_i < 1$$

if and only if the picture is exactly X_0.

Note, however, that this scheme works just on a particular object in a particular position. It cannot be generalized to recognize a particular object in any position (or even, in general, in two positions). In fact we show in the next section that even the simplest possible figure, namely one that consists of just one point, cannot be recognized independently of position!

(f) *Convexity*. The remarks in §5, Example c, footnote apply to the diameter-limited case.

Limitations of Diameter-limited Perceptron. Now we consider some of the basic limitations of the diameter-limited perceptron, by exhibiting and analysing some patterns they cannot recognize.

(g) *The figure containing one single black point*. This is the fundamental counter-example. We want a machine

$$\sum \alpha_i \phi_i > 0$$

to accept figures with area 1, but reject figures with area 0 or area greater than 1. Clearly this can be defined by two area cuts (i.e., area > 0 AND area < 2), but it cannot be realized by a linear threshold function with the area-restriction.

To see that this cannot be done, suppose that $\{\phi_i\}$, $\{\alpha_i\}$ and θ have been selected. Present first the blank picture, X_0. Then, defining $f(X) = \sum \alpha_i \phi_i(x)$ we have $f(X_1) < \theta$. Now present a figure, X_1, containing only one point, x_1. We must then have

$$f(X_1) \geqq \theta.$$

The change in the sum must be due to a change in the values of some of the ϕ's. In fact, it must be due to changes only in ϕ's for which $x \in S(\phi)$, since nothing else in the picture has changed. In any case,

$$f(X_1) - f(X_0) > 0. \tag{1}$$

Now choose another point x_2 which is farther than D away from x_1. Then no $S(\phi)$ can contain both x_1 and x_2. For the figure X_2 containing only x_2 we must also have

$$f(X_2) = \sum \alpha_i \phi_i \geqq \theta. \tag{2}$$

Now consider the figure X_{12} containing both x_1 and x_2. The addition, to X_2, of the point x_1 can affect only ϕ's for which $x \in S(\phi)$, and these are changed exactly as they are changed when the all-blank picture X_0 is changed to the picture X_1. Therefore

$$f(X_{12}) = f(X_2) + [f(X_1) - f(X_0)]$$

and by (1) and (2),

$$f(X_{12}) > \theta,$$

but we require that

$$f(X_{12}) \leqq 0.$$

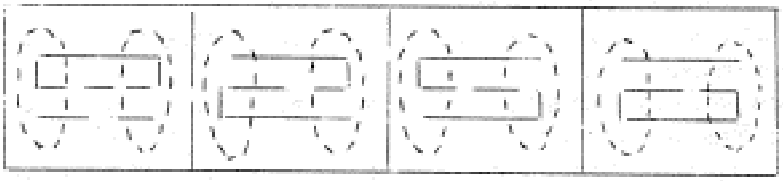
REMARK. Of course, this is the same phenomenon noted in the introduction to §II.

(h) *Area segments.* The diameter-limited perceptron cannot recognize the class of figures whose areas A lie between two bounds $A_1 \leqq A \leqq A_2$.

PROOF. This follows from the method of (a) above, which is a special case of this, with $A_1 = 1$ and $A_2 = 1$. But using the method of §I, Example (vii), this recognition is possible with order 2 if the diameter-limitation is relaxed.

(i) *Connectedness.* The diameter-limited perceptron cannot decide when the picture is a single, connected whole, as distinguished from two or more disconnected pieces.

PROOF. Consider the four pictures



and suppose that the diameter D is of the order indicated by the dotted circle. Now figures X_{01} and X_{10} are connected, but X_{00} and X_{11} are disconnected. Suppose that there were a set of ϕ's and α's and a such that

$$\sum \alpha_i \phi_i(X_{00}) < \theta, \qquad \begin{array}{c} \sum \alpha_i \phi_i(X_{01}) \geqq \theta, \\ \sum \alpha_i \phi_i(X_{10}) \geqq \theta, \end{array} \qquad \sum \alpha_i \phi_i(X_{11}) < \theta$$

so that these four figures were correctly separated. But then, just as in the previous argument we would have for all ϕ_i,

$$\phi_i(X_{11}) = \phi_i(X_{10}) + \phi_i(X_{01}) - \phi_i(X_{00})$$

because the two changing regions are more than D apart, hence

$$\sum \alpha_i \phi_i(X_{11}) \geqq \theta + \theta - \theta = \theta$$

contradicting the separation requirement.

Acknowledgement: We would like to thank many students and associates, particularly William Henneman, Dona Strauss and John White, for their contributions to this work.

BIBLIOGRAPHY

1. W. W. Bledsoe and I. Browning, *Pattern recognition and reading by machine*, Proc. Eastern Joint Computer Conference, 1959; reprint, *Pattern recognition*, Uhr, 1966.

2. A. Gamba, *Optimum performance of learning machines*, Proc. I.R.E. **49** (1961), 349; *Further experiments with PAPA*, Nuovo Cimento Suppl. Ser. X **20** (1961), 112-115.

3. N. Nilsson, *Learning machines*, McGraw-Hill, New York, 1965.

4. A. B. J. Novikoff, *Integral geometry as a tool in pattern perception*, Principles of self-organization, Pergamon, New York, 1961.

5. W. Pitts and W. S. McCulloch, *How we know universals*, Bull. Math. Biophys. **9** (1943), 127-147ff reprinted in *Embodiments of mind*, M.I.T. Press, Cambridge, Mass., 1965.

6. F. Rosenblatt, *Principles of neurodynamics*, Spartan, Washington, D.C., 1962.

7. C. J. Zeeman, "Topology of the brain" in *Topology of 3-manifolds*, M. K. Fort, ed., Prentice-Hall, Englewood Cliffs, N. J., 1961.

8. M. Minsky and O. G. Selfridge, *Learning in random nets*, Proc. London Information Theory Sympos., Butterworth, London, 1961.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS