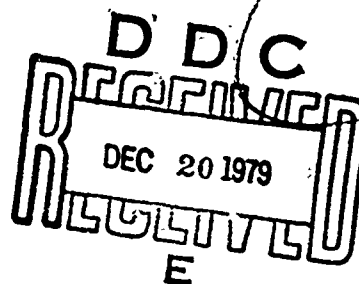


AD A078863

LEVEL



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence
Memo. No. 140.

(14) MAC-M-358,
AI-M-148

(6) Perceptrons and Pattern Recognition
by

(10) Marvin L. Minsky and Seymour Papert

(11) Sep 67

(12) 113

This monograph includes most of the results to date of our analysis of the geometric ability of linear separation machines. We expect soon to publish the complete work as a book: the complete version will contain further results on learning and on some relation between parallel and serial algorithms. To include

The authors invite correspondence and discussion, from the global theorem level to the local misprint level, in the hope of excising as much falsehood as possible from the final version.

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

Work reported herein was supported (in part) by Project MAC, an M.I.T. research program sponsored by the Advanced Research Projects Agency, Department of Defense, under Office of Naval Research Contract Number Nonr-4102 (01) - (02)

(15)

THE RUTH H. HOOKER
TECHNICAL LIBRARY

SEP 11 1979
MAC-M-358
September 1967

NAVAL RESEARCH LABORATORY

THE RUTH H. HOOKER
TECHNICAL LIBRARY

SEP 13 1979

NAVAL RESEARCH LABORATORY

DDC FILE COPY

79 12 18 12

407 018

LB

CONTENTS

0 Introduction

PART I: Algebraic Theory --

1 Theory of Boolean Linear Separation Functions,

2 Group Theory of Linear Predicates,

3 Some Special Predicates, and

4 The 'And - Or' Theorem ;

PART II: Geometric Theory --

5 Connectivity and Topological Properties,

6 Geometric Patterns of Low Order,

7 Normalization and Stratification, and

8 The Diameter - Limited Perceptron; and

Part III: Other Topics

9 Magnitude of the Coefficients,

10 Learning

11 Serial Algorithms

Accession For	
NWIS G.A.M.I	☒
DDC TAB	☐
Unpublished	☐
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Available/or special
A	

REFERENCES

Although there is a vast literature of publications on the subjects of linear threshold functions, perceptron experiments, and statistical theory concerning pattern recognition, we have found almost no references that are concerned with the problem of characterizing the patterns that can be recognized under different constraints on the machine's connections.

Those who want to work in this area could begin with the following papers:

Bledsoe, W. W. and Browning, I., "Pattern Recognition and Reading by Machine," Proc. 1959 Eastern Joint Computer Conference, pp. 225--232.
Practical experiments with finite-order conjunctively local predicates.

Dertouzos, M., Threshold Logic: a synthesis approach, MIT Press, 1965.
Theory of synthesis of low-order predicates.

Nilsson, N. J., Learning Machines: Foundations of trainable Pattern Classifying Systems, McGraw-Hill, 1965.
Review of the field of linear-separation learning systems:
much easier to read than the previous literature.

Pitts, W. and W. S. McCulloch, "How We See Universals," Bull. Math. Biophysics 9, 1947, pp. 127--147.
Discusses recognition of group-invariant visual patterns.

Rosenblatt, F., Principles of Neurodynamics, Spartan Books, 1962.
Introduces and discusses many parallel-network machines, with some biological speculation.

Acknowledgments

The development of this theory owes much to a number of places and people. The places include mountains in the Alps, Sierras and Rockies, swamps and beaches of California, Florida and Puerto Rico, and other locations far enough away from "work" to encourage undistracted thinking.

Among the people are Terry Beyer, Woodrow W. Bledsoe and Dona Strauss, whose influences affect much of the global style and orientation of the paper. More local contributions on particular results and methods are due to Manuel Blum, William Henneman, David Huffman, John White, and a number of other colleagues and students.

And most local of all we thank Lucy Sloan for untangling innumerable drafts and revisions of the manuscript.

CHAPTER 0

INTRODUCTION

0.1 General Introduction

The context of this study was the desire for a better understanding of a set of concepts we believe important for the theory of computation. Distinctions like "serial vs. parallel computation," "local vs. global properties," "addressed vs. associative memory," "iterative vs. recursive algorithms," are frequently used to refer to these concepts, often as if they were well-defined technical terms used with substantial knowledge about the conditions under which these forms of computation would be advantageous. But despite their wide currency in an intuitive form, they have not as yet received any satisfactory formal definitions, nor are they at all well understood even in their intuitive forms.

We felt that our inability to formulate satisfactory definitions was due mainly to the unavailability of thoroughly analyzed special cases that could serve as models for thinking about the broader issues. Good theories develop rarely outside of the context of well-understood real problems, and it is perhaps not surprising that work directed sharply toward obtaining an "abstract theory of computation"--e.g., the mathematical developments in current theories of recursive function, automata, formal linguistics, and the like--has been disappointing in the extent of its practical illumination, despite its often elegant mathematical quality. Accordingly, we have become engaged in a number of attempts to clarify the nature of computations in some problems of independent interest. In the present study we explore the properties of the

simplest class of automata we know that have no loops or feedback but are nevertheless capable of some non-trivial computations. Fortunately they are also rich enough to be the object of an interesting mathematical theory.

The mathematical results described herein permit analysis, to a certain level, of the range and limitations of a class of computing machine that have been widely investigated, by empirical methods, for possible use on problems of pattern recognition.

The characteristic feature of these machines is that they make their decisions--about whether or not a certain event fits a certain "pattern"--by "adding up" evidence obtained from many separate small experiments. This is a very important concept because it is so clear and simple. Most, and perhaps all, more complicated machines for making decisions will have a little of this character. In any case, until we understand this simple concept very thoroughly, we certainly can expect trouble with more advanced ideas. Generally, in Science and Mathematics, one advances by understanding first the "linear" systems, and these machines are our candidate for the "linear case of the parallel machine in general." We will bring forward a number of arguments, at various points in the text to support this view (which is a methodological position rather than a technical matter).

These devices, defined below in §0.3 and §1.2 are most fittingly known as perceptrons in recognition of Rosenblatt's contributions toward formulating mathematically clear definitions. Under this name, the machines have been widely investigated, but with generally inconclusive and puzzling results. The empirical tests have been, by and large, unconstrained by theoretical analyses of the machines' limitations: such analysis as was attempted was

committed chiefly along certain statistical directions that failed to shed much light on the relation between the structures of the "patterns" and the ability of perceptrons to "recognize" those patterns.

Our theory does not completely characterize these limits. For, any such theory must mediate between some a priori classification of the patterns themselves, and their recognition by perceptrons. It would be too much to ask for an absolute classification of "patterns," for this depends ultimately on what is one's goal, i.e., what one is interested in. We have chosen to study some patterns that are definable in terms of familiar geometric concepts, for these are both of great practical interest, and are profoundly well understood mathematically. For each class of geometric patterns, we have to attack the problem of what conditions must be met if a perceptron is to make an appropriate recognition. To do this we need to develop analytic tools, and often new ones for each new problem.

Our experience has been that such problems are by no means trivial. Some of them baffled us for a long time before we found suitable analytic concepts for treating them. Some of them led to solutions quite the opposite of our intuitive expectation. Above all, we were repeatedly surprised at the curious, and various, mathematical paths we were led--or rather, forced--along. We have made some attempt to leave traces of these paths (thus running against today's mathematical style of covering completely one's intellectual tracks) and we hope the reader will try to share this by reading the book -- re as a novel in which characters develop and interact, than as a sequence of theorems and proofs.

0.2 Local Properties

One of the most powerful themes in cybernetic discussions of pattern recognition is related to the discovery that seemingly complex "patterns" can often be characterized or generated by "local" processes*. The following examples are meant to indicate the meaning of the quoted words and to show why it is difficult to find a formal definition.

Let R be a region in the ordinary two-dimensional Euclidean plane. Let X be a figure drawn on R , e.g., a circle or a pair of circles or a black and white sketch of a man's face. In general we think of X as merely a subset of the points R . When we talk of a "pattern" we usually have in mind some class of figures, e.g., all circles, or all connected figures, or all smiling faces. We shall discuss a number of kinds of algorithms that examine figures to decide whether they belong to a given class.

To talk about these algorithms we will have to introduce some auxiliary concepts. We adopt the word "predicate" for any function $\phi(X)$ which has the value 1 for some figures, and the value 0 for all other figures. In general, to compute $\phi(X)$ we must look at every point of R to check whether it is in X or not. In some special cases $\phi(X)$ can be computed by looking only at a proper subset of points R . In any case we call the required subset of R the support of ϕ . The simplest kind of predicate looks only at a single point of R : we

* A powerful and productive theme in the study of animal behavior, "ethology," is the explanation of highly selective "recognition" behavior on the basis of multiple sequential selection steps, each relatively simple in itself.

define

$$\begin{aligned}\varphi_p(X) &= 1 \text{ if the point } p \text{ is in } X, \\ &= 0 \text{ if not.}\end{aligned}$$

The support of φ_p is then just the unit set containing the single point p . Given any set $\{p_1 \dots p_n\}$ containing n points there are 2^{2^n} predicates whose supports are subsets* of $\{p_1 \dots p_n\}$, viz., all the Boolean functions of the predicates $p_1 \in X, p_2 \in X, \dots, p_n \in X$. Thus the support of a predicate is simply the minimum set of points on which it really depends.

Predicates of finite support are "local" in a very strong sense, but so strong as to exclude all examples of direct geometric interest. However, we will put this notion to use in defining a weaker but more interesting sense of localness. We will begin with an important geometric predicate, convexity.

The convexity predicate $\downarrow \text{CONVEX}(X)$.

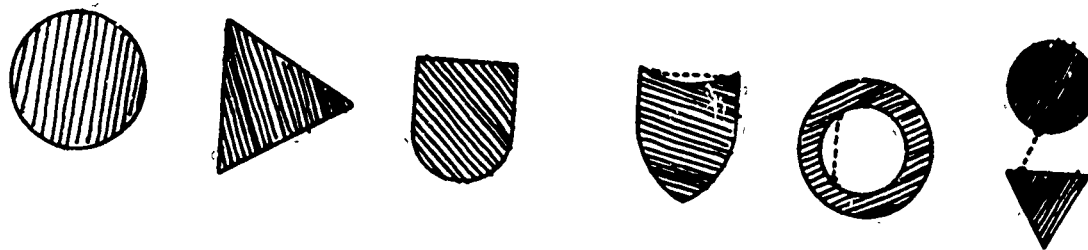


Fig. 0.2-1

We say that a figure X is convex if, given any pair of its points, the line segment between them lies entirely within X . This is true of each

* Not all of the 2^{2^n} have the whole set $\{p_1 \dots p_n\}$ as their support, but most of them do.

figure on the left. Each figure on the right has exceptions, as indicated by the dotted lines. Now we wish to compute the predicate $\psi_{\text{CONVEX}}(X)$ which has value 1 if X is a convex figure of the plane and value 0 if X is not convex. Clearly ψ_{CONVEX} does not have a finite support, for its value can depend on what happens anywhere in the (infinite) plane. We ask instead: is it possible to find a collection of simpler predicates each with small support, together with some simple way to combine them to synthesize ψ_{CONVEX} ?

To be more specific: We shall say that a predicate $\psi(X)$ is conjunctively local if there is a number k and a collection Φ (perhaps infinite) of predicates whose supports each contain no more than k points. Furthermore Φ must have the property that

$$\psi(X) = 1 \text{ if and only if } \phi(X) = 0 \text{ for every } \phi \text{ in } \Phi.$$

We ask whether ψ_{CONVEX} is conjunctively local. (The point is that we are trying to develop ways of building complicated predicates out of simple ones.)

The answer is yes! We can set $k = 3$ and choose some predicates ϕ_{xyz} that depend each on only three points, as follows:

Let x , y , and z be any three distinct points that lie, in that order, along any straight line and define

$$\begin{aligned} \varphi_{xyz}(X) &= 1 \text{ if and only if} \\ &\begin{cases} x \text{ is in } X, \text{ and} \\ z \text{ is in } X, \text{ but} \\ y \text{ is not in } X; \end{cases} \\ &= 0 \text{ in any other case.} \end{aligned}$$

The only way an X can escape having $\varphi(X) = 1$ for some φ is by being convex: otherwise there will exist a line segment (call it $[x,z]$) whose ends lie in X but which does not lie entirely within X . Choose y to be one of the points of $[x,z]$ not in X ; then $\varphi_{xyz} = 1$.

We note in passing that there is a simple formal way to describe $\forall \text{CONVEX}$ now; we can write

$$\forall \text{CONVEX}(X) = 1 \iff \sum_{y \in [x,z]} \varphi_{xyz}(X) < 1,$$

for the sum of any collection of zeros will be zero, while any exceptions will makes the sum at least unity.

Many other geometric predicates are conjunctively local. Another example, discussed also in §0.4, is the predicate

$$\forall \text{CIRCLE}(X) = 1 \iff X \text{ is the perimeter of a complete circle.}$$

Theorem: $\forall \text{CIRCLE}(X)$ is conjunctively local with $k = 4$.*

Proof: The proof is based on the fact that any three points x, y, z , not in a straight line, determine a circle, C_{xyz} .

* But see note below for the degenerate case of 2 or fewer points in X .

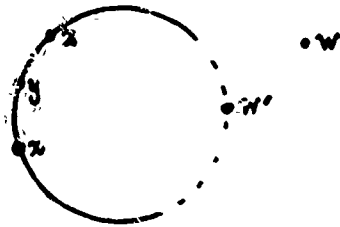


Fig. 0.2-2

Now obviously no predicate of limited support can tell whether the whole figure X is exactly a circle, because some points of X may be outside its support set. But if X is not a circle then at least one of the following two kinds of events must happen:

- i) there are four points x, y, z and w in X which do not lie on the same circle, or
- ii) there are three points x, y , and z in X and one point w' , not in X , which do lie on the same circle.

To see this, choose any 3 points x_0, y_0, z_0 in X . They determine a circle C_0 . If (i) is false for all points w not in C_0 , then all other points of X must lie in C_0 and we can conclude that all of X is contained in a certain circle C_0 . But now if (ii) is false for all points w' in C_0 this means that all of the circle C_0 is contained in X . So X is C_0 . It follows that $\forall$ CIRCLE is conjunctively local "with order r ," i.e., can be described by the simultaneous truth of a lot of predicates each with ≤ 4 support points.

Q.E.D.

NOTE: If X contains only 0 or 1 or 2 points, it will pass the test for circle-ness. There is absolutely nothing that can be done to prevent this! If we are prepared to ignore 0, 1, and 2 point figures, or to consider them "degenerate circles" (which is hard to swallow for the 2-case) then $\forall \text{CIRCLE}$ is conjunctively local. If we must reject the 2-point figures as non-circles, then $\forall \text{CIRCLE}$ is not conjunctively local. To see this, we note first that the only way a non-circular X can escape $\forall \text{CIRCLE}(X)$, as defined in 0.2, is by having fewer than 3 points. We now prove that there is no way to repair this. (This is interesting, not so much because of the theorem, which is unimportant, but because it is a simple example of an impossibility proof.) Suppose that we had a conjunctively local definition for $\forall \text{CIRCLE}(X)$, i.e., a number k and a set φ_α of predicates each of support $\leq k$ for which:

$$X \text{ is a circle} \iff \varphi_\alpha(X) = 0 \text{ for all } \alpha.$$

Then let X consist of two points x_1 and x_2 . Since X isn't a circle there must be some φ_{α_0} for which $\varphi_{\alpha_0} = 1$. Let φ_{α_0} 's support set be $P_1, P_2, P_3, \dots, P_k$, where at least $P_3, \dots, P_k$ are not in X . Then $\varphi_{\alpha_0}(X)$ will be 1 for any other set X' which contains x_1 and x_2 but none of $P_3, \dots, P_k$. But we can always find such an X' which is a circle, because there are an infinite number of circles through x_1 and x_2 , and we have to avoid only a finite set of points.

On the other hand certain properties are essentially not conjunctively local: e.g., the property of being connected. (See Chapter 5.) As we

shall understand later, no minor modification of that property will make it conjunctively local, or even local in the broader sense to be defined in the next section. We are building up to the idea that certain properties are profoundly global in that separate local observations cannot be combined in simple ways to yield conclusive evidence for them.

To obtain the new sense of local let us now try to separate essential from arbitrary features of the definition of conjunctive localness. The intention is clear: to divide the computation of a predicate ψ into two stages:

Stage 1: The computation of many properties or features which are easy to compute either because they depend only on a small subset of the whole input space R , or are very simple in some other interesting way.

Stage 2: A decision algorithm which expresses ψ as a function of the results of Stage 1 computations. For the exercise to be meaningful this decision function must also be particularly homogeneous, or easy to program, or easy to compute.

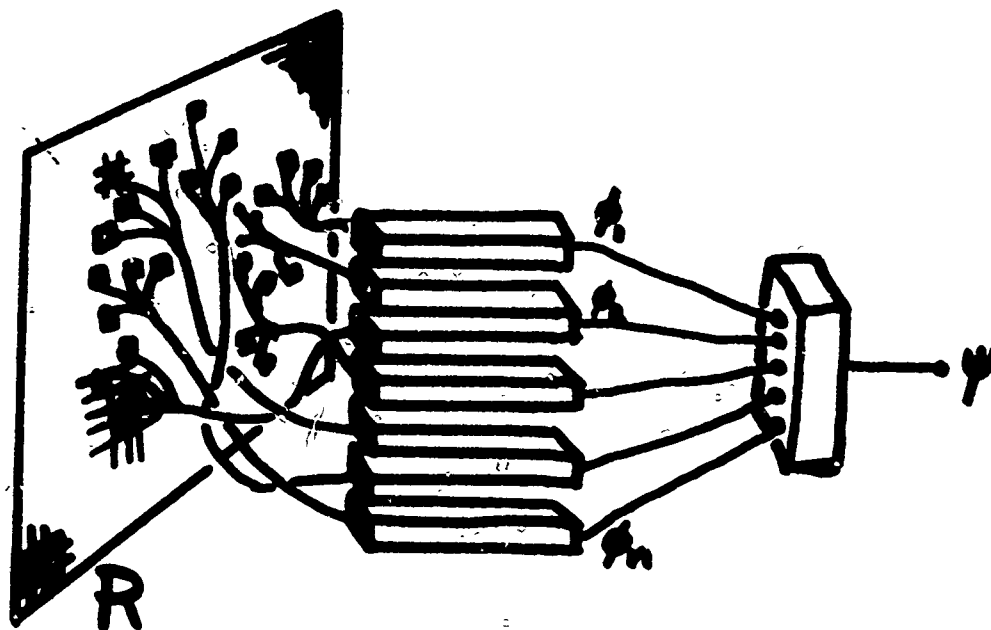


Fig. 0.2-3

The particular way in which this intention was realized in our example is extremely arbitrary. For Stage 1 we might, instead of restricting the number of points in the support sets of the local functions, have restricted, for example, their diameter (as in Chapter 8).

For Stage 2 there are any number of candidates to replace unanimity as the decision criterion, with greater claim to generality and very little loss in computational simplicity. A general theory would have to undertake the difficult task of characterizing the complexity of all possible algorithms. Without such a characterization, the requirement of Stage 2 must retain a heuristic character that makes formal definition difficult.

In this study we shall confine attention to a class of decision functions that includes unanimous decision as a particular case: that is, the definition

of perceptron in the next section can be thought of as derived from the above scheme by replacing unanimity by majority or, more generally, weighted voting.

0.3 Perceptrons--Definition

Let Φ be a set of predicates. We say the predicate ψ is linear in the set Φ if it can be expressed in the form: $\psi(X) = 1$ if and only if

$$\sum_{\varphi \in \Phi} \alpha_{\varphi} \varphi(X) \geq \theta$$

where the "coefficients" α_{φ} and the "threshold" θ are real numbers. The unanimity condition used in the previous section to define conjunctive localness can be expressed in this form by letting $\alpha_{\varphi} = -1$ for all φ , and $\theta = 0$, provided we do not mind the sum becoming infinite. For $\sum \alpha_{\varphi} \varphi(X) = -\sum \varphi(X) \geq 0$ is true exactly when no $\varphi(X) = 1$. So $\psi(X) = 1$ if and only if $\varphi(X) = 0$ for all $\varphi \in \Phi$.

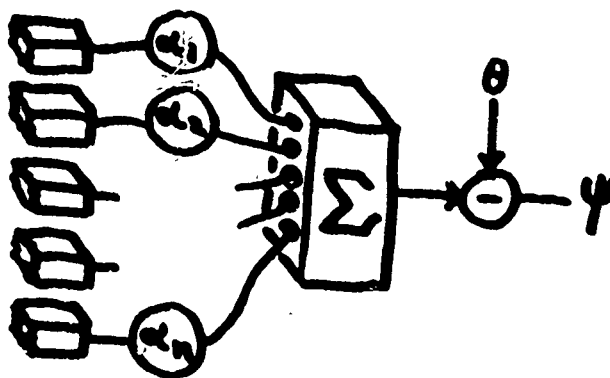


Fig. 0.3

The possibility of infinite sums leads to more fussy complications in proofs than it is worth. To prevent it happening we shall "quantize" the plane by assuming it to be made up of discrete little squares. This is equivalent in effect to identifying figures which differ by less than some "tolerance." Moreover we shall consider only bounded figures X , and choose ϵ so that, for a given X , only a finite number of φ 's make $\varphi(X) = 1$. With these stipulations (which will be set out more carefully later) we define:

A perceptron is a device capable of computing all predicates which are linear in some given set of ϵ of "partial predicates."* We obtain families of perceptrons by imposing restrictions on the members of ϵ .

The following families seem to be particularly interesting:

- (a) Diameter-limited perceptrons: the support sets of members of ϵ are restricted not to exceed a fixed diameter in the ordinary metric of the plane.
- (b) Order-restricted perceptrons: we say that a perceptron has $\leq n$ if no member of ϵ has more than n points in its support.

- (c) Gamba perceptrons: the members of ϵ have unrestricted support but must be "linear threshold functions" (i.e.,

* That is, we are given a set of φ 's, but can select freely their weights, the α_φ 's, and also the threshold θ .

have order 1). This is equivalent to saying that each φ in Φ is defined by a signed measure on R , and a threshold θ_φ .

- (d) Random perceptrons: These are the form most extensively studied by Rosenblatt's group: they are order-restricted and Φ is generated by a stochastic process according to an assigned distribution function.

To give a preview of the kind of results we will obtain, we present here a simple example of a negative result:

Theorem 8.2.3: A diameter-limited perceptron cannot determine whether or not all the parts of a geometric figure are connected to one another. The proof requires us to consider just four figures:

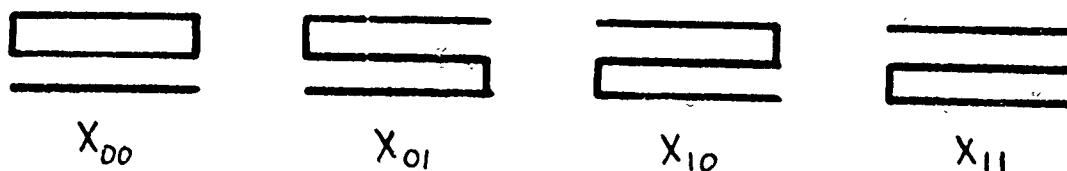
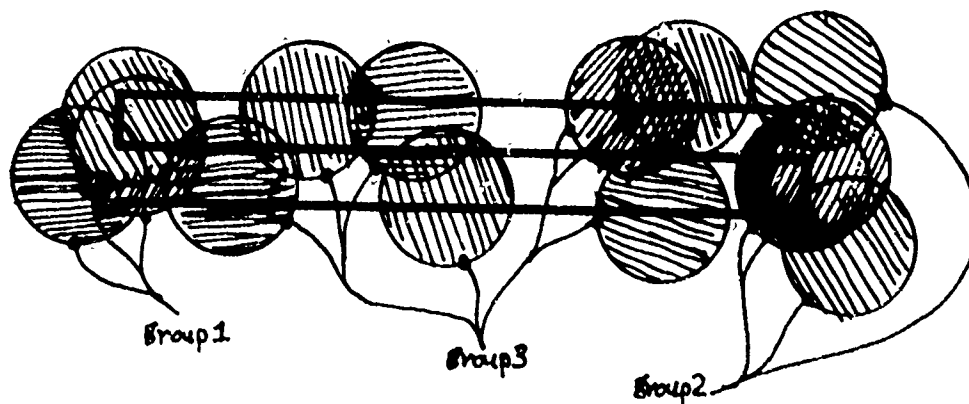


Fig. 8.2.3

and a diameter-limited perceptron Φ whose support sets have diameters like those indicated by the circles below:



Suppose that a perceptron could distinguish the disconnected figures X_{00} and X_{11} from the connected figures X_{10} and X_{01} , i.e., by whether or not

$$\sum \alpha_{\varphi} \varphi > \theta$$

that is, according to whether or not

$$\sum_{\text{group 1}} \alpha_{\varphi} \varphi + \sum_{\text{group 2}} \alpha_{\varphi} \varphi + \sum_{\text{group 3}} \alpha_{\varphi} \varphi > \theta$$

Then for X_{00} the sum of the three Σ 's is negative. In changing X_{00} to X_{10} only $\sum_{\text{group 1}}$ is affected, and its value must increase enough to make

the total exceed θ . If we change X_{00} to X_{01} similarly $\sum_{\text{group 2}}$

must increase. But if we change X_{00} to X_{11} then, both $\sum_{\text{group 1}}$ and

$\sum_{\text{group 2}}$ will have these increases; $\sum_{\text{group 3}}$ is unchanged in every case, so

the full increase must be even more on the positive side, and the perceptron must accept X_{11} as connected!

Q.E.D.

0:4 Seductive Aspects of Perceptrons, I:

Homogeneous Programming and Learning

The purest vision of the perceptron as a pattern-recognizing device is the following:

The machine is built with a fixed set of computing elements for the partial functions φ , usually obtained by a random process. To make it recognize a particular pattern (set of input figures) one merely has to set the parameters α_φ to suitable values. Thus "programming" takes on a pleasingly homogeneous form. Moreover since "programs" are representable as points $(\alpha_1, \alpha_2, \dots, \alpha_n)$ in an n -dimensional space, they inherit a metric which makes it easy to imagine a kind of automatic programming which people have been tempted to call learning: by attaching feedback devices to the parameter controls they propose to "program" the machine by providing it with a sequence of input patterns and an "error signal" which will cause the parameters to change in the right direction when the machine makes an inappropriate decision. The perceptron convergence theorems* define conditions under which this procedure is guaranteed to find, eventually, a correct set of values.

To separate reality from wishful - thinking, we begin by making a number of distinctions. Let Φ be the set of partial predicates of the perceptron and $L(\Phi)$ the set of predicates linear in Φ . Thus $L(\Phi)$ is the repertoire of the perceptron -the set of predicates it can compute as the parameters α_φ range over all possible

* See Chapter 10.

values. Of course $L(\Phi)$ could in principle be the set of all predicates (on 2^R): but this is universally recognized as being impossible in practice, since Φ would have to be astronomically large. So any real perceptron has a limited repertoire. The ease and uniformity of programming have been bought at a cost. We contend that the traditional investigations of perceptrons do not realistically measure this cost. In particular they neglect the following crucial points:

- i. The translation of geometric patterns or predicates on the input plane R into n -dimensional vectors $(\alpha_1 \dots \alpha_n)$ loses the geometric individuality of the patterns and has only led to a theory which can do little more than count the number of predicates in $L(\Phi)$! As a result not many people seem to have observed or suspected that there might be particular geometrically meaningful and intuitively simple predicates which belong to no practically realizable set $L(\Phi)$. We have already given an example of this for the diameter-limited case and will later extend it to the order-limited cases. At the same time we shall show that certain predicates which might intuitively seem to be difficult for these devices can, in fact, be recognized by low-order perceptrons.

- ii. Little attention is paid to the size, or more precisely, the information content, of the parameters $\alpha_1, \dots, \alpha_n$. We shall give examples (which we conjecture to be typical rather than exceptional) where the ratio of the largest to the smallest of the coefficients is meaninglessly big. Under these conditions it is of no (practical) avail that a predicate be

in $L(\Phi)$. In some cases the information capacity needed to store $\alpha_1 \dots \alpha_n$ is greater than that needed to store the whole class of figures in the pattern!

iii. Closely related to the previous point is the problem of time-of-convergence in a "learning" process. Practical perceptrons are essentially finite-state devices. It is therefore vacuous to cite a perceptron convergence theorem (see Chapter 10) as an assurance that a perceptron will eventually find a correct setting of its parameters (if one exists). It could do so trivially by cycling through all its states, e.g. by trying all coefficient assignments. The significant question is: how fast the perceptron converges relative to the time taken by this homeostat-like random procedure? It will be seen that there are situations of some geometric interest for which the convergence time can be shown to increase more than exponentially with the size of the set R .

Perceptron theorists are not alone in neglecting these precautions. A perusal of any typical collection of papers on "self-improving" systems will provide a generous sample of schemes for "learning" or "adaptive" machines which lack even the degree of rigor and formal definition to be found in the literature on perceptrons. The proponents of these schemes never provide any analysis of the range of behavior which can be learned, nor show any awareness of the price paid to make learning easy, by restricting this range with hidden assumptions about the environment in

which the device is to operate. One is tempted to detect a mystique of unintelligibility: the murkier the mechanism the greater the virtue, as though it were better to try to find unanalysable analogues for biological systems than to try to find mechanistic explanations for them.

These critical remarks must not be read as suggestions that we are opposed to making machines than can "learn." Exactly the contrary. But we do believe that significant learning at a significant rate presupposes some significant prior structure. Simple learning schemes based on adjusting coefficients can indeed be practical and valuable when the partial functions are closely matched to the task, as they are in Samuel's checker player. A perceptron with a set of partial functions properly designed for a discrimination known to be of suitably low order will have a good chance to improve its performance adaptively. Our deep objection is to the concept of giving a high-order problem to a quasi-universal perceptron whose partial functions have not been constructed with any particular task in mind.

It may be argued that people are universal learning machines and so a counter-example to this thesis. But our brains are sufficiently structured to be programmable in a much more general sense than the perceptron and our culture is sufficiently structured to provide, if not actual program, at least a rather complex set of interactions which govern the course of whatever the process of self-programming may be. Moreover it takes time for us to become universal learners; the slow transition from infancy to

intellectual maturity is rather a confirmation of the thesis that the rate of acquisition of new cognitive structure (i.e., learning) is a sensitive function of the level of existing cognitive structure.

0.5 Seductive Aspects of Perceptrons, II:

Parallel Computation

The perceptron was conceived as a parallel-operating device in the physical sense that the partial predicates are computed simultaneously. From a formal point of view the important aspect is that they are computed independently of one another. The price paid for this is that all the ϕ_i must be computed, although only a minute fraction may in fact be relevant to the final decision. The total amount of computation may become vastly greater than that which would be carried out in a sequential process that can decide what next to compute, conditionally on the outcome of earlier computation. Thus the choice between parallel and serial methods in any particular situation must be based on the relative value of reducing the (total elapsed) time against the cost of the additional computation involved. In the case of the perceptron the concept of order provides a basis for the estimation of the latter quantities.

Even low order predicates may involve large amounts of wasteful computation of information which would be irrelevant to a serial computation. But the numbers can remain within physically realizable bounds, especially if a large tolerance (or "blur") is acceptable. High order predicates

create a completely different situation. An instructive example is provided by the essentially global predicate of connectivity:

$\psi_{\text{con}}(X) = 1$ if and only if X is a connected figure. As shown in Chapter 5 a perceptron for this predicate on a 100×100 toroidal retina would need partial functions that look at (a most conservative minimum of) more than 800 points.* In this case the computation of local functions is irrelevant to a perceptron-like linear threshold decision: the partial functions are themselves global. Moreover, the number of possible partial functions with such large support makes nonsense of any hope that a realizable randomly-generated set of them would be sufficiently dense to span the appropriate space of functions. To make this point even sharper we shall show that for certain predicates and classes of partial functions, the number of partial functions would exceed physically realizable limits even for a perceptron designed specifically for the particular predicate.

The general conclusion to be drawn is that the appraisal of any particular scheme of parallel computation cannot be undertaken rationally without a theory of the corresponding dichotomy of problems as local and global. The lack of a general theory of what is global and local is no excuse for avoiding the problem in particular cases. The analyses below

* Unless the predicates are specially designed, their number may turn out to be of the order of 2^{1000} , and so may their coefficients. This would make it necessary to compute serially, anyway, since all the power of Niagara Falls would not be enough to run them in parallel.

show that it is not impossibly difficult to develop such a theory for a limited class of computing devices such as the perceptrons.

C.6 Seductive Aspects of Perceptrons, III:

The Use of Simple Analogue Devices

An attractive feature of the perceptron is the idea that the linear threshold decision function can be computed by a very simple analogue device. It is perhaps generally appreciated that the utility of the scheme is limited by the sparseness of linear threshold functions in the set of all logical functions. However, almost no attention has been paid to the possibility that the set of linear functions which are practically realizable may be rarer still. To illustrate this problem we shall compute the minimal ratio between largest and smallest coefficients in the linear representations of certain predicates. It will become apparent that this ratio can increase faster than exponentially with the number of distinguishable points in R. It follows that for "big" input sets--say larger than 20--no simple analogue storage device can be made with enough information capacity to store the whole range of coefficients!

To avoid misunderstanding perhaps we should repeat the qualification made in connection with our critique of the perceptron as a model for "learning devices." We have no doubt that analogue devices of this sort have a role to play in pattern recognition. But we do not see that any good can come of experiments which pay no attention to the limiting factors

which will assert themselves as soon as the small model is scaled up to a usable size.

0.7 Mathematical Plan: Introduction to Part I

Part I (Chapters I--IV) contains a series of definitions and general theorems required for Part II. It would be difficult to struggle through this material without a preliminary picture of the roles these mathematical devices are destined to play. We can give such a picture by outlining how we will prove the following theorem:

Theorem 4.1 (Chapter 3): Informal Version

Suppose the retina R has a finite number, $|R|$, of points.

Then there is no perceptron $\sum \alpha_{\varphi} \varphi(X) > 0$ that will tell us whether or not the "number of points in X is odd or even" unless at least one of the φ 's has support $= |R|$. Thus no bound can be placed on the orders of perceptrons that solve this problem for arbitrarily large retinas.

The proof uses several steps:

Step 1. In §1.1--§1.4 we define "perceptron," "order," etc. more precisely, and show that certain details of the definitions can be changed without serious effects, i.e., that $\sum \alpha_{\varphi} \varphi \geq 0$ can always be replaced by $\sum \alpha_{\varphi} \varphi > 0$.

Step 2. In §1.3 we define the particularly simple φ -functions called "masks." For each subset A of the

retina define the mask $\mu_A(X)$ to have value 1 if the figure X contains or "covers" all of A , value 0 otherwise. Then we prove the simple but important theorem (§1.5) that if a predicate has order $\leq k$ (see §1.3) in any set of φ -functions, there is an equivalent perceptron that uses only masks of support $\leq k$. (See §0.2.)

Step 3. To really "get at" the parity--the "odd-even" property--we ask: what rearrangements of the input space R leave it unaffected? That is, we ask about the group of transformations that have no effect on it. This seems like using excessively high-powered mathematics, but it seems necessary for the more difficult problems so we should get used to it here. In this case the group is the whole permutation group on R --the set of all rearrangements of its points.

Step 4. In Chapter 2 we show how to use this group to reduce its perceptron to a simple form. In the present case, the group-invariance theorem (see section 2.2) shows that for the parity perceptron all masks with the same support-size--that is, all that look at the same numbers of (though different sets of) points--can be given identical coefficients. Let β_j be the weight assigned to masks that have support-size = j .

Step 5. It then follows that the parity perceptron can be written in the form

$$\sum_{j=0}^K \beta_j \binom{|X|}{j} > 0$$

where K is the largest support and $\binom{|X|}{j}$ is the number of subsets of X that have j elements. (Define $|X|$ to be the number of points in picture X .)

Step 6. Because

$$\binom{n}{j} = \frac{n(n-1) \dots (n-j+1)}{1 \cdot 2 \cdot 3 \dots j} = \frac{n^j}{j!} + \dots \pm \frac{n}{j}$$

which is a polynomial of degree j in n , we can put our predicate in the form

$$P_K(|X|) > 0$$

where P_K is a polynomial in $|X|$ of algebraic degree $\leq K$. Now if $|X|$ is even, $P_K(|X|) > 0$ while if $|X|$ is odd, $P_K(|X|) \leq 0$ so that in the range $0 \leq |X| \leq |R|$, P_K must change its direction $|R| - 1$ times. But a polynomial must have degree $\geq |R|$ to do that, so we conclude that $K \geq |R|$. This completes the proof exactly as done in Chapter 3.1.

This shows how the algebra works in. For some of the more difficult connectedness theorems of Chapter 5, we need somewhat more algebra and group theory. In Chapter 4 we push the ideas about the geometry of algebraic degrees a little further to show that some surprisingly simple predicates require unbounded-order perceptrons.

To see some simpler, but still characteristic results, the reader might turn directly to Chapter 8, which is almost self-contained because it does not need the algebraic theory.

CHAPTER 1: THEORY OF LINEAR BOOLEAN SEPARATION FUNCTIONS

1.0

In this section we develop the theory 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.

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:

1.1 Conventions

(i) Let R be an arbitrary set and F a family of subsets of R . Using the letters $A, B, C, \dots, X, Y, Z$ for subsets of R it is natural to associate with F a predicate $\varphi_F(X)$ which is TRUE if and only if $X \in F$.

(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 adopt the notation $\lceil \psi(X) \rceil$ 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 $\lceil 3 < 5 \rceil = 1$, which is true and $3 < \lceil 5 = 1 \rceil$, which is false.

(iii) Occasionally it will be convenient in examples to use the

traditional representation of $\varphi(X)$ as a function of n "boolean variables" where $n = |R|$. If the elements of R are $x_1, \dots, x_n$, it is traditional to think of a subset X of R as an assignment of the values 1 or 0 to the 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 i^{th} point in the given enumeration of R , and for the set function $[x_i \in X]$. This notation is particularly convenient when φ 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

$$\varphi(X) = [x_i \in X \text{ or } x_j \in X].$$

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

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

We call $S(\varphi)$ the support of φ .

1.2 Functions Linear with Respect to a Class of Predicates

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

* For an infinite space R , some predicates will have $S(\varphi)$ undefined. For example, suppose that $\varphi(X) = 1 \iff \{X \text{ contains an infinite set of points}\}$. For then one can determine $\varphi(X)$ by examining the intersection of X with any set of S that is the complement of R in some finite set. And there is no minimal such S .

$$\psi(X) = \left[\sum_{\varphi \in \Phi} \alpha_{\varphi} \varphi(X) > \theta \right].$$

This can be written more simply as

$$\psi = \left[\sum_{\varphi} \alpha_{\varphi} \varphi > \theta \right].$$

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

Proposition: The following formal modifications lead to equivalent definitions:

- (1) If Φ is assumed to contain a constant function, θ can be taken to be zero.
- (2) The inequality sign "<" can be replaced by ">," "<="," ">=".
- (3) It can be assumed that the exact equality $\sum \alpha_{\varphi} \varphi = \theta$ never arises.
- (4) We can restrict the α_i to be rationals or integers.
- (5) The choices allowed under (1)--(4) can be made independently.

Proof: Most points are obvious. To prove (3) for general real coefficients note that there are countably many X's and so countably many values of $\sum \alpha_{\varphi} \varphi$. To prove (4), for integer α 's, note that we could make all the α 's even and put $\theta = 1$.

Our most common choice will be to take the α_i as integer and θ as zero.

Remark on notation: In view of this definitional invariance, it might be useful to abbreviate the $\left[\sum \alpha_{\varphi} \varphi(X) > \theta \right]$ notation to simply $\left[\sum \alpha_{\varphi} \varphi \right]$.

* But some don't hold in the infinite-retina case.

in view of option 3. Then one could go on and use inner-product notations like $\langle \alpha_\varphi \cdot \varphi \rangle$ or even $\alpha_i \varphi^i$. However, we have often found the form with explicit θ more convenient.

1.3 The Central Concept of Order*

The order of ψ is the smallest k for which there is a $\sharp$ satisfying

$$\psi \in L(\sharp)$$

$$\varphi \in \sharp \implies |S(\varphi)| \leq k$$

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

Functions of order 1 appear in the literature under the name of "linear threshold functions." Note that the concept of order would be unaffected by imposing the condition that $\sharp$ contain a constant function.

It follows that it would not be changed by assuming $\theta = 0$ in the definition of $L(\sharp)$. Clearly neither can any of the other options of 1.2 affect the value of the order of a predicate.

Definition: φ is called a mask if there is a set A such that

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

We denote this function by φ_A .

* We emphasize that the order of a predicate is defined absolutely--not simply relative to a particular $\sharp$ -class.

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

$$y_1 \wedge y_2 \wedge \dots \wedge y_t.$$

where $\{y_i\}$ is the subset A of R.

In particular, the constant function $\varphi(X) = 1$ is the mask with support = 0.

1.4 Examples of Linear Representation

Proposition: All masks are of order 1.

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

$$\varphi_A = \left\lceil \sum_{x \in A} \varphi_x \geq |A| \right\rceil.$$

In particular the individual point-function 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, $x = y$, 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:

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

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

$$x_2 \vee \bar{x}_1 = [x_2 + (1 - x_1) > 0] = [x_2 - x_1 > -1], \text{ which is}$$

also $x_1 \supset x_2$, is of order 1.

Example (ii)

$$x_1 \equiv x_2, \text{ which is}$$

$$\begin{aligned} x_1 x_2 \vee \bar{x}_1 \bar{x}_2 &= [x_1 x_2 + (1 - x_1)(1 - x_2) > 0] \\ &= [2x_1 x_2 - x_1 - x_2 > -1] \end{aligned}$$

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

Example (iii)

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

$$\psi^M(X) = [|X| = M],$$

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

Proof: consider the representation

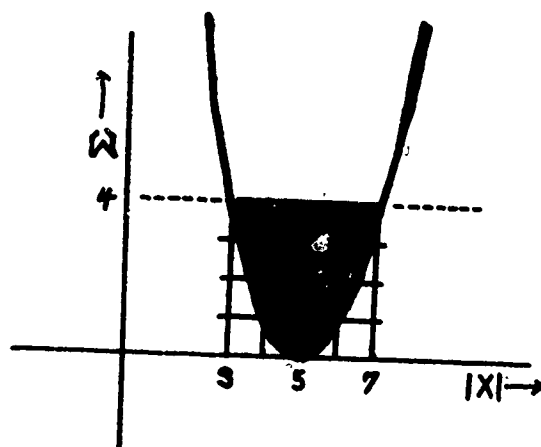
$$\varphi^M(X) = [(2M - 1) \sum_{\text{all } i} x_i + (-2) \sum_{i \neq j} x_i x_j \geq M^2].$$

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

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

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

Observe that, by decreasing the constant term, we can modify the predicate to accept counts within an arbitrary interval instead of a single value.



$$\lceil (|X| - 5)^2 \leq 4 \rceil = \lceil 3 \leq |X| \leq 7 \rceil$$

Fig. 1.4

Note that the linear form for the counting function does not contain R explicitly. Hence it works as well as 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 $\lceil \sum x_i \leq M \rceil$.

We note in passing that we can obtain an arbitrary function $f(|X|)$ of the area of a figure from these predicates by writing

$$\begin{aligned} f(|X|) &= f(0) + \sum_{k=1}^R (f(k)) \cdot \lceil |X| \geq k \rceil \\ &= f(0) + \sum_{k=1}^{|X|} (f(k)) - f(k-1)) \\ &= f(|X|). \end{aligned}$$

This fact is used in §8.2.

1.5 The "Positive Normal Form Theorem"

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

Theorem 1. Every φ is a linear threshold function with respect to the set of all masks.

Any Boolean function $\psi(x_1, \dots, x_n)$ can be written in the disjunctive normal form

$$\psi(x) = C_1(X) \vee C_2(X) \vee \dots \vee C_p(X)$$

where each $C_i(X)$ has the form

$$z_1 \wedge z_2 \wedge \dots \wedge z_q$$

and each z is either an x_i or an $\bar{x}_i$. And since at most one of the $C_i(X)$ can be true for any X , we can rewrite ψ using the arithmetic sum:

$$\psi(X) = C_1(X) + C_2(X) + \dots + C_p(X)$$

(even for infinite sums). The bars over the letters can be eliminated by using the equation

$$\bar{\alpha x_j} \beta = \alpha(1 - x_j)\beta = \alpha\beta - \alpha x_j \beta$$

where α and β are conjunctions, so that any bar on a term within a conjunction can be removed.

When all the bars have been eliminated and like terms have been collected together we have:

$$\psi(X) = \sum \alpha_i \varphi_i(X) \quad (A)$$

where each φ_i is a mask, and each α_i is an integer. Since $\psi(X)$ is 0 or 1, (A) is equivalent to

$$\psi(X) = \left[\sum \alpha_i \varphi_i(X) > 0 \right]. \quad (B)$$

Example: $\left[X_1 + X_2 + X_3 \text{ is odd} \right] = X_1 + X_2 + X_3 - 2X_1X_2 - 2X_2X_3 - 2X_3X_1 + 4X_1X_2X_3$
 Theorem 1.5.2: The representation (A) is unique.

To see this let $\{\varphi_i\}$ be a set of masks and $\{\alpha_i\}$ a set of non-zero numbers. Choose a k for which $S(\varphi_k)$ is minimal, i.e., there is no j such that $S(\varphi_j) \subset S(\varphi_k)$. Then:

$$\varphi_k(S(\varphi_k)) = 1$$

$$\varphi_j(S(\varphi_k)) = 0 \quad j \neq k.$$

It follows that $\sum \alpha_i \varphi_i(X)$ is not identically zero since it has the value α_k for $X = S(\varphi_k)$.

Now if $\sum \alpha_i \varphi_i(X) \equiv \sum \beta_i \varphi_i(X)$ for all X , then $\sum (\alpha_i - \beta_i) \varphi_i(X) = 0$ for all X . It follows that all $\alpha_i = \beta_i$. This proves the uniqueness of

the representation (A) which we shall call the positive normal form of ψ . Note that the positive normal form has the values 0 and 1 as ordinary arithmetic sums; i.e., without the $\lceil \cdot \rceil$ device of interpreting the validity of an inequality as a predicate

Theorem 1.5.3: ψ is of order k if and only if k is the smallest number for which there exists a set Φ of masks satisfying

$$\varphi \in \Phi \implies |S(\varphi)| \leq k$$

and

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

Proof: In $\psi = \lceil \sum \alpha_i \varphi_i > 0 \rceil$ each φ_i can be replaced by its positive normal form. If $|S(\varphi_i)| < k$, this will be true of all the masks that appear in the positive normal form.

Example:

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

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

illustrating how the negations are removed without raising the order. This particular order-3 form appears later (§5) in a perceptron that recognizes convex figures.

Theorem 1.5.4: If ψ_1 has order 0_1 and ψ_2 has order 0_2 , then $\psi_1 \oplus \psi_2$ and $\psi_1 \equiv \psi_2$ have order $\leq 0_1 + 0_2$.

Proof: Let $\psi_1 = [\sum \alpha_i \varphi_i > 0]$ and $\psi_2 = [\sum \beta_j \varphi_j > 0]$ and assume that the coefficients are chosen so that equality never arises. Then

$$\begin{aligned}\psi_1 \oplus \psi_2 &= [(\sum \alpha_i \varphi_i)(\sum \beta_j \varphi_j) > 0] \\ &= [\sum \alpha_i \beta_j \varphi_i \varphi_j > 0]\end{aligned}$$

But

$$|S(\varphi_i \varphi_j)| < |S(\varphi_i)| + |S(\varphi_j)|.$$

The other conclusion follows from $[\psi \equiv x] = 1 - [\psi \neq x]$.

Example:

Since $\psi^M(x) = [x \geq M] \equiv [x \leq M]$ we conclude that ψ^M has order ≤ 2 , the result of example (iii).

Question: What can be said about the orders of $[\psi_1 \& \psi_2]$ and $[\psi_1 \sim \psi_2]$? The answer to this question may be surprising, in view of the simple result of the previous theorem: It is shown in §4 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 $(\psi_1 \sim \psi_2)$ have order $> n$. In fact, suppose that $K = A \cup B \cup C$ where A , B , and C are large disjoint subsets of K . Then $\psi_1 = [X \cap A > X \cap C]$ and $\psi_2 = [X \cap B > X \cap C]$ each have order 1 because they are represented by

$$\left[\sum_{x_i \in A} x_i - \sum_{x_i \in C} x_i > 0 \right] \text{ and } \left[\sum_{x_i \in B} x_i - \sum_{x_i \in C} x_i > 0 \right]$$

but we shall see in §4 that $(\psi_1 \wedge \psi_2)$ and $(\psi_1 \vee \psi_2)$ are not even of finite order in the sense described in §1.6 below. On the other hand one can be surprised in special cases: see, e.g., Theorem §5.5.

1.6 Predicates of Finite Order

Strictly, a predicate is defined for a particular set R and it makes no formal sense to speak of the same predicate for different R 's. However the motivation of our work was entirely from "predicates" defined independently of R --e.g., the number of elements in the set X , or other geometric properties of figures in a real Euclidean plane to which X and R provide mere approximations. To be very precise we could use a phrase such as predicate scheme to refer to a general construction which defines a predicate for each of a large class of sets R . In general (except in this section) we shall use "predicate" in this wider sense.

Suppose we are given a predicate scheme Ψ which induces a predicate ψ_R for each of a family $\{R\}$ of retinas. We shall say that Ψ is of finite order if the orders of the ψ_R are uniformly bounded for all R 's in the appropriate family. Two examples will make this clearer:

(i) Let $\{R_i\}$ be a sequence of sets with $|R_i| = 1$. For each R_i there is a predicate ψ_i defined by the predicate scheme $\psi^{\text{PAR}}(X)$ which asserts, for $X \subset R_i$, that " $|X|$ is an even number." As we will see in §3.1, the order of any such ψ_i must be ≥ 1 . Thus ψ^{PAR} is not of finite

order.

(ii) Now let ψ_i be the predicate defined over R_i by the predicate scheme ψ_{TEN} :

$$\psi_i(X) = [|X| = 10].$$

We shall show in 1.4 that ψ_i is of order 2 for all R_i with $i > 10$ and (obviously) of order zero for $R_1 \dots R_9$. Thus the predicate scheme ψ_{TEN} is of finite order. We shall say in this case that it is of order 2.

In these cases one could obtain the same dichotomy by considering infinite sets R : on an infinite retina the predicate

$$\psi_{TEN}(X) = [|X| = 10]$$

is of finite order, in fact of order = 2, while

$$\psi_{PAR}(X) = [|X| \text{ is even}]$$

has no order. We shall often look at problems in this way and in §7 will discuss formalization of the concept of an infinite perceptron. It should be noted, however, that the use of infinite perceptrons does not cover all cases. For example the predicate

$$\dagger(X) = \lceil |X| > \frac{1}{2} |R| \rceil$$

is well-defined and of order 1 for any finite R. It is meaningless for infinite R, yet we would like to consider the corresponding predicate-scheme to have finite order.

CHAPTER 2: THE GROUP THEORY

In this chapter we consider linear threshold functions that are invariant under groups of transformations of the points of the base-space R . The purpose of this, realized finally in Chapter 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.

2.1 Example: Coefficients Averaged Over A Symmetry

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 $x_1 x_2 \vee \bar{x}_1 \bar{x}_2$ 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

$$\forall (x_1, x_2) \quad x_1 x_2 \vee \bar{x}_1 \bar{x}_2 = \alpha x_1 + \beta x_2 > \theta. \quad (1)$$

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

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

In this simple case it is easy enough to deduce the contradiction:

$$\left. \begin{array}{l} \alpha \leq \theta \text{ and } \theta < 0 \implies \alpha < 0 \\ \beta \leq \theta \text{ and } \theta < 0 \implies \beta < 0 \end{array} \right\} \implies \alpha + \beta < \alpha \implies \alpha + \beta < \theta$$

while by hypothesis $\alpha + \beta > \theta$. 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

$$\alpha x_1 + \beta x_2 > \theta$$

$$\alpha x_2 + \beta x_1 > \theta$$

hence

$$\left(\frac{\alpha + \beta}{2}\right) x_1 + \left(\frac{\alpha + \beta}{2}\right) x_2 > \theta$$

by adding the inequalities.

Similarly

$$\alpha x_1 + \beta x_2 \leq \theta$$

$$\alpha x_2 + \beta x_1 \leq \theta$$

yields

$$\left(\frac{\alpha + \beta}{2}\right) x_1 + \left(\frac{\alpha + \beta}{2}\right) x_2 \leq \theta.$$

* One can say the same of geometric arguments about hyperplanes separating vertices of the n-dimensional unit-cube.

It follows that if we write γ for $\frac{\alpha + \beta}{2}$, then

$$\psi(x_1, x_2) = [\gamma x_1 + \gamma x_2 > \theta]$$

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

$$\psi(X) = [\gamma |X| > \theta] = [\gamma |X| - \theta > 0]$$

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

Now consider three values of X ,

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

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. See 2.5. In fact, for order 1, it is the same method.)

2.2 Equivalence-Classes of Predicates

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

Let G be a group of transformations of R onto itself. If $g \in G$ and $X \subset R$ we define Xg , the result of transforming X by g , to be the set obtained by applying g to each member of X :

$$Xg = \{y | x \in X \wedge y = xg\}.$$

Then we can define an equivalence relation $\varphi \equiv_G \varphi'$ of predicates with respect to the group by

$$\varphi \equiv_G \varphi' \text{ if and only if, } (\exists g)(\forall X)(\varphi(X) = \varphi'(Xg)).$$

That is, φ is equivalent to φ' if there is a transformation g such that $\varphi(X)$ and $\varphi'(Xg)$ are always the same. Our main theorem shows that if a perceptron is to classify patterns in a way that is invariant under group G , then its computation can, and in a sense, can only depend on the G -equivalence classes of its φ -functions.

2.3 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., $\varphi \in \Phi, g \in G \implies \varphi g \in \Phi$, where we define φg so that
 $\varphi g(X) \equiv \varphi(Xg)$.

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

Then there exists a linear representation of ψ ,

$$\psi = \left[\sum_{\varphi \in \Phi} \beta_{\varphi} \varphi > 0 \right]$$

for which the coefficients β_{φ} depend only on the G -equivalence class of φ , i.e.,

$$\varphi \sim_G \varphi' \implies \beta_{\varphi} = \beta_{\varphi'}.$$

Remarks:

(1) These conditions are stronger than they need be. To be sure, the theorem is not true in general for infinite groups. A counterexample will be found in § 11.4. However, in special cases we can prove the theorem for infinite groups. An example with interesting consequences will be discussed later. (See § 11.2 .) It will also be seen that the assumption that G be a group can be relaxed slightly.

(2) We have not investigated relaxing condition (ii), and this would be interesting. However, it does not interfere with our methods for showing certain predicates to be not of finite order. When the theorem is applied to show that a particular ψ is not in $L(\Phi)$ for a particular Φ , it shows also that ψ is not even in the possible larger L of Φ closed under G . If one found a useful notion similar to but more delicate than "order," one

might have to find a correspondingly sharper invariance theorem for it.

Proof: Let the given linear representation of ψ be:

$$\psi = [\sum \alpha(\varphi) \varphi > 0].$$

The new representation will be:

$$\psi = [\sum \beta(\varphi) \varphi > 0].$$

where

$$\beta(\varphi) = \sum_{g \in G} \alpha(\varphi g).$$

It is clear that $\beta(\varphi)$, so defined, depends only on the equivalence class of φ . For if $\varphi \equiv \varphi'$, then $\exists g_0$ such that $\varphi' = \varphi g_0$ and

$$\beta(\varphi') = \sum_{g \in G} \alpha(\varphi' g) = \sum_{g \in G} \alpha(\varphi g_0 g) = \sum_{g \in g_0^{-1} G} \alpha(\varphi g) = \beta(\varphi).$$

It is equally easy to see that ψ is indeed given by

$$\psi = [\sum \beta(\varphi) \varphi > 0].$$

Choose any X . Suppose that $\psi(X) = 1$. Then $\psi(Xg) = 1$ for all $g \in G$, i.e.,

$$\sum_{\varphi \in \mathfrak{F}} \alpha(\varphi) \varphi(Xg) = \sum_{\varphi \in \mathfrak{F}} \alpha(\varphi) \varphi_g(X) > 0 \text{ for all } g \in G.$$

If we substitute ϕ' for ϕ_g this can be written

$$\sum_{\varphi' \in \mathfrak{F}} \alpha(\varphi_g^{-1}) \varphi'(X) > 0 \text{ for all } g \in G$$

which is the same as

$$\sum_{\varphi \in \mathfrak{F}} \alpha(\varphi_g) \varphi(X) > 0 \text{ for all } g \in G.$$

Adding these equations term by term:

$$\sum_{g \in G} \sum_{\varphi \in \mathfrak{F}} \alpha(\varphi_g) \varphi(X) > 0$$

i.e.,

$$\sum_{\varphi \in \mathfrak{F}} \left(\sum_{g \in G} \alpha(\varphi_g) \right) \varphi(X) > 0$$

i.e.,

$$\sum_{\varphi \in \mathfrak{F}} \beta(\varphi) \varphi(X) > 0.$$

Similarly, if $\psi(X) \leq 0$ we obtain

$$\sum_{\varphi \in \mathfrak{F}} \beta(\varphi) \varphi(X) \leq 0.$$

This proves the theorem. For readers to whom these ideas seem difficult to work with abstractly, some concrete examples of the equivalence classes will be useful; the geometric "spectra" of §5.2 and especially §5.5 should be helpful.

We shall often use this theorem in the following form:

Corr. 1: Under the conditions of the theorem ψ has a representation

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

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

Proof: For masks, $\phi_A \equiv \phi_B$ if and only if $A = Bg$ for some $g \in G$.

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

$$\psi = \left\lceil \sum_i \alpha_i N_i(X) > 0 \right\rceil$$

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

Proof: ψ can be represented as

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

$$= \left[\sum_i \sum_{\varphi \in \Phi_i} \alpha_{\varphi} \varphi > 0 \right]$$

$$= \left[\sum_i \alpha_i \sum_{\varphi \in \Phi_i} \varphi > 0 \right] = \left[\sum_i \alpha_i N_i(X) > 0 \right]$$

2.4 The Triviality of Invariant Predicates of Order 1

Theorem: 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:

$$\left. \begin{aligned} \psi(X) &= [|X| > m] \\ \psi(X) &= [|X| \geq m] \\ \psi(X) &= [|X| < m] \\ \psi(X) &= [|X| \leq m] \end{aligned} \right\} \text{ for some } m.$$

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

$$\psi(X) = \left[\sum_{p \in X} \alpha \varphi_{\{p\}} > \theta \right] \text{ (or with some other inequality sign),}$$

i.e., the coefficient α is independent of p . But $\sum_{p \in X} \alpha \varphi_{\{p\}} > \theta$ can be transformed into $\sum_{p \in X} \varphi_{\{p\}} > \frac{\theta}{\alpha}$ (for $\alpha > 0$; for $\alpha \leq 0$ a similar argument proves the corresponding assertion). But $\sum_{p \in X} \varphi_{\{p\}} = |X|$. Thus order-1 invariant predicates can do nothing more than define a count on the

cardinality of "area" of figures.

2.5 Relation to Haar Measure

The last theorem is closely related to Haar's theorem on the unicity of invariant measures. For measures on finite sets the unique Haar measure is, in fact, the counting function $\mu(X) = |X|$. The following remarks make this relation a little more precise.

We first note that the set function defined by:

$$\mu(X) = \sum_{x_i \in X} \alpha_i$$

is a measure, i.e., satisfies:

$$\mu(X) + \mu(Y) = \mu(X \cup Y) + \mu(X \cap Y).$$

If we defined invariance by:

$$\mu(X) = \mu(Xg)$$

it would follow immediately from Haar's theorem that $\mu(X) = c|X|$, where c is a constant. Our hypothesis on μ is slightly weaker since we merely assume:

$$\mu(X) > 0 \iff \mu(Xg) > 0.$$

and deduce a correspondingly weaker conclusion, i.e.,

$$(\mu(X) > 0) \iff (c|X| > 0).$$

In the general case the relation between an invariance theorem and the theory of Haar measures is less clear since the set function $\Sigma_{i,k} \alpha_{i,k}(X)$ is not in general a measure. This seems to suggest some generalization of measure (perhaps on simplicial complexes rather than sets) but we have not tried to pursue the problem. Readers interested in the history of ideas might find it interesting to pursue the relation of these results to those of Pitts and McCulloch (1947).

CHAPTER 3: SOME SPECIAL PREDICATES

3.0

In this chapter we study the order of two particularly interesting predicates.

3.1 The Parity Functions

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

$$\psi_{\text{PAR}}(X) = [|X| \text{ is an odd number}].$$

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 3.1.1: 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 $\psi_{\text{PAR}} = [\sum \alpha_i \varphi_i > 0]$ where φ_i are the masks with $|S(\varphi_i)| \leq K$ and the α_i depend only on the equivalence classes defined by

$\equiv$
 G

Since masks with the same support are identical, and sets with the same cardinality can be transformed into one another by elements of G ,

$$\varphi_i \equiv \varphi_j \iff |S(\varphi_i)| = |S(\varphi_j)|.$$

Thus

$$\psi_{\text{PAR}} = \left[\sum_{j=0}^K (\alpha_j \sum_{\varphi \in \Phi_j} \varphi) > 0 \right]$$

where Φ_j is the set of masks whose supports contain exactly j elements.

We now calculate for an arbitrary subset X or R ,

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

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

$$C_j(X) = \binom{|X|}{j} = \frac{|X|(|X| - 1) \dots (|X| - j + 1)}{j!}$$

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

It follows that $\sum_{j=0}^K \alpha_j C_j(X)$ is a polynomial of degree $\leq K$ in $|X|$, say $P(|X|)$.

Now consider any sequence of sets $X_0, X_1, \dots, X_{|R|}$ such that X_1 contains 1 points, i.e., $|X_1| = 1$.

Then the sequence of values of $P(|X|)$,

$$P(|X_0|) \leq 0, P(|X_1|) > 0, P(|X_2|) \leq 0, \dots, P(|X_{|R|}|),$$

changes direction $|R|$ times as $|X|$ increases from 0 to $|R|$. But since P is a polynomial of degree K , it follows that $K \geq |R|$.

q.e.d.

From this we obtain the

Theorem 3.1.2: If $\psi_{\text{PAR}} \in L(\mathfrak{E})$ and if $\mathfrak{E}$ contains only masks, then $\mathfrak{E}$ contains every mask.

Proof: Suppose, if possible, that $\psi_{\text{PAR}} \in L(\mathfrak{E})$, that $\mathfrak{E}$ contains only masks, and the mask whose support is A does not belong to $\mathfrak{E}$. Let

$$\psi_{\text{PAR}} = \left[\sum_{\varphi \in \mathfrak{E}} \alpha_{\varphi} \varphi > 0 \right].$$

Define, for any ψ , $\psi^A(X) = \psi(X \cap A)$.

Clearly ψ_{PAR}^A , the parity function for subsets of A , is of order $|A|$ by the previous theorem.

Now consider φ^A for $\varphi \in \mathfrak{E}$. If $S(\varphi) \subset A$, clearly $\varphi^A = \varphi$.

If $S(\varphi)$ is not a subset of A , φ^A is identically zero since

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

It follows that either $S(\varphi^A)$ is a proper subset of A or φ^A is identically zero.

Let $\mathfrak{E}^A$ be the set of masks in $\mathfrak{E}$ whose supports are subsets of A .

Then $\psi_{PAR}^A = \lceil \sum_{\varphi \in \mathcal{A}} \alpha_{\varphi} \varphi > d \rceil$

But for all $\varphi \in \mathcal{A}$, $|S(\varphi)| < |A|$. It would follow that the order of ψ_{PAR}^A is less than $|A|$, which contradicts theorem 3.1.1. Thus the hypotheses are impossible and the theorem follows.

q.e.d.

Corr. 1: If $\psi_{PAR} \in L(\mathcal{A})$ then $\mathcal{A}$ must contain at least one φ for which $|S(\varphi)| \geq |R|$.

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

Corr. 2: Let $\mathcal{A}$ be the set of all ψ_{PAR}^A for proper subsets A of R. Then $\psi_{PAR} \notin L(\mathcal{A})$.

The further analysis of ψ_{PAR} , in Chapter 9.1 shows how functions that might be recognizable, in principle, by a very large perceptron, might not actually be realizable in practice because of impossibly huge coefficients. For example, it is shown that the ratio of the largest to the smallest coefficients of ψ_{PAR} must be $2^{|R|} - 1$.

3.2 *The "One-in-a-box" Theorem

Another predicate of great interest is associated with the geometric property of "connectedness." Its application and interpretation is deferred to Chapter 5; the basic theorem is proved now.

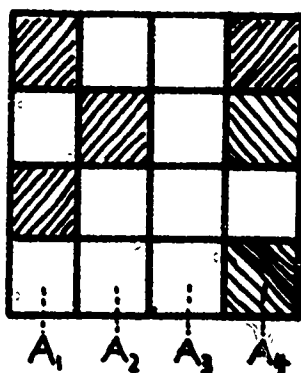
Theorem: Let $A_1, \dots, A_n$ be disjoint subsets of R and define the predicate

$$\psi(X) = \lceil (\forall i)(|X \cap A_i| > 0) \rceil$$

* This theorem is used to prove the theorem in Chapter 5.1. Because Chapter 5.7 gives an independent proof (using Theorem 3.1.1) this section can be skipped on first reading.

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 $\geq m$.

Corr.: If $R = A_1 \cup A_2 \cup \dots \cup A_m$, the order of ψ is at least $\left(\frac{|R|}{4}\right)^{1/3}$.



The retina is a square array of 16 cells, and A_i is the i -th column. The one-in-a-box predicate is not satisfied by the figure to the left, because no cell in the third column is occupied.

Fig. 3.2

Proof: For each $i = 1, \dots, m$ 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 $\varphi \in \Phi$ consider the ordered set of occupancy numbers.

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

Note that $\varphi_1 \stackrel{\Delta}{=} \varphi_2$ if and only if $|S(\varphi_1) \cap A_i| = |S(\varphi_2) \cap A_i|$ for each i .

Let $\hat{\varphi}_1, \hat{\varphi}_2, \dots, \hat{\varphi}_m$ be the equivalence classes.

Now consider an arbitrary set X and an equivalence class $\hat{\varphi}_j$. We wish to calculate the number $N_j(X)$ of members of $\hat{\varphi}_j$ satisfied by X , i.e.,

$$N_j(X) = |\{\varphi | \varphi \in \hat{\varphi}_j \wedge S(\varphi) \subset X\}|.$$

A simple combinatorial argument shows that

$$N_j(X) = \binom{|X \cap A_1|}{|S(\varphi) \cap A_1|} \binom{|X \cap A_2|}{|S(\varphi) \cap A_2|} \dots \binom{|X \cap A_m|}{|S(\varphi) \cap A_m|}$$

where $\binom{y}{n} = \frac{y(y-1) \dots (y-n+1)}{n!}$ and φ is an arbitrary member of $\hat{\varphi}_j$.

Since the numbers $|S(\varphi) \cap A_i|$ depend only on the classes $\hat{\varphi}_j$ 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 $x_i = |X \cap A_i|$:

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

Now let $\psi = [\sum \alpha_\varphi \varphi > 0]$ 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

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

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 ,

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

with the property that

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

i.e., that if,

$$\text{for all } i, 0 \leq x_i \leq 4m^2,$$

then

$$Q(x_1, \dots, x_m) > 0 \iff (\forall i)(x_i > 0).$$

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

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

Then $Q(x_1, \dots, x_m)$ becomes a polynomial of degree at most $2K$ in t . Now let t take on the values $t = 0, 1, \dots, 2m$.

Then $x_i = 0$ for some i if t is odd, in fact, for $i = \frac{t+1}{2}$

but $x_i > 0$ for all i if t is even.

Hence, by the definition of the ψ predicate, 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 $2K \geq 2m$ i.e., $K \geq m$. This completes the proof.

CHAPTER 4: THE AND/OR THEOREM

4.0 Introduction

In this chapter we prove the And/Or theorem stated in §1.5.

Theorem: There exist predicates, ψ_1 and ψ_2 of order 1 such that $\psi_1 \vee \psi_2$ and $\psi_1 \wedge \psi_2$ are not of finite order.

We prove the assertion for $\psi_1 \wedge \psi_2$. The other half can be proved in exactly the same way. The results will not be used in the sequel and could be omitted on a first reading.

4.1 Lemmas

We have already remarked in §1.5 that if $R = A \cup B \cup C$ the predicate

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

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 the formal statement is:

If $\psi_M(X)$ is the predicate of the stated form
for $|R| = 3M$, then the order of ψ_M increases
without bound as $M \rightarrow \infty$.

The proof follows the pattern of proofs in §3. We shall assume that the order of $\{\psi_M\}$ is bounded by a fixed integer, N , for all M , and derive a

1a

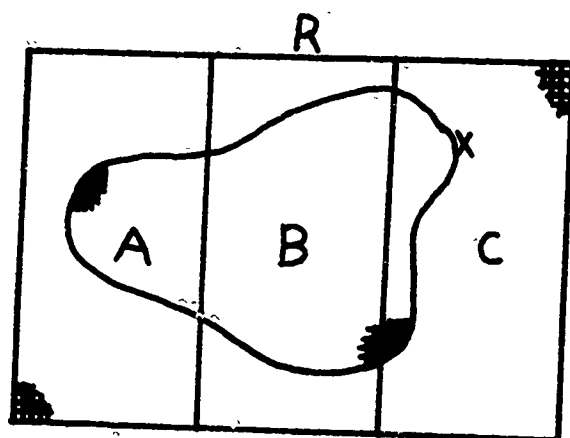


Fig. 4.1

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 M . We do this by choosing the group of permutations that leave the sets A , B , and C fixed but allow arbitrary permutations within the sets. The equivalence classes of masks are then characterized by three numbers, i.e., $|A \cap S(\varphi)|$, $|B \cap S(\varphi)|$ and $|C \cap S(\varphi)|$. For any given φ the number $N_\varphi(X)$ of masks in its equivalence class satisfied by a given set X is

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

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

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

Let $\mathcal{E}$ be the set of masks with $|\text{support}| \leq N$. Enumerate the equivalence classes of $\mathcal{E}$ and let $N_i(X)$ be the number of masks of the i th class satisfied by X . The group invariance theorem allows us to write:

$$\psi_M(X) = \left[\sum_i N_i(X) > 0 \right].$$

The sum $\sum_i N_i(X)$ is a polynomial of degree at most N in x, y, z . Call it

$P_M(x, y, z)$.

Now, by definition, for those values of x, y, z which are possible occupancy numbers, i.e., non-negative integers $\leq M$:

$$P_M(x, y, z) > 0 \text{ if and only if } x > z \text{ and } y > z.$$

We shall show, through a series of lemmas, that this cannot be true for all M .

Lemma 1

Let $P_M(x, y, z)$ be an infinite sequence of non-zero polynomials of degree at most N , with the property that for all positive integers x, y, z less than M

$$x > z \text{ and } y > z \implies P_M(x, y, z) \geq 0$$

and

$$x \leq z \text{ or } y \leq z \implies P_M(x, y, z) \leq 0.$$

(A)

Then there exists a single non-zero polynomial $P(x, y, z)$ of degree at most N with the property that the implications (A), with P replacing P_M , hold for all positive integral values of x, y, z . This follows by a straightforward compactness argument. (It should be observed that we have had to weaken the separation conditions (A) by allowing equality in both conditions since inequality would not be preserved in the limit. Consequences of this will make themselves felt in

the proof of lemma 2.) For the sake of completeness we include the following elementary proof. Write:

$$P_M(x,y,z) = \sum_{i=1}^T C_{M,i} m_i(x,y,z)$$

where $m_1, m_2, \dots, m_T$ is an enumeration of the monomials of degrees $\leq N$ in x, y, z .

Since the conditions on P_M are preserved under multiplication by a positive scaling factor, we can assume that

$$\sum C_{M,i}^2 = 1.$$

Now consider the set of points in T-space:

$$C_M = (C_{M,1}, C_{M,2}, \dots, C_{M,T}), \quad M = 1, 2, \dots$$

These all lie in a compact set--the surface of the unit T-dimensional sphere. There is, therefore, a subsequence C_{M_j} which converges to a limit:

$$C_{M_j} \rightarrow C = (c_1, c_2, \dots, c_T)$$

in the sense that, for each i ,

$$\lim_{j \rightarrow \infty} c_{M_j, i} = c_i.$$

The polynomial

$$P(x, y, z) = \sum_{i=1}^T c_i m_i(x, y, z)$$

inherits the properties (A) for all positive integral values of x, y, z . That it is not identically zero follows from the fact that the c_i inherit the condition $\sum c_i^2 = 1$.

Lemma 2

In order to prove our main theorem, we first establish a corresponding result for polynomials in two variables, and later (Lemma 3) adapt it to $P(x, y, z)$.

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

$$\alpha > 0 \text{ and } \beta > 0 \implies f(\alpha, \beta) \geq 0 \quad (B)$$

$$\alpha \leq 0 \text{ or } \beta \leq 0 \implies f(\alpha, \beta) \leq 0. \quad (C)$$

Proof:

Assume, if possible, that $f(\alpha, \beta)$ satisfies these conditions and is not identically zero. We can write

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

where $r(\alpha, \beta)$ is of degree less than N in β , and $g(\alpha)$ is not identically zero. Then we can find a value $\alpha_0 > 0$ such that $g(\alpha_0)$ and $g(-\alpha_0)$ are both non-zero. We can then find $\beta_0 > 0$ so large that

$$\beta_0^N g(\alpha_0) > |r(\alpha_0, \beta_0)|.$$

We have:

$$f(\alpha_0, \beta_0) > 0 > f(-\alpha_0, \beta_0)$$

so that

$$g(\alpha_0) > 0 > g(-\alpha_0).$$

It follows that $(-\beta_0)^N g(\alpha_0)$ and $(-\beta_0)^N g(-\alpha_0)$ have opposite signs. But this contradicts the hypothesis: $\beta < 0 \implies f(\alpha, \beta) \leq 0$, which implies:

$$f(\alpha_0, -\beta_0) = (-\beta_0)^N g(\alpha_0) \leq 0$$

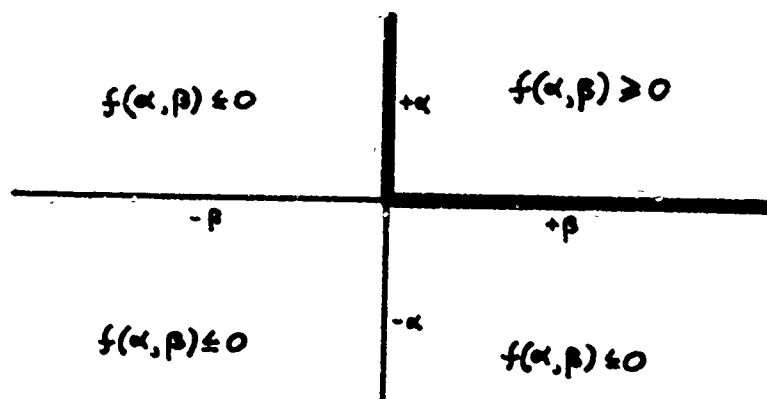
$$f(-\alpha_0, -\beta_0) = (-\beta_0)^N g(-\alpha_0) \leq 0.$$

This contradiction establishes the lemma.

4.2 A Digression on Bezout's Theorem

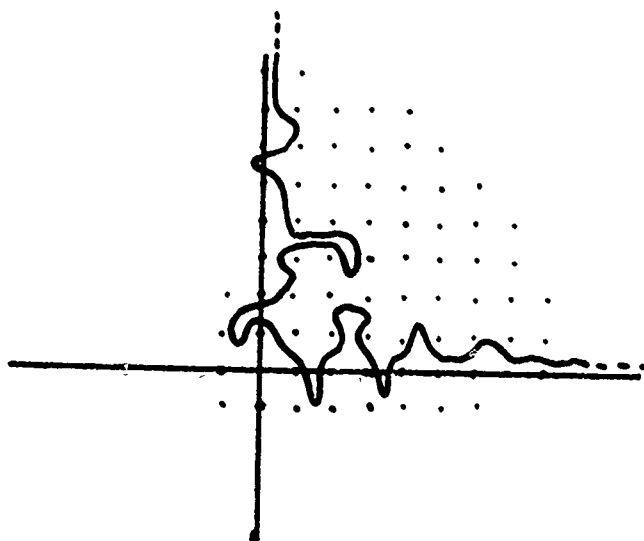
Readers familiar with elementary algebraic geometry will observe that the lemma would follow immediately from Bezout's theorem if the conditions could be stated for all real values of α and β . We would then merely have to

prove that the doubly infinite L of the figure is not an algebraic curve:

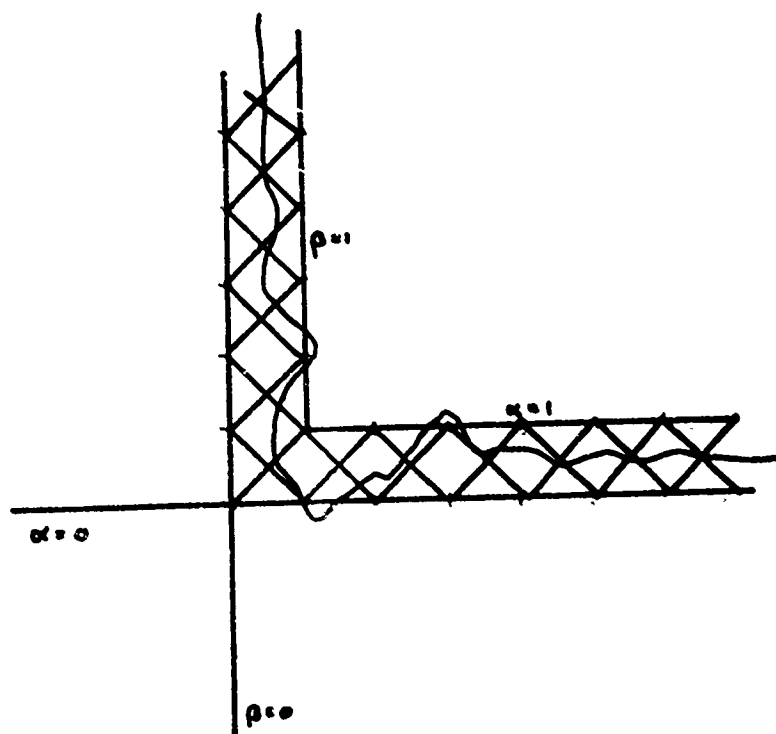


Bézout's theorem tells us that if the intersection of an algebraic curve, L , with an irreducible algebraic curve, Y , contains an infinite number of points, it must contain the whole of Y . But the L contains the positive half of the y -axis. Straight lines are irreducible, so it would have to contain the entire y -axis if it were algebraic.

Unfortunately, because our conditions hold only on integer lattice-points, we must allow for the possibility that $f(\alpha, \beta) = 0$ takes a more contorted form, for example as in the next figure:



Part of the pathological behavior of this curve is irrelevant. Since a polynomial of degree N can cut a straight line only N times, the incursions into the interiors of the quadrants can be confined to a bounded region. This means that the curve $f(\alpha, \beta) = 0$ must "asymptotically occupy" the parts of the "channel" illustrated by



It seems plausible that a generalization of Bezout's theorem could be formulated to deduce from this that the curve must enter the negative halves in a sense that would furnish an immediate and more illuminating proof of our lemma. We have not pursued this conjecture, but believe it indicates a valuable direction for future research.

Lemma 3

If a polynomial $P(x,y,z)$ satisfies the following conditions for all positive integral values of x, y , and z , then it is identically zero.

$$x > z \text{ and } y > z \implies P(x,y,z) \geq 0$$

$$x \leq z \text{ or } y \leq z \implies P(x,y,z) \leq 0.$$

Proof:

Suppose that $P(x,y,z)$ had these properties, but were not identically zero. Define $Q(\alpha, \beta, z) \equiv P(z + \alpha, z + \beta, z)$ and write

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

where r is of degree less than M in z , and $f(\alpha, \beta)$ is not identically zero. Then we can show that f must satisfy the conditions in Lemma 2: Choose any α_0 and β_0 for which $f(\alpha_0, \beta_0) \neq 0$. Choose a sufficiently large positive z_0 , for

- (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 $f(\alpha_0, \beta_0) \geq 0 \iff Q(\alpha_0, \beta_0, z_0) \geq 0$,

i.e., if and only if $P(z_0 + \alpha_0, z_0 + \beta_0, z_0) \geq 0$.

Thus $\alpha_0 > 0$ and $\beta_0 > 0 \implies z_0 + \alpha_0 > z_0$ and $z_0 + \beta_0 > z_0$

$\implies P(z_0 + \alpha_0, z_0 + \beta_0, z_0) \geq 0$

$\implies f(\alpha_0, \beta_0) \geq 0$

and similarly $\alpha_0 < 0$ or $\beta_0 < 0 \implies f(\alpha_0, \beta_0) \leq 0$.

But this is true for all α_0, β_0 . Thus by the Lemma 2, $f(\alpha, \beta) \equiv 0$.

It follows that $P(x, y, z)$ is of degree zero in z , which is only possible if it is identically zero.

This concludes the proof of the And/Or theorem.

INTRODUCTION TO PART II : Geometric Theory of Linear Inequalities

In Chapters 5 - 8 we will study the problem of building linear predicate machines to "recognize" patterns that are geometrically interesting. We will study chiefly two-dimensional patterns, and will ask questions like:

- (1) Is the problem of deciding whether the input figure is convex, or is connected, of finite order (in the sense of § 1.6)? This is studied in Chapter 5.
- (2) What is the smallest order of a perceptron that can recognize triangles, or circles? Studied in Chapter 6.
- (3) Can a finite-order perceptron tell when the input picture contains two figures that are congruent or similar (in the Euclidean sense). Can one determine which figures are symmetrical? See Chapter 7.
- (4) What can be done with the more restricted diameter-limited perceptrons? See Chapter 8.

Our discussion of these questions is not very systematic or thorough, because our knowledge is still based on too few well-understood particular cases. Furthermore, we are reluctant to propose any very rigid classifications of the knowledge we have obtained, because at almost every turn so far the results have been unexpected. Thus, the generally strong negative results described in Chapter 5 left us unprepared for the apparently positive results in Chapter 7, which in any practical sense are again reversed by the considerations in Chapter 9.

For a preview of the general situation, before immersion in mathematical detail, we suggest reading first the introductory sections of Chapters 5 - 8.

Geometrical Patterns

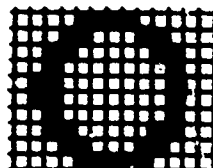
We are about to study a number of interesting geometrical predicates. But as a first step, we have to provide the underlying space $\mathbb{R}$ with the topological and metric properties necessary for defining geometrical figures; this was not necessary in the case of predicates like Parity and others related to counting, for these were 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_E 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_E . Thus to any subset X_E of E^2 corresponds the subset X of R defined by

$$X = \{x \in R \mid x \cap X_E \neq \emptyset\}.$$

Now, although X and X_E 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 $X_E \rightarrow X$. 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



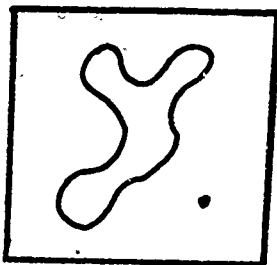
will not look very round.

If it were necessary to distinguish between E^2 and R we would say that two figures X_E, X_E' of E^2 are in the same R -tolerance class if $X = X'$. In this we would follow the general mathematical approach proposed by E.C. Zeeman [1963] for treating this kind of problem; so far, we have not had to do so. There is no problem with the translation groups play the main roles in Chapters 6, 7 and 8. There is a serious problem of handling the tolerances when discussing, as in § 7.6, dilations or rotations. The problem is even more serious when discussing general topological equivalence and it is only because we use very special restrictions on our figures, in Chapter 5, that the tolerance theory can be avoided.

CHAPTER 5: CONNECTIVITY, A GEOMETRIC PROPERTY WITH UNBOUNDED ORDER

5.0 Introduction

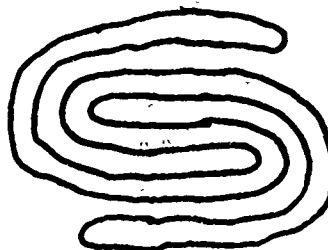
In this chapter we begin the study of connectedness. A figure X is connected if it is not composed of two or more separate, non-touching, parts. While it is interesting in itself, we chose to study the connectedness property especially because we hoped it would shed light on the more basic, though ill-defined, question of local vs. global property. For connectedness is surely global. One can never conclude that a figure is connected, from isolated local experiments. To be sure, in the case of a figure like



one would discover, by looking locally at the neighborhood of the isolated point in the lower right corner, that the figure is not connected. But one could not conclude that a figure is connected, from the absence of any such local evidence of disconnectivity. If we consider figures like the following,



and



it is difficult to imagine any local event that could bias a decision toward one conclusion or the other. Now, this is easy to prove, for example, in the narrow framework of the diameter-limited concept of local (see §0.3 and §8). It is harder to establish for the order-limited framework. But the diameter-limited case gives us a hint: by considering a particular subclass of figures we might be able to show that the problem is equivalent to that of recognizing a parity, or something like it, and this is what we in fact do.

5.1* The Connectedness Theorem

We define connectedness as follows:

Two points of R are adjacent if they are squares (in the map $X_E \rightarrow X$ 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 points from p_1 to p_2 .

Theorem: The predicate

$$\psi(X) = [X \text{ is connected}]$$

is not of finite order (§1.6), i.e., it

has arbitrarily large orders as $|R|$ grows

in size.

* The proof in §5.1 is needed for the theorem of §5.3. Otherwise the proof in §5.7 of theorem §5.1 yields a better result.

** We can't allow corner contact, as in , to be considered as connection. For this would allow two "curves" to cross without "intersecting": and not even the Jordan Curve Theorem would be true. The problem could be avoided by dividing E^2 into hexagons instead of squares!

Proof 1: Suppose that $\psi(X)$ could have order $< m$. Consider an array of squares of R arranged in $2m + 1$ rows of $4m^2$ squares each.

row B_1								
row B_2	b_{21}							
row B_3								
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
								
row B_{2m}								
row B_{2m+1}								

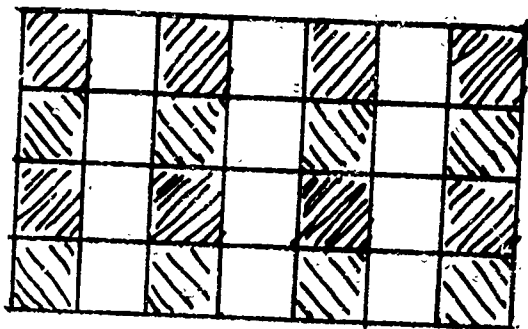
Let G_0 be the set of points shaded in the diagram; that is, the array of points whose row indices are odd, and let G_1 be the remaining squares of the array. Let $\mathcal{F}$ be the family of figures obtained from the figure G_0 by adding subsets of G_1 , i.e., $F \in \mathcal{F}$ if it is of the form $G_0 \cup F_1$, 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 of §3.2. 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 $\mathcal{F}$ as a subset of all possible figures on R . Clearly, if we had an order- k predicate ψ_{CON}^k that could recognize connectivity on R , we could have one that works on $\mathcal{F}$; namely the same predicate with constant zero inputs to all variables not in $G_0 \cup G_1$. And since all points of the odd rows have always value 1 for figures in $\mathcal{F}$, 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 unity inputs to the points in G_1 . Thus each Boolean function of the original predicate ψ_{CON}^k is replaced by the function obtained by fixing certain 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 original ψ_{CON}^k . This proof shows that ψ_{CON} has order at least $C \cdot |R|^{1/3}$. In 5.7 we show it is at least $C \cdot |R|^{1/2}$.

5.2 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
box.

Now consider a ψ of degree 2; we will show that it cannot characterize the connectedness of pictures of this kind. Suppose that $\psi = [\sum \alpha_i \phi_i > 0]$ and consider the equivalent form, symmetrized under the full group of permutations that interchange the rows and permute within rows*. 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)

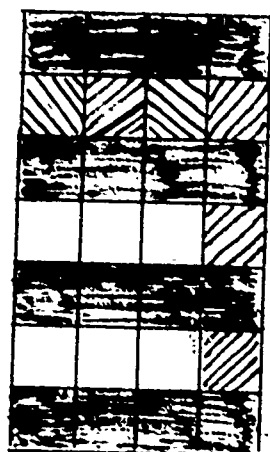
hence any order 2 predicate must have the form

$$\psi = \alpha_1 n^1(X) + \alpha_{11} N^{11}(X) + \alpha_{12} N^{12}(X) > 0 \quad (1)$$

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

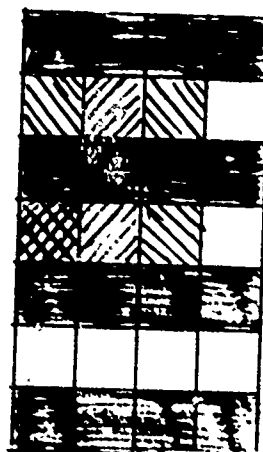
Now consider the two figures:

* Note that this is not the same group used in proving theorem §3.2. There we did not use the row-interchange part of the group.



X_1

CONNECTED



X_2

DISCONNECTED

In each case one counts:

$$N^1 = 6$$

$$N^{11} = 6$$

$$N^{12} = 9$$

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 $|A_1| = 4$, while the general proof required $|A_1| = 4m^2 = 36$. (It is known also that if $k=6$, we can get a similar result with $|A_1| = 16$. This was shown by Dona Strauss.)

The case of $k=2$, $m=3$, $|A_1| = 3$ is of order 2, since one can in fact express the connectivity predicate for that space as

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

(This was found by brute force)

The proof method used in this example is an instance of use of what we call the "geometric n-tuple spectrum," and the general principle is further developed in Chapter 6.

5.3 Slice-wise Connectivity

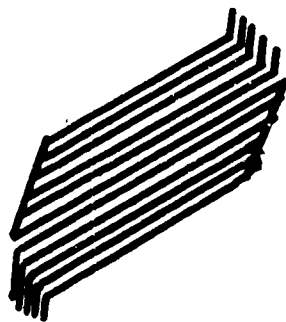
It should be observed that the proof in §5.1 applies not only to the property of connectivity in its classical sense but to the stronger predicate defined by:

A figure X is "slice-wise disconnected" if there is a straight line L such that:

X 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.

It is fairly clear that human ability to discern connectivity is limited, if the available time is restricted, suggesting a non-parallel process. Thus it takes a certain time to decide which of these figures are connected, even in the simple cut-wise sense:



5.4 Reduction of One Order Problem to Another

The study of the order of predicates is often facilitated by the reduction of a given predicate to an order 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 sub-perceptron system of P if the basic set R' is a sub-set of R and if its set of predicates Φ' is that obtained by relativising the members of Φ to R' , i.e., each predicate $\phi \in \Phi'$ satisfies

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

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 $f(X) = \{p \in R' \mid q \in R; f(q) = p\}$).

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

$$p \in R' \implies f(p) \subset R$$

$$p \neq q \implies f(p) \cap f(q) = \Lambda.$$

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 5.4.1: Collapsing Theorem:

If f is a collapsing map from R to R' and ψ' is obtained from ψ by f , then the order of ψ' is greater than that of ψ .

Proof: Let $\psi = \sum_{\phi \in \hat{\Phi}} \alpha_{\phi} \phi > 0$ where $\hat{\Phi}$ is the set of masks of degree less than k on r .

Now for any $X' \subset R$,

$$\begin{aligned} \psi'(X') &= \psi(f(X')) \\ &= \left[\sum_{\phi \in \hat{\Phi}} \alpha_{\phi} \phi(f(X')) \right] > 0. \end{aligned} \tag{1}$$

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

$$S(\phi) \not\subset f(R') \text{ then } \phi(f(X')) = 0 \text{ for all } X' \subset R.$$

Now for $\phi \in \hat{\Phi}'$ 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,

$$X' \supset \{p \mid f(p) \cap S(\varphi) \neq \Lambda\}$$

$$\Rightarrow f(X') \supset \bigcup \{f(p) \mid f(p) \cap S(\varphi) \neq \Lambda\} \supset S(\varphi),$$

$$\text{i.e., } X' \supset \{p \mid f(p) \cap S(\varphi) \neq \Lambda \Rightarrow f(X') \supset S(\varphi) \Rightarrow \varphi(f(X'))\}.$$

On the other hand, if $\varphi(f(X'))$, i.e., $f(X') \supset S(\varphi)$, it follows that

$$f(p) \cap S(\varphi) \neq \Lambda \Rightarrow p \in X'$$

since $f(p) \cap f(q) = \Lambda$ for $p \neq q$.

$$\text{Thus } \varphi(f(X')) = [X' \supset \{p \mid f(p) \cap S(\varphi) \neq \Lambda\}].$$

In other words $\varphi(f(X'))$ is a mask on R' with support

$$\{p \mid f(p) \cap S(\varphi) \neq \Lambda\}.$$

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

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

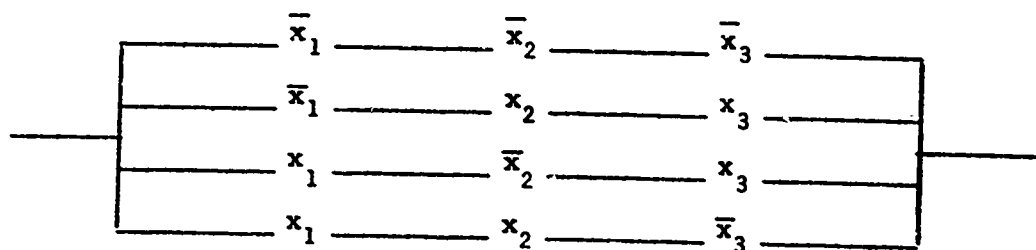
Going back to equation (1) we see, then, that ψ' is represented as a linear function of masks of degree less than k .

q.e.d.

5.5 Huffman's Construction for ψ_{CON} : Proof 2 of Theorem 5.1

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 $n = 3$.



The interpretation of the symbols x_i and $\bar{x}_i$ is as follows: when x_i is in the "on" position contact is made whenever x_i appears, and broken whenever $\bar{x}_i$ appears; when x_i is in the "off" position contact is made where $\bar{x}_i$ 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 present case can be written:

$$\psi_{\text{PAR}}(x_1, x_2, x_3) = \bar{x}_1 \bar{x}_2 \bar{x}_3 \vee x_1 x_2 \bar{x}_3 \vee \bar{x}_1 \bar{x}_2 x_3 \vee x_1 x_2 x_3$$

(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 reduction argument, in intuitive form, is as follows: the Huffman switching net can be regarded as defining a class $\mathcal{F}$ of geometric figures which are connected or not depending on the parity of a certain set, the set of switches that are 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 can use the collapsing theorem to formalize this argument. But before doing so we note that a certain price will be 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 as $\log |R|$. This gives a weaker lower bound, $\log |R|$ compared with $|R|^{1/3}$, if we wish to estimate the order. In fact, in order to escape this penalty we

have decided not to apply the collapsing theorem after all to this case; instead we shall construct a related but more complex switching net to obtain a sharper bound.

5.6 Connectivity on a Toroidal Space $|R|$

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

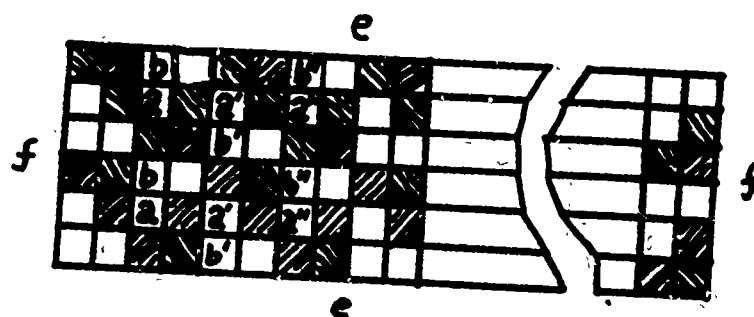


Figure 5.6.1

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 $X \in \mathcal{F}$
- (ii) for each $X \in \mathcal{F}$ and each i , either both points marked a_i or both points b_i are in $\mathcal{F}$, but no other combinations are allowed.

Then it can be seen, for each $X \in \mathcal{F}$, that X has either one connected component 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 Theorem §3.1.1, we find that ψ_{CON} has order $\geq \frac{1}{12} |R|$.

The idea for this proof came from the attempt to reduce connectivity

to parity directly by representing the switching diagram:

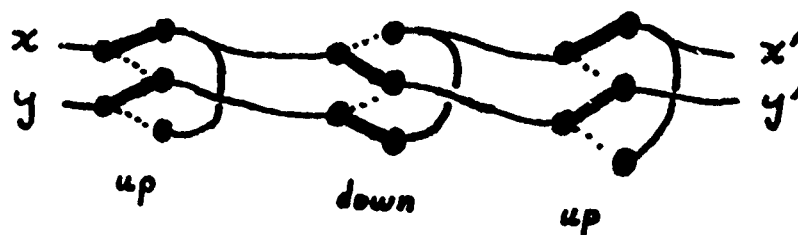


Figure 5.6.2

If an even number of switches are in the "down" position then x is connected to x' and y to y' . If the number of down switches is odd, x is connected to y' and x' to y . This diagram can be drawn in the plane (see §5.7) by bringing the vertical connections around the end; then one finds that the predicate $[x \text{ is connected to } x']$ has for order some constant multiple of $|R|^{1/2}$. If we put the toroidal topology on R , the order can be shown to be greater than a constant multiple of $|R|$; this is also true for a 3-dimensional Euclidean R . These facts strongly suggest that our bound for the order of ψ_{CON} is too low in the 2-dimensional plane case. The following section improves the situation somewhat by replacing $|R|^{1/3}$ by $|R|^{1/2}$.

5.7 Reduction of ψ_{CON} to ψ_{PAR} in the Plane

The following construction shows that the order of ψ_{CON} is $\geq O(|R|^{1/2})$ for two-dimensional plane figures. It results from modifying Figure 5.6.2 so as to connect x to x' . This is easy for the torus, but for a long time we thought it was impossible in the plane.

We first define a "4-level switch" to be a pair of figures with the following two connection diagrams.

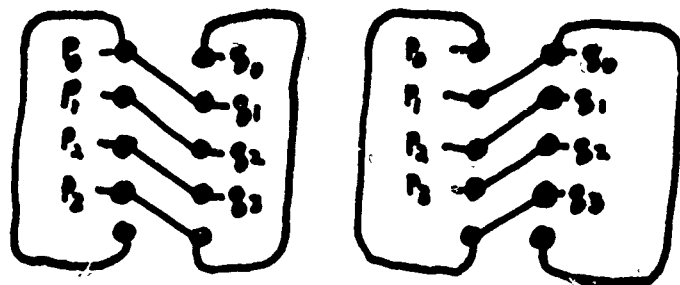


Figure 5.7.1

In the "down" state we have

p_0 connected to q_1

p_1 connected to q_2

p_2 connected to q_3

p_3 connected to q_0

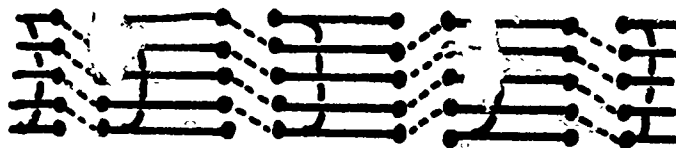
and we write

$$p_i \Rightarrow q_{(i+1) \bmod 4}.$$

With the switch in the "up" state we have, similarly,

$$p_i \Rightarrow q_{(i-1) \bmod 4}.$$

Then, changing a switch will have the effect of adding $2(\text{mod } 4)$ to the index of the q connected to any given p , because subtracting 2 has the same effect as adding 2. Now consider the effect of n switches in cascade:



This simply iterates the effect: each switch that is "down" adds 1 to the q -index and each "up" switch subtracts $1(\text{mod } 4)$ so that if k switches are down we have

$$P_i \Rightarrow q_{i+k-(n-k)\text{mod } 4}$$

Then that there are only two possible mappings since if an even number of switches are down we have

$$P_i \Rightarrow q_{i-n(\text{mod } 4)}$$

and if an odd number are down we have

$$p_i \Rightarrow q_{i+2-n(\bmod 4)}$$

Finally, we add fixed connections tying together p_1 , p_2 and p_3 and q_{1+n} , q_{2+n} , and q_{3+n} . The overall effect in the two cases is then either



Figure 5.7.3

and we see that in the one case the network is disconnected and in the other case it is connected. We illustrate the $n = 6$ case:

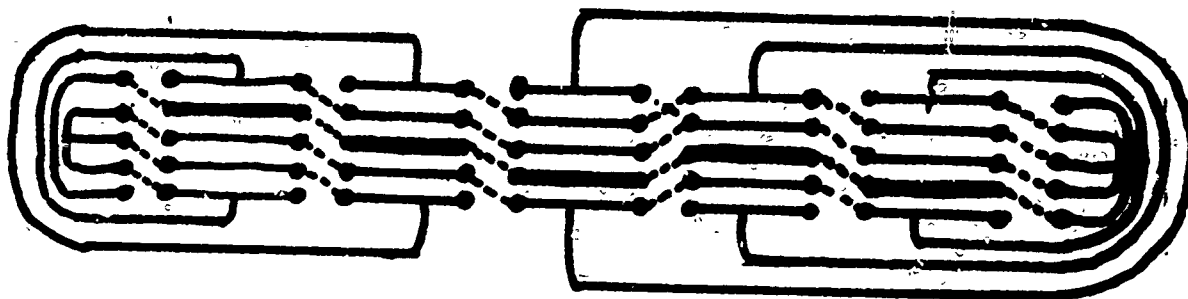
(See Figure 5.7.4, p. 16c.)

and we can state:

Theorem 5.7.1: This network is connected when an odd number of switches are down and disconnected when an even number are down.

It remains only to realize the construction of the switches. Define a switch to be the two configurations:

-16c-



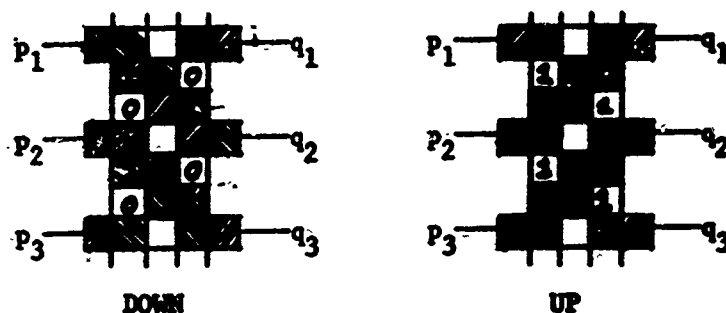


Figure 5.7.5

Remember that  is not a connection. When the entire construction is completed, for n switches, the network will be about $5n$ squares long and about $2n + 12$ squares high, so that the number of switches can grow proportionally to $|R|^{1/2}$. It follows that the order of t_{ON} grows at least as fast as $|R|^{1/2}$.

The idea for the proof comes from observing that in the planar version of Figure 5.6.2

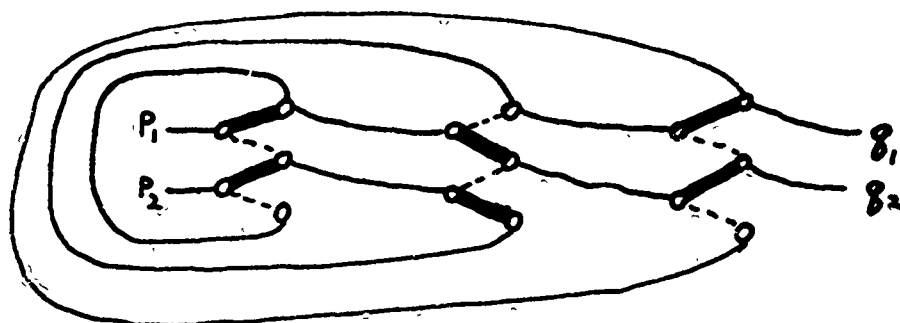


Figure 5.7.6

we have $p_1 \leftrightarrow q_1$ and $p_2 \leftrightarrow q_2$ for one parity and $p_1 \leftrightarrow q_2$ and $p_2 \leftrightarrow q_1$ for the other. If we could make a permanent additional direct connection from p_1 to q_1 then the whole net would be connected or disconnected according to the parity. But this is topologically impossible, and because the construction appeared incompleteable we took the long route through proving and applying the One-in-a-box theorem. Only later did we realize that the $p_1 \leftrightarrow q_1$ connection could be made "dynamically," if not directly, by the construction in Figure 5.7.1. This figure is made by superimposing two copies of Figure 5.7.2, and using the second copy only to insure that p_2 and q_2 are always connected in the first copy.

5.7.2 The Order of ψ_{CON} As a Function of $|R|$

What is the true order? Let us recall that at the root of the proof methods we used, was the device (§5.0) of considering not all the figures but only special subclasses with special combinatorial features. Thus even the order $|R|/12$ of §5.6 is only a lower bound. Our suspicion is that the order cannot be less than $|R|/2$. As for the number of ϕ 's required, Theorem 3.1.2 and the toroidal results give us $\geq 2^{|R|/12}$, but this too, is a low bound, and one suspects that nearly all the masks are needed. Another line of thought suggests that one could get by with the order of the number of connected figures, but that has probably not much smaller exponent. As for the coefficients, the results of §9 will apply immediately.

Examination of the toroidal construction in §5.6 might make one suspect that the result, $\psi_{\text{CON}} \geq \frac{1}{12} |R|$ is an artifact resulting from the use of a long, thin torus. Indeed, for a "square" torus we could not get this result because of the area that would be covered by the connecting bridge lines. This clouds the conclusion a little. On the other hand, if we consider a three dimensional R , then there is absolutely no difficulty either in the torus or in ordinary E^3 of showing that $\psi_{\text{CON}} \geq \frac{1}{K} |R|$. We leave unresolved the problem of finding precisely the lower bound of ψ_{CON} in E^2 , content with showing that it is not of finite order, and that it grows at least as fast as $|R|^{1/2}$.

5.8 Predicates Related to the Euler Formula

Curiously enough the predicate

$$\lceil X \text{ is connected} \rceil \vee \lceil X \text{ contains at least one hole} \rceil$$

has finite order, even though neither disjunct does--an instance of the opposite of the And/Or phenomenon. This will be shown by a construction involving the Euler relation for orientable geometric figures.

5.8.1 The Euler Polygon Formula

Two-dimensional objects have a topological invariant $G(X)$ that in polygonal cases is given by

$$G(X) = |Faces(X)| - |Edges(X)| + |Vertices(X)|.$$

Some examples:

$G = +1$				
$G = 0$				
$G = -1$				

Topologically, $G(X)$ is in general given by

$$G(X) = |\text{connected components}| - |\text{holes}|.$$

It is possible to make low-order perceptrons that realize predicates like $\lceil G(X) = n \rceil$ and $\lceil G(X) < n \rceil$ as follows.

For each point x_i of K choose weight

$$\alpha_i = +1 \quad (\text{"vertices"}).$$

For each adjacent pair , choose weight

$$\alpha_{ij} = -1 \quad (\text{"edges"}).$$

For each "square" , choose weight

$$\alpha_{ijkl} = +1 \quad (\text{"faces"})$$

For all other masks, choose weight

$$\alpha = 0.$$

Then it is claimed that

$$G(X) = \sum x_i - \sum x_i x_j + \sum x_i x_j x_k x_l.$$

We sketch the proof by (inductive) analysis of what happens when a point (square in K^2) is added to a figure.

Adding an adjacent square to a figure, that touches on only one side does not change G , and adds $1 - 1 = 0$ to our sum;



Adding a square that touches two others normally decreases G by unity, and appropriately adds $1 - 2 = -1$ to our sum:

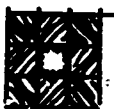


This normally connects two components or (if the two were already connected) adds a hole. The exception is:



which does not change G and appropriately adds $1 - 2 + 1 = 0$ to the sum.

Case-analyses of the 3-neighbor and 4-neighbor situation complete the proof: these include partial fills like



which add $1 - 3 + 2 = 0$ and $1 - 4 + 4 = 1$, the latter representing the increase in G when a hole is finally filled-in. All this corresponds to an argument in algebraic topology concerning addition of edges and cells.

to chain-complexes.

It follows immediately that the predicate

$$\lceil G(X) < n \rceil$$

is realized with order ≤ 4 . This leads to some curious observations: If we are given that the figures X are restricted to from the connected (= one-component) figures then an order-4 machine can recognize

$$\lceil X \text{ is simply-connected} \rceil = \lceil G(X) > 0 \rceil$$

or

$$\lceil X \text{ has less than 3 holes} \rceil = \lceil G(X) > -2 \rceil.$$

But of course we cannot conclude that these can be recognized unconditionally by a finite order perceptron.

Note that this topological invariant is thus seen to be highly "local" in nature--indeed all the φ 's satisfy (a very tight) diameter-limitation! Now returning to our initial claim we note that

$$\lceil G(X) = n \rceil \equiv (\lceil G(X) \leq n \rceil \equiv \lceil G(X) \geq n \rceil)$$

hence by Theorem 1.5.4 we can conclude that $\lceil G(X) = N \rceil$ has order ≤ 8 . But the proof of that theorem involves constructing product- φ 's that are not

diameter-limited, and we believe that this predicate cannot be realized by diameter-limited perceptrons. The situation seems similar to the relation between §8.2 and §8.9, but we have not explored it. If the conjecture is true, we have the intriguing possibility of using the theory to classify topological invariants into different categories of "localness." What would be the analogy, if any, within topology, of the distinction between diameter- and order-limited?

5.8.2 Uniqueness of Topological Invariants

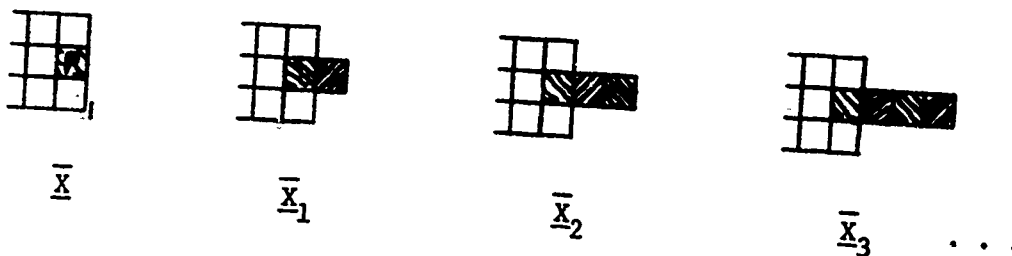
We have shown that the topological invariant function $G(X)$ lies in the class

$$L(\{x_i, x_i x_j, x_i x_j x_k x_l\}) = L(\Phi_0)$$

where Φ_0 contains the masks enumerated in 5.8.1. We now show that, assuming bounded coefficients (see §9.)::

Theorem 5.8.2: The only topological invariants in $L(\Phi_0)$ are functions of $G(X)$.

Proof: Consider any figure X and let p be one of its points maximally to the right. Then the sequence of figures



are topologically equivalent.

Suppose that $\psi \in L(\mathbb{Z}_0)$ is topologically invariant. Since this includes translation and rotation (by 90°) invariance, we can write

$$\psi = [\alpha \Sigma p_{\square} + \beta \Sigma p_{\blacksquare} + \gamma \Sigma p_{\boxplus} > 0].$$

Define $f(n)$ to be the sum of the summations for figure X_n . Since the direction of the inequality must be the same for all X_n we must have, for all n ,

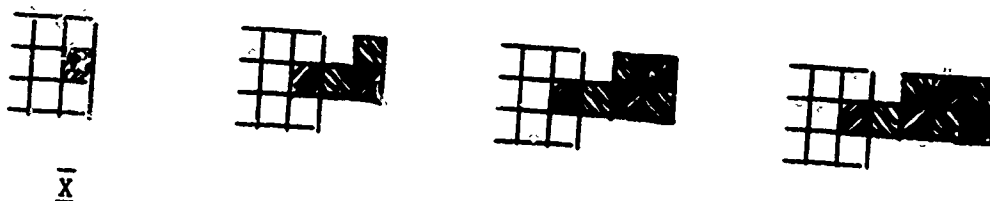
$$f(X) + n(\alpha + \beta) > 0$$

and

$$f(Y) + n(\alpha + \beta) < 0$$

for any figures X and Y for which $\psi(X)$ and $\sim \psi(Y)$. But then we must have $\alpha + \beta = 0$, because otherwise ψ would be trivial (constant).

Similarly, by appending to X the figures



$\bar{X}$

one finds that $2\alpha + 3\beta + \gamma = \beta + \gamma = 0$ and we conclude that

$$\alpha + \beta = \gamma$$

and hence that

$$\psi = \left[G(X) > \frac{\theta}{\alpha} \right]:$$

Hence $L(\phi_0)$ contains only the Euler invariants!

Q.E.D.

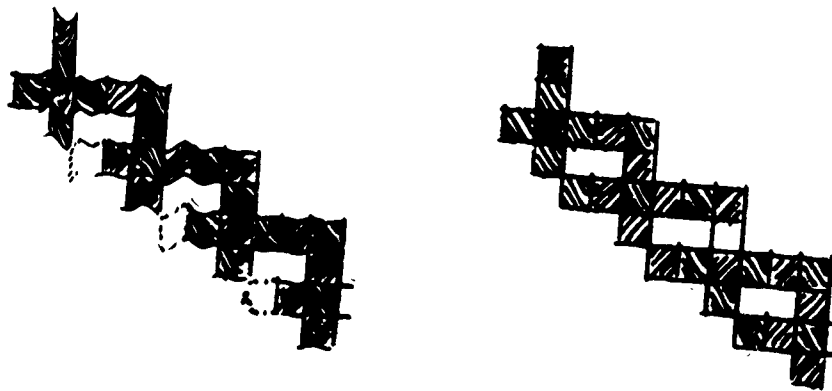
We can push the argument into larger ϕ 's. Suppose that the masks of the form  are adjoined to ϕ_0 . Then by using the figure extensions



we find that $2\alpha + 2\beta + \delta = \delta = 0$, hence the new masks will have zero coefficient in ψ . Even if we further adjoin the mask , we must then* have (from the same extensions)

$$\alpha_{\square} + \alpha_{\blacksquare} + \alpha_{\blacklozenge} + \alpha_{\blacksquare} = 0 = \alpha_{\blacksquare} + \alpha_{\blacksquare}$$

and a more complex sequence of extensions, comparing



shows that $\alpha_0 + \alpha_1 + \alpha_2 = \alpha_3 = 0$ and hence
 that α_0 and hence α_1 are both zero (else α_2 would
 change connectivity, contradicting the definition in §5.1).

All this suggests a conjecture we have not investigated at all; that
 the only realizable topological invariants of finite order are of
 Eulerian type. If this were proved, it would directly imply the connectedness
 theorem, though it wouldn't give a magnitude estimate on order-growth.

* There are many other topological invariants besides the number of components
 of X and $G(X)$, for example, a component of X lies within a hole within another
 component of X .

CHAPTER 6: GEOMETRIC PATTERNS OF SMALL ORDER

6.0

Chapters 6 and 7 will demonstrate some techniques for constructing low-order predicates of geometric character. The results are in one sense more positive than those of the previous two chapters. We were frequently surprised to find that certain predicates are of much lower order than we originally expected. However, there are also some negative results of a new kind. In §6.6 we shall face the problem of recognizing patterns "in context." A low-order perceptron can tell, for example whether a given figure is a square (as shown in §7.2.5) but the problem of deciding whether a figure contains a square (and perhaps something else) is not of finite order!

In Chapter 7 we describe a very powerful technique, "stratification," that shows how to construct finite-order functions for many patterns invariant under important geometric groups, e.g., translation and dilatations. But this method seems to give rise to extremely large coefficients. In Chapter 9 we shall see that the occurrence of coefficients which increase without bound, as the retina size increases, is not an accidental by-product of this method of construction; it will be shown that predicates as simple as recognizing a single given figure independently of translation on the retina necessarily lead to unbounded coefficients in any realization by a family of perceptrons of limited order.

The division of material between Chapters 6 and 7 corresponds formally to two general methods we have found for constructing geometric predicates of finite order; "difference-vector spectra" (Chapter 6) and "Stratification" (Chapter 7). A deeper difference is related to the status of group invariance

in the two cases.

The geometric predicates we want to consider are Δ -invariant under the group of Euclidean transformations, and sometimes we want invariance under size-dilatation as well. Unlike the groups of permutation used up to now, dilatation is not easy to define in the context of finite quantized retinas. The difficulties are, at least superficially, of two kinds: those that come from discreteness and those that seem related only to finiteness. An extreme example of the first kind is posed by the group of dilatations of plane figures. We can double the size of a figure on a discrete retina, but we may be prevented by the raster size from halving it. Worse problems of the same kind occur when we consider rotations (other than those that preserve the retinal lattice). Problems of the second kind, which we shall treat at some length, arise in connection with the more innocuous group of translations.

We anticipate the details of Chapters 6 and 7 by a brief summary of the main results. In Chapter 6 we use results that, in some sense, depend only on the "local" structure of the group. The theorems we prove then remain true whether R is the full infinite plane, or if we force it to be finite, e.g., as we did in §5.6 by cutting-out a finite portion and sewing its edges together with a toroidal connection. In all cases, the Group Invariance theorem is applicable, and the coefficients are equal on equivalent φ 's, etc. The price seems to be that in each case we are accepting a diameter-limitation, either in the predicates or in the class of figures to be accepted. (Thus we recognize the set of geometric rectangles in §6.3.2, by their "conjunctively local" diameter-limited property of having only four corners. But we cannot recognize geometric squares (presumably) without the methods of Chapter 7.)

In Chapter 7 we use more global properties of the transformation group; in particular, that all the translations of the plane can be ordered in an enumeration under which sufficiently large elements always dominate sufficiently small ones. These enumerations enable us to show that some less local properties can be realized on the full infinite plane: the price is unbounded coefficients and loss of equal coefficients for equivalent predicates. These procedures do not survive toroidal closure of a finite portion of the plane because the required property of the ordering is destroyed, and this appears to be irreparable if the group contains cyclic groups in its global structure.

The predicates and methods of representation of Chapter 6 depend only on the local structure of the group. In the next chapter we use more global properties of the transformation group; for example, we will use a complete ordering of all translations to obtain low-order representations for certain predicates. This result doesn't carry over to rotation because these procedures do not survive a toroidal closure of a finite part of the plane. (The ordering relation is destroyed!) This seems irreparable because the group now contains a cyclic subgroup in its global structure. Indeed we find (in §7.10) that certain simple translation-invariant predicates do not satisfy the group invariance theorem on the infinite plane. It will turn out that we can use this to advantage. We eventually shall show (in §9.4) that the G.I. theorem doesn't hold if bounds can be placed on the coefficients, and so deduce that the delinquent predicates cannot possibly be represented with bounded coefficients. But despite this consolation prize we feel deeply dissatisfied

with our state of understanding of the inter-relations of these phenomena. We believe there is some deeper fact connected with the global structure of the group, but we have been unable to guess precisely what it is. Thus, although the next two chapters contain many amusing constructions and some intriguing theorems, we see them as posing more questions than they answer.

Because translation-invariance is required throughout Chapter 6, theorem 9.4 assures us that we can use the group-invariance theorem provided we assume that the coefficients are bounded in each equivalence-class. This is always (trivially) true for finite R . But there will then exist infinite preceptron counter-examples to some of the statements in Chapter 6; for example, §7.10 discusses an infinite exception to statements in §6.2. We felt that it was desirable to leave the results of Chapter 6 in their present form, even after discovering the methods of Chapter 7, because in real life, the conclusions based on the finite case are the most important, however interesting the infinite case might be, mathematically.

In §6.1--§6.4 we show that certain patterns have orders $= 1$, $= 2$, ≤ 3 , ≤ 4 respectively. In most cases we usually have not established the lower bound on the orders and have no systematic methods for doing so.

F.1 Geometric Patterns of Order 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 treat alike the figures that are "similar" in Euclidean Geometry. For order 1 we know that all

coefficients can be assumed to be equal, since the translation group satisfies the condition for Theorem 2.4. Therefore, the only patterns that can be of order 1 are those defined by a single cut in the cardinality or area of the set:

$$\psi = [|X| > A] \text{ or } \psi = [|X| < A].$$

Note: If translation invariance is not required, then perceptrons of order 1 can compute other properties, e.g., concerning moments about particular points or axes. However, these are not "geometric" in the sense of being suitably invariant so while they may be of considerable practical importance, we will not discuss them here*.

6.2 Patterns of Order 2, Distance Spectra

For $k = 2$ things are more complicated. As shown in §1.4, ex. iii, it is possible to define a double cut, or segment, in the area of the set; that is, we can do the counting trick, and recognize the figures whose areas are

$$\psi = [A_1 < |X| < A_2].$$

In fact, in general we can always find a function of order 2 k that recognizes the sets whose areas lie in any of k intervals.) But let us return to patterns

* See, e.g., McCulloch and Pitts (1947) for an eye-centering servomechanism.

invariant under geometric groups. First, consider only the group of translations, and masks of order 2. Then two masks $x_1 x_2$ and $x_1' x_2'$ are equivalent if and only if the difference vectors

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

are equal, possibly with opposite sign. Thus, with respect to the translation group, any order 2 predicate can depend only on a figure's "difference-vector spectrum," defined as the sequence of the numbers of pairs of points separated by each angle and distance pair. The two figures:



and



have the same difference-vector spectra, i.e.,

<u>"vector"</u>	<u>number of pairs</u>
0	4
	1
	2
	1
	1
	1

Hence no order 2 predicate can make a classification which is both translation invariant and separates these two figures. In fact, an immediate consequence of the group invariance theorem is:

Theorem: Let $\psi(X)$ be a translation-invariant predicate of order 2. Define $n(v)$ to be the number of pairs of points in X that are separated by the vector v . Then $\psi(X)$ can be written

$$\psi = \lceil \sum_v \alpha_v n(v) > \theta \rceil.$$

Proof: $n(v)$ predicates in the class $\varphi(v)$ are satisfied by any translation of X .

Corollary^{*}: Two figures with the same translation spectrum $n(v)$ cannot be distinguished by a translation-invariant order 2 perceptron.

Example: the figures



and



are indistinguishable, while

* Conversely if the spectra are different, e.g., $n(v_1)(A) < n(v_1)(B)$ then the translations of two figures can be separated with $\lceil n(v_1)(X) < n(v_1)(A) \rceil$. But classes made of different figures may not be separable.

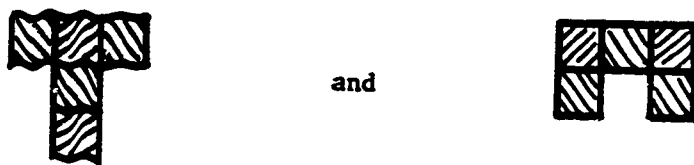


and

have different difference-vector spectra, and can be separated.

If we add the requirement of invariance under rotation, the last pair above becomes indistinguishable, because the equivalence-classes now group together all differences of the same length, whatever their orientation*.

An interesting pair of figures rotationally distinct, but still indistinguishable, for $k = 2$, is the pair



and

which have the same (direction-independent) distance-between-point-pair spectra through order 2, namely:

* 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. There is of course no difficulty for the 90° rotations, the only rotation group used here.

$|x_i - x_j| = 1$ from 4 pairs



$|x_i - x_j| = \sqrt{2}$ from 2 pairs



$|x_i - x_j| = 2$ from 2 pairs



$|x_i - x_j| = \sqrt{5}$ from 2 pairs



and each has 5 points (the order 1 spectrum).

The group-invariance theorem, §2.3, tells us that any group-invariant perceptron can not distinguish between members of equivalence classes of masks, but can depend only on the pattern's "occupancy numbers," i.e., exactly the "geometric spectra" discussed here. Many other proposals for "pattern-recognition machines"--not perceptrons, and accordingly not representable simply as linear forms--might also be better understood after exploration of their relation to the theory of these geometric spectra. But it seems less likely that this kind of analysis would bring a great deal to the study of the more "descriptive" or, as they are sometimes called, "syntactic" scene-analysis systems that the authors secretly advocate (Chapter 13).

6.3 Patterns of Order 3

6.3.1 Convexity

A particularly interesting predicate is

$$\text{CONVEX}(X) = [X \text{ is a single, solid convex figure}].$$

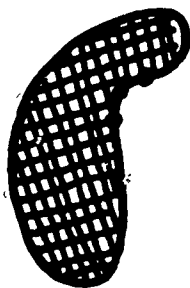
That this is of order ≤ 3 can be seen from the definition of "convex": X is convex if and only if any line-segment whose end points are in X lies entirely within X . Given a suitable definition of tolerance approximation, it follows that X is a convex if and only if

$$a \in X, b \in X \Rightarrow \text{midpoint}([a, b]) \in X$$

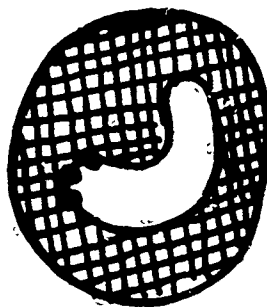
hence

$$\text{CONVEX}(X) = [\sum_{a, b} [a \in X \wedge b \in X \wedge \text{mid}([a, b]) \notin X] < 1]$$

is of order ≤ 3 , and presumably of order $= 3$. This is a "conjunctively local" condition of the kind discussed in §0.2. Note that if a connected figure is not convex one can further conclude that it has at least one "local" concavity, as in

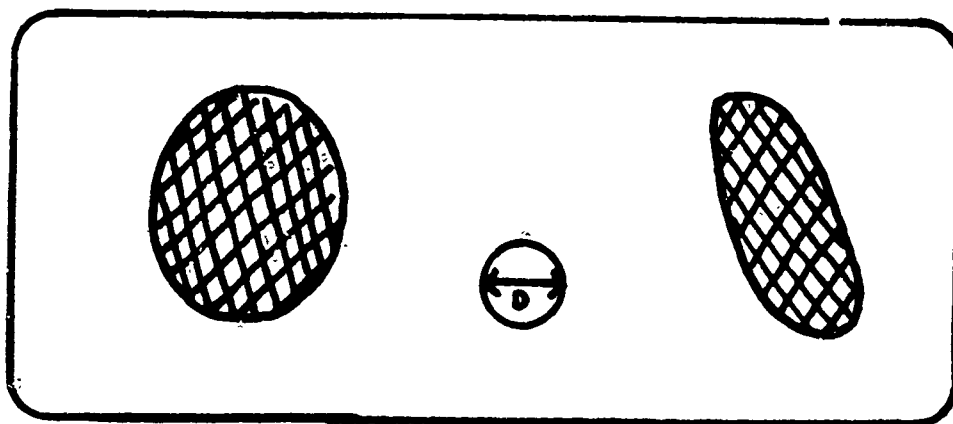


or



with the three points arbitrarily close together. Thus, if we are given that X is connected, then convexity is also diameter-limited order 3.

If we are not sure X is connected, then the preceding argument fails in the diameter-limited case because a pair of convex figures, widely separated, will be accepted:



Indeed, convexity is probably not order 3 diameter-limited, but one should not jump to the conclusion that it is not diameter-limited of any order, because of the following "practical" consideration:

Even given that a figure is connected, its "convexity" can be defined only relative to a precision of tolerance. If this precision is not infinite, and it cannot be in the diameter-limited case, then either there will be a bound on the size of the acceptable figures, or some small negative curvature will have to be tolerated. But within this constraint, one can approximate an estimate of curvature, and define "convex" to be $\int \text{curvature} \leq 4\pi$. We will discuss this further in §8.3.

6.3.2 Rectangles

Within our square-array formulation of $E^2 \rightarrow R$ we can define with order 3 the set of solid axis-parallel rectangles. This can even be done with

diameter-limited φ 's, by

$$[\sum \varphi_{\text{hollow}} + \sum \varphi_{\text{solid}} \leq 4]$$

where all φ 's equivalent under 90° rotation are included. The hollow rectangles are caught by

$$[2\sum \varphi_{\text{hollow}} + \sum \varphi_{\text{solid}} \leq 12]$$

where the coefficients are chosen to exclude the case of two or more separate points. These examples are admittedly weakened by their dependence on the chosen square lattice, but they have an underlying validity in that the figures in question are definable in terms of being rectilinear with no more than four corners, and we will discuss this slightly more than "conjunctively-local" kind of definition in Chapter 8.

One would suppose that the sets of hollow and solid squares would have to be of order 4 or higher, because the comparison of side-lengths should require at least that. It is surprising, therefore, to find that they have order 3. The construction is distinctly not conjunctively-local, and we will postpone it to Chapter 7, even though it satisfies the bounded-coefficient stipulation for the present chapter.

Another example of an order 3 predicate is

$$[X \text{ lies within a line and has } \leq n \text{ segments}]$$

which can be defined, up to a tolerance, by

$$|\sum \varphi_{\square} + \sum \varphi_{\square} - \sum \varphi_{\square} + n \sum (\text{all non-collinear triples}) \leq n|$$

6.3.3 Higher-order Translation Spectra

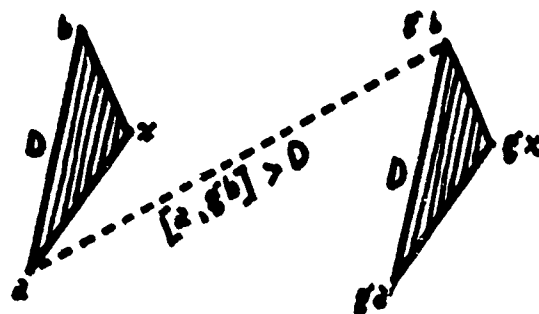
If we define the 3-vector spectrum of a figure to be the set of numbers of three-point masks satisfied in each translation equivalence-class, it is interesting to note the following fact (which is about geometry, and not about linear separation).

Theorem 6.3.3: Figures are uniquely characterized (up to translation) by their 3-vector spectra, even in higher dimensions.

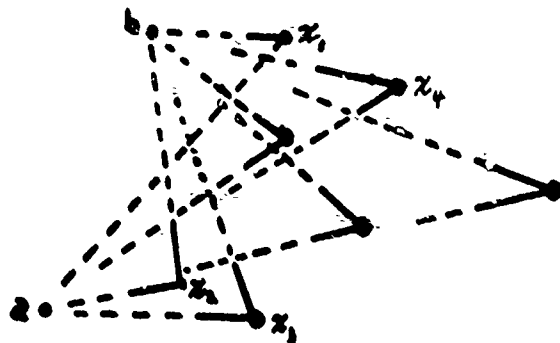
Proof: Let F be a particular figure. The figure F has a "diameter"; the maximal distance between two of its points. Choose a pair (a,b) of points of F with this distance and consider the set $\mathcal{M}_{ab} = \{\varphi_{a,b,x}\}$ of masks of order 3 that contain a,b , and any other point x of F . These masks must have coefficients equal to unity in the translation spectrum of F , for if F contained two translation-equivalent masks

$$\varphi_{a,b,x} \text{ and } \varphi_{ga,gb,gx}$$

then one of the distance $[a,gb]$ or $[ga,b]$ would exceed D , for they are diagonals of a parallelogram with one side equal to D .



Thus any translate of F must contain a unique pair parallel to (a,b) and the part of its spectrum corresponding to $\hat{\phi}_{ab}$ allows one to reconstruct completely the figure.



The fact that a figure is determined by its 3-translation spectrum, not, of course, imply that recognition of classes of figures is order 3. (It does imply that the translations of two different figures can be so separated. In fact, the method of §7.3, Application 7, shows this can be done with order 2, but only outside the bounded-coefficient restriction.)

The study of the relation between spectra and ordinary geometric concepts is presumably the domain of integral geometry, a subject with which we are not familiar. One would want to know, for example, how much of the spectrum is necessary to characterize figures invariantly under rotation--and what perceptron orders are needed to exploit this spectrum. (For, as in §5.8.1, the perceptron order may be higher than the spectral discrimination order.) The remarks about rotation, in Chapter 7, suggest that the problems about spectra under rotation are quite a bit deeper.

6.4 Patterns of Order 4 and Higher

As shown in §0.2, we can use the fact that any three points determine a

circle to make an order 4 perceptron for the predicate;

[X is the perimeter of a complete circle]

by using the form

$$\left[\sum_{d \notin C_{abc}} x_a x_b x_c x_d + \sum_{d \in C_{abc}} x_a x_b x_c \bar{x}_d < 1 \right]$$

where C_{abc} is the circle* through x_a , x_b , and x_c .

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 may be 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. The coefficient sizes are often fatally large, as shown in Chapter 9.
- c) a very slight change in the pattern definition** often destroys its order of recognizability. With low orders, it may not be possible to define tolerances for reasonable performance.

* Again there is a tolerance problem; what is a circle in the discrete retina? See §8.3.

** See note at end of Chapter 0.

6.5 Spectral Recognition Theorem

A number of the preceding examples are special cases of the following theorems. The ideas introduced here are not used later.

The group invariance theorem (§2.2) shows that if a predicate ψ is invariant with respect to a group G then if $\psi \in L(\Phi)$ for some Φ it can be realized by a form

$$\psi = [\sum_i \beta_i N_i(X)]$$

where the N_i are the numbers of φ 's in each G -equivalence class satisfied by X . In §5.2 we touched on the "difference vector spectrum" for geometric figures under the group of translations of the plane. These spectra are in fact the $N_i(X)$ numbers up to order $k = 2$. If a G -invariant ψ cannot be described by any condition on the N_i 's for a given Φ , then obviously ψ is not in $L(\Phi)$. The following results show some conditions on the N_i that imply that ψ is of finite order.

Suppose that ψ is defined by simultaneous satisfaction of m equalities:

$$\psi(X) \equiv [N_1(X) = n_1 \text{ and } N_2(X) = n_2 \text{ and } \dots N_m(X) = n_m]$$

where $n_1, \dots, n_m$ is a finite sequence of integers. Then ψ has $\leq$ twice the maximum order of the φ 's associated with the N_i 's. We will state this more precisely as

Theorem 6.5.1: Let

$$\Phi = \Phi_1 \cup \Phi_2 \cup \dots \cup \Phi_m,$$

and

$$N_i(X) = |\{\varphi; \varphi \in \Phi_i \text{ and } \varphi(X) = 1\}|$$

$$= \sum_{\varphi \in \Phi_i} \varphi(X).$$

Then the order of

$$\Psi(X) = \lceil N_i(X) = n_i, \text{ for } 1 \leq i \leq m \rceil$$

is at most twice

$$\max\{|\Phi|; \max_{\varphi \in \Phi} |\varphi|\}.$$

The goal of the proof is to show that the definition of Ψ can be put in the form of a linear threshold expression, viz.:

$$\Psi(X) = \lceil \sum (N_i(X) - n_i)^2 < 1 \rceil.$$

As it stands this is not a linear threshold combination of predicates. To recast it into the desired shape we introduce some ad hoc conventions that will not be used elsewhere. Given any set Φ of predicates we construct

a new set of predicates $\Phi^{(2)}$ by first listing all pairs of (φ_i, φ_j) of predicates in Φ and defining

$$\varphi_{i,j}(X) = \varphi_i(X) \wedge \varphi_j(X).$$

Many of the predicates so constructed will be logically equivalent, for example:

$$\varphi_{ij} = \varphi_{ji},$$

but we make the convention that these are to be counted as distinct members of $\Phi^{(2)}$. (This means that in a very strict sense $\Phi^{(2)}$ is a set of "predicate forms" rather than of predicates.)

The effect of the convention is to simplify the arithmetic and logic of the counting argument we are about to use. Let X be a figure for which exactly N predicates in Φ are satisfied. Obviously N^2 predicates of $\Phi^{(2)}$ will be satisfied by X , i.e.,

$$\sum_{\varphi \in \Phi^{(2)}} \varphi(X) = N^2.$$

Now let $\Phi_1, \Phi_2, \dots$ be an enumeration of the equivalence classes of Φ . Since the number of predicates of Φ_i satisfied by X is

$$N_i(X) = \sum_{\varphi \in \Phi_i} \varphi(X);$$

then, as we have seen,

$$\sum_{\varphi \in \Phi_i(2)} \varphi(X) = N_i^2(X).$$

Thus

$$\sum_i \left\{ \sum_{\varphi \in \Phi_i(2)} \varphi(X) - 2 n_i \sum_{\varphi \in \Phi_i} \varphi(X) + n_i^2 \right\} = \sum_i \left\{ (N_i(X) - n_i)^2 \right\}.$$

To represent the left hand side of this equation in the standard form for linear threshold predicates we define $\Phi' = \Phi(2) \cup \Phi \cup \{\text{the constant function}\}$. The linear form we want is

$$\sum \alpha(\varphi) \varphi$$

where

$$\alpha(\varphi) = 1 \text{ for } \varphi \in \Phi(2)$$

$$\alpha(\varphi) = -2n_i \text{ for } \varphi \in \Phi_i$$

$$\alpha(\text{constant}) = \sum n_i^2,$$

and then

$$\Phi(X) = \lceil \sum \alpha(\varphi) \varphi(X) < 1 \rceil.$$

To complete the proof of the theorem we have only to observe that

$$\begin{aligned} |S(\varphi_{i,j})| &= |S(\varphi_i) \cup S(\varphi_j)| \\ &< |S(\varphi_i)| + |S(\varphi_j)| \\ &< 2(\text{Max } |S(\varphi)|; \varphi \in \Phi). \end{aligned}$$

Q.E.D.

6.5.2 Extended Exact Matching

An obvious generalization of Theorem 6.5.1 is this:

Suppose that ψ is defined by

$$\psi(X) \equiv \bigvee_{i=1}^n \bigwedge_{j=1}^m (N_i = n_{ij}),$$

i.e., ψ satisfies any one of a number of exact conditions on the N_i . Then ψ is of finite order, for we can realize the polynomial form

$$\lceil \sum_{i=1}^n \sum_{j=1}^m (N_j(X) - n_{ij})^2 < 1 \rceil$$

methods like those in the previous paragraph. However, the extension now requires Boolean products of predicates of different equivalence classes, and the maximal order required will be $\leq 2 \cdot \sum_{k=1}^m |S_k|$ where $|S_k|$ is the support size associated with N_k .

6.5.3 Mean Square Variation

Instead of the predicates discussed in §5.5.1, we could increase θ to higher values:

$$\lceil \sum (N_i - n_i)^2 < \theta \rceil.$$

Then the system will accept figures for which the sum of the squares of the differences of the N_i 's and the n_i 's are bounded by θ . Any pattern-classification machine will be sensitive to certain kinds of distortion, and this observation hints that it might be useful to study such machines, and

perceptrons in particular, in terms of their spectrum-distortion characteristics. Unfortunately we don't have any good ideas concerning the geometric meaning of such distortions. The geometric nature of this sort of "invariant noise" is an interesting subject for speculation, but we have not investigated it.

6.6 Figures in Context

For practical and theoretical reasons it is interesting to study the recognition of figures "in context," for example:

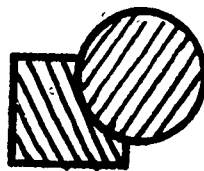
$$\psi(X) = [\text{a } \underline{\text{subset of } X} \text{ is a square}],$$

or

$$\psi(X) = [\text{a } \underline{\text{connected component of } X} \text{ is a square}],$$

or even, to begin to consider three dimensional projection problems:

$$\psi(X) = [X \text{ contains a significant portion of the outline of a partially-obscured square}].$$



The examples show that there is more than one natural meaning one could give to the intuitive project of recognizing instances of patterns embedded in "contexts." We do not know any general definition that might cover all natural senses, and are therefore unable to state sharp theorems. We do, nevertheless, claim that the general rule is for low order predicates to lose their property of finite order when embedded in context in any natural way. To illustrate the thesis we shall pick a particularly common and apparently harmless interpretation:

$$\psi_{\text{in context}}(X) = [\psi(Y) \text{ for some component, } Y, \text{ of } X].$$

It will be obvious that the techniques we use can be adapted trivially to many other definitions.

Intuitively, we would expect $\psi_{\text{in context}}$ to be much harder for a perceptron since the context acts as noise and the parallel operation of the device allows little chance for this to be separated from the essential component. The point appears particularly clearly in the cases where ψ uses rejection rules. These cannot be transferred over to $\psi_{\text{in context}}$ for very obvious reasons. Similarly, we lose the stratification methods of Chapter 7 and, indeed, most of our technical tricks used to obtain low order representations of predicates. The next two theorems show how this intuitive idea can be given a rigorous form. It should, however, be observed that no simple generalization is possible about the relation of ψ to $\psi_{\text{in context}}$ since some ψ 's become degenerate in context. For example, every set has a connected subset of odd parity and every set has a connected component!

Theorem 1: Let R be a finite square retina and let $\psi(X)$ be

$$\psi(X) = [X \text{ is exactly one horizontal line across the retina}].$$

Then ψ is of order 2 but $\psi_{\text{in context}}$ is not of finite order.

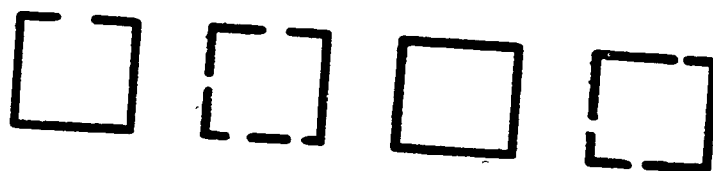
Proof: We leave as an exercise the proof that ψ as defined has order 2. To show that $\psi_{\text{in context}}$ is not of finite order we merely observe that it is the negation of the negative the one-in-a-box predicate, ψ_1 . Let G_1 be the $m \times m$ array of unshaded points discussed in §5.1. Taking this as our retina, the predicate ψ_1 asserts that there is not horizontal white line across the retina. Its negative, in the sense of §1.7, asserts that there is no horizontal black line. Since ψ_1 is not of finite order, the argument of §1.7 shows that the same is true of its negative. And by reversing the predicate's inequality we find the same is true for the desired

$$\psi_{\text{in context}} = [X \text{ contains a horizontal line across the retina}].$$

Theorem 2: Let $\psi(X)$ be

$$[X \text{ is a hollow square}].$$

Then $\psi_{\text{in context}}$ is not of finite order.



Proof: The proof is exactly the same as the previous except that the "boxes" or horizontal lines are folded into squares and mapped without overlap into a larger retina. Again, it can be shown that ψ itself is of finite order, in this case, order 3.

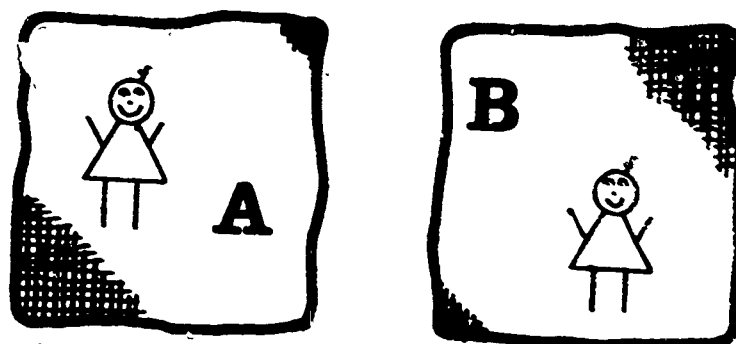
Note: An alternative proof method is to fold the lines of switching elements used in the Huffman construction for connectivity (§5.5).

CHAPTER 7: NORMALIZATION AND STRATIFICATION

7.1 Equivalence of Figures

In previous chapters we discussed the recognition of patterns--classes of figures--closed under the transformations of some group. We now turn to the related question of recognizing the equivalence, under a group, of an arbitrary pair of figures. The results below were surprising to us, for we had supposed that such problems were not generally of finite-order. A number of questions remain open, and the superficially positive character of the following constructions are clouded by the apparently enormous coefficients they require, and the manner in which they increase with the size of the retina.

A typical problem has this form: The retina



is presented* as two equal parts A and B and we ask: is the figure in part B a rigid translation of the figure in part A? More generally, is there an element g from some given group G of transformations for which B is the result of g operating on A? What order predicates are required to make such distinctions? The results of this chapter all derive from use of a technique we call stratification. Stratification makes it possible, under certain conditions, to simulate a sequential process by a parallel process, in which

* All the theorems of this Chapter apply directly to perceptrons on infinite retinas; that is, without having to consider limiting processes on sequences of finite retinas.

the results are so weighted that, if certain conditions are satisfied, some computations will numerically outweigh effects of others. The technique derives from the following theorem:

Theorem 7.2: (Stratification Theorem)

Let $\Pi = \{\pi_1, \pi_2, \dots, \pi_j, \dots\}$ be a sequence of predicates and define a sequence $C_1, \dots, C_j, \dots$ of classes by

$$X \in C_j \iff [\pi_j(X) \wedge (k > j \implies \sim \pi_k(X))]$$

Thus X is in C_j if j is the last index for which $\pi_i(X)$ is true.

Let $\Phi = \{\varphi_i\}$ be a family of predicates and let $\psi_1, \dots, \psi_j, \dots$ be an ordered sequence of predicates in $L(\Phi)$ that are each bounded in the sense that for each ψ_j there is a linear form with integer coefficients

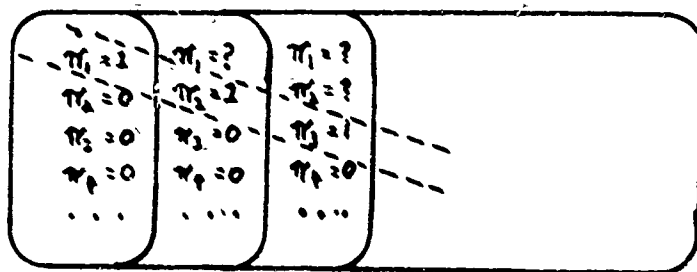
$$\Sigma_j = \sum_i \alpha_{ij} \varphi_i - \theta \text{ such that}$$

$$\psi_j = [\Sigma_j > 0]$$

and that there exist bounds B_j such that $|\Sigma_j(X)| < B_j$ for all finite $|X|^*$. Then the predicate $\psi(X) = [X \in C_j \implies \psi_j(X)]$ obtained by taking, on each C_j , the values of the corresponding ψ_j , lies in $L(\Phi \cup \Pi)$, that is, can be written as a form

* The proof actually requires only that $|\Sigma_j(X)|$ be bounded on each C_k , i.e., non-uniformly.

$$\psi(X) = \left| \sum \alpha_{jk} (\pi_j \wedge \varphi_k) \right| > \theta.$$



The partition into $C_j(X)$.

Fig. 7.2.1

Usually it will be the case that for any finite $|X|$, X will lie in one of the C_j . Otherwise we will be interested only in the values of $\psi(X)$ for $X \in \bigcup_j C_j$.

The proof is by an inductive construction. Define

$$S_1 = \pi_1 \cdot (E_1)$$

and

$$\begin{cases} M_j = \max_{C_j} |S_{j-1}| \\ S_j = S_{j-1} - \pi_j M_j + (2M_j + 1) \cdot \pi_j \cdot \varphi_j. \end{cases}$$

The bounds B_j assure the existence of the M_j 's. Now write the formal sum generated by this infinite process as

$$S = \sum_{j,k} \beta_{jk} \pi_j \wedge^k$$

and we will show that $\psi(X) = [S(X) > 0]$. The infinite sum is well-defined because, since each X is in some C_j , there will be only finite sum associated with each π_j term. **BASE:** It is clear that if X is in C_1 then $\pi_1 = 1$ so $\psi(X) = [S_1(X) > 0]$. **INDUCTION:** Assume that if X is in C_{j-1} then $\psi(X) = [S_{j-1}(X) > 0]$. Now the coefficients are integers, so if $X \in C_j$, $\pi_j = 1$ and

- i) if $\psi(X)$ then $\Sigma_j \geq 1$ so $S_j \geq -M_j - M_j + 2M_j + 1 = 1$
- ii) if $\sim \psi(X)$ then $\Sigma_j \leq 0$ so $S_j \leq M_j - M_j = 0$.

Q.E.D.

Corollary 7.2: The order of $\psi(X)$ is no larger than the sum of the degree in ψ and the maximum support in Π . This follows because the predicates in ψ occur only as conjuncts with predicates in Π .

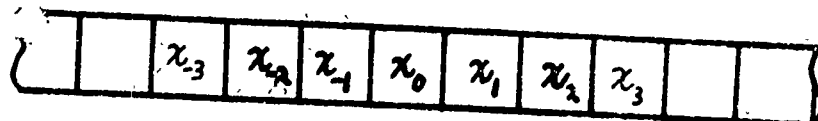
Thus the construction assumes that the domain of $\psi(X)$ can be divided into classes by its intersections with the disjoint "strata," C_j . Within each stratum the $-\pi_j M_j$ term negates decisions made on lower strata until the ψ_j test is passed. In all applications below, the strata represent, more or less, the different possible deviations of a figure from a "normal" condition.

Hence there is a close connection between the possibility of constructing "stratified" predicates, and the conventional "pattern recognition" concept of identifying a figure first by normalizing it and then by comparing the normalized image with a prototype.

It should be noted that predicates obtained by the use of this theorem will have enormous coefficients, growing exponentially or faster with the stratification index j . Thus the results of this chapter should not be considered of practical interest. They are more of theoretical interest in showing something about the relation--or perhaps non-relation--of the structure of the transformation groups to the order of certain predicates invariant under those groups.

7.3 Application 1: Symmetry along a Line

Let $R = \dots, x_0, \dots$ be the points of an infinite linear retina, i.e., $-\infty < x_S < \infty$:



Suppose that X is a figure in R with finite $|X|$. We ask whether the predicate

$$\psi_{\text{SYM}}(X) = [X \text{ is symmetrical under reflection}^*]$$

The important thing is that the symmetry center is not specified in advance, and may occur anywhere along the infinite line! If the reader has any difficulty with this section, he should read § 7.9 first.

is of finite order.

We will "stratify" ψ_{SYM} by finding sequences $\pi_1, \dots$ and $\psi_1, \dots$ that allow us to test for symmetry, using the following trick: the π_i 's will "find" the two "endpoints" of X and each of the ψ_i 's will test the symmetry of a figure for the corresponding pair of endpoints. Our goal, then, is to define the π_i 's so that each C_j will be the class of figures with a certain pair of endpoints. To do this we need $\pi_1, \dots$ to be an enumeration of all segments (x_s, x_{s+d}) for any s and for any $d \geq 0$, with the property that any term (x_s, x_{s+d}) must follow any term (x_{s+a}, x_{s+a+b}) with $0 \leq a \leq a+b \leq d$. There do indeed exist such sequences, as shown in Fig. 7.2.1-1.

$$\pi_1 = x_0 x_0 = x_0$$

$$\pi_2 = x_1 x_1 = x_1$$

$$\pi_3 = x_0 x_1$$

$$\pi_4 = x_{-1}$$

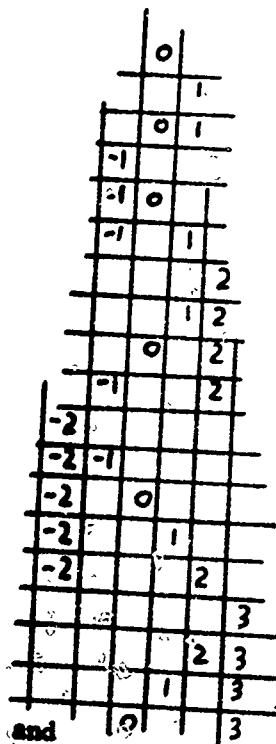
$$\pi_5 = x_{-1} x_0$$

$$\pi_6 = x_{-1} x_1$$

$$\pi_7 = x_2 x_2 = x_2$$

$$\pi_8 = x_1 x_2$$

$\vdots$



and it can be seen that

- i) each segment occurs eventually and
- ii) no segment is ever followed by another that lies within it.

Therefore, if x_s, x_{s+d} are the extreme left and right points of X , then X will lie in precisely the C_j for that (x_s, x_{s+d}) . Now define ψ_j to be

$$\psi_j = \lceil x_{s+1} = x_{s+d-1}, i = 0, \dots, d \rceil$$

or, equivalently,

$$\psi_j = \lceil \sum_{i=0}^d (x_{s+1})(1 - x_{s+d-1}) \leq 0 \rceil$$

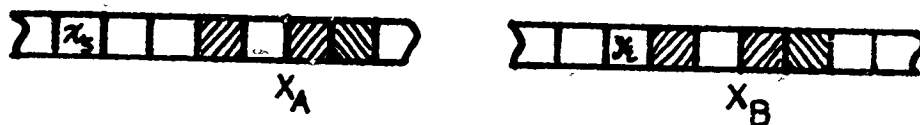
showing that it is a predicate of order 2 which is bounded, with $B_j < \frac{d}{2}$.

So, finally, application of the stratification theorem shows that

ψ_{SYM} has order ≤ 4 , since the ψ 's have order ≤ 2 and the π 's have support ≤ 2 .

7.4 Application 2: Translation-Congruence along a Line

Let $\dots, x_s, \dots$ and $\dots, y_t, \dots$ be the points of two infinite linear retinas, i.e., $-\infty < x_s, y_t < +\infty$:



Let X be a figure composed of a part X_A in the left retina and a part X_B in the right retina. We want to construct $\psi_{TRANS}(X) = \lceil \text{the (finite) pattern in A is a translate of the pattern in B} \rceil$.

To "stratify" ψ_{TRANS} we have to find a sequence π_i that allows us to test, with appropriate ψ_i 's, whether the A and B parts of X are congruent.

We will do this by a method like that used in § 7.2.1 but we have now to handle

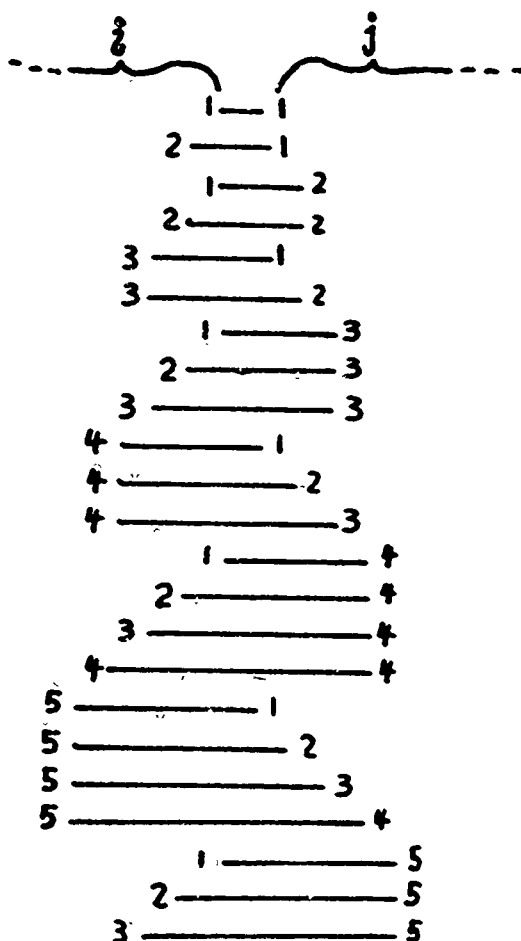
two segments simultaneously. That is, we need a sequence of π_j 's that enumerate all quadruples in such a way that a figure lies in C_j if and only if the endpoints of its A and B parts are precisely the corresponding values of x_s, x_{std_x}, y_t and y_{ttd_y} . There does indeed exist such a sequence (!, and one can be obtained from the π_i 's of § 7.2.1 as follows (the reader might try to find one himself):

Define π_{jk} to be the four-point mask obtained by

$$\pi_{jk}(X) = \pi_j(X_A) \cdot \pi_k(X_B),$$

that is, by choosing according to i two points of A and according to j two points of B. The master sequence requires us to enumerate all π_{ij} 's under the condition that no π_{ab} can precede any π_{cd} if both $a \geq c$ and $b \geq d$.

A solution is:



$\pi_{11}; \pi_{21}, \pi_{12}, \pi_{22}; \pi_{31}, \pi_{32}, \pi_{13}, \pi_{23}, \pi_{33}; \pi_{41}, \pi_{42}, \pi_{43}, \pi_{14}, \pi_{24}, \dots$
 and for the π_{jk} term in this sequence, an appropriate predicate $\psi(jk)$ is:

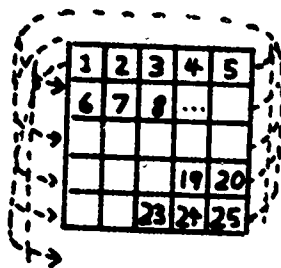
$\psi(jk) =$ [the segments defined by π_j and π_k have
 the same lengths, and the x 's and y 's
 in those intervals have the same values
 at corresponding points].

This is an order 2 predicate, and bounded (by the segment lengths). The π_j 's now have support 4, so $\psi_{\text{TRANS}}(X)$ has finite order ≤ 6 . Actually, having found both extrema of X_A , it is necessary only to find one end of X_B , so a slightly different construction using the method of §7.9 shows that the order of ψ_{TRANS} is ≤ 5 .

7.5 Application 3. Translation on the Plane.

The method of application 2 can be applied to the problem of the two-dimensional translations of a bounded portion of the plane by using the following trick:

Let each copy of the retina be an $(m \times m)$ array. Arrange the squares into a sequence $\{x_i\}$ with the square at (a,b) having index $ma + b$. In effect, we treat the retina as a cylinder and index its squares so:



This maps each half of the retina onto a line like that of application 2 in such a way that for limited translations that do not carry the figure λ over the edge of the retina, translations on the plane are equivalent to translations on the line, and an order-5 predicate can be constructed. In § 7.6 we will show how the ugly restriction just imposed can be eliminated!

Application 4. 180° rotation about undetermined point on the plane.

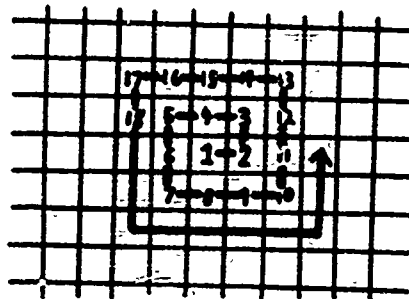
With the same kind of restriction, this predicate can be constructed (with order=4) from application 1 by the same route that derived application 3 from application 2. Similarly, we can detect reflections about arbitrary vertical axes.

7.6 Repeated Stratification.

In the conditions of the Stratification Theorem, the only restriction on the ψ_j 's is that they be suitably bounded. In certain applications, there is no

reason the ψ_j 's themselves cannot be obtained by stratification. This is particularly easy to do when the support of ψ_j is finite, for then boundedness is immediate. To illustrate this repeated stratification we will proceed to remove the finite restriction in Application 3 of §7.5.

First enumerate all the points of each of two infinite plane retinas A and B according to the more or less arbitrary pattern:



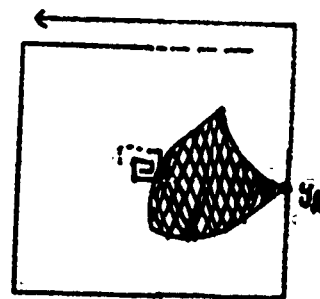
to obtain two sequences $x_1, \dots, x_5, \dots$ and $y_1, \dots, y_5, \dots$.

We had a similar problem there, only now it is a two-dimensional version.

Now we will invoke precisely the same enumeration as in 7.4, but defining

$$\pi_{jk}(X) = (x_j \in X_A \wedge y_k \in X_B) = x_j \cdot y_k$$

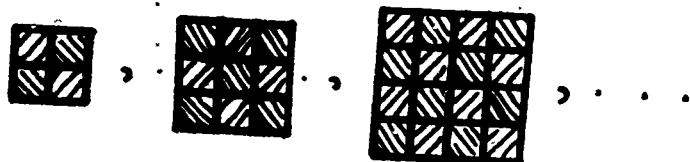
Then $C_{(jk)}$ is the class of pairs of figures whose highest (in the x-enumeration) point of X_A is x_j and highest point of X_B is y_k



We need only a (bounded) $\psi_{(jk)}$ that decides whether X_A is a translate of X_B for figures in C_{jk} . But the figures in C_{jk} all lie within bounded portions of the planes, in fact within squares of about $[\max(j,k)]^{1/2}$ on a side around the origins! Within such a square - or better, within one with twice the dimensions, to avoid "edge-effects" - we can apply directly the result of application 3, Chapter 7.5 to obtain a predicate $\psi_{(jk)}$ with exactly the desired property, and with finite support! The resulting order is $\leq 5 + 2 = 7$. (We have another slightly fallacious construction that yields order-4, so we suppose the true value to be somewhere in between). The same arguments can be used to lift the restrictions in application 4, Chapter 7.5.

7.7 Application 5. The Axis-parallel Squares in the Plane

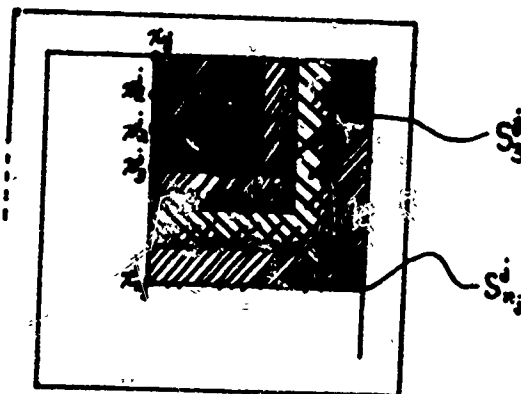
We digress a moment to apply the method of the last section to show that the predicate $\psi_{\square}(X) = [X \text{ is a solid (hollow) axis-parallel square}]$, (that is, of the form anywhere in the plane) has order ≤ 3 .



(We consider this remarkable because informal arguments, to the effect that two sides must be compared in length while the interior is also tested, suggest orders of at least 4. The result was discovered and proven by another method by our student, John White).

We enumerate the points $x_1, \dots$, of a single plane, just as in § 7.6 and simply set $\pi_j = x_j$. Then C_j is the set of figures whose "highest" point is x_j .

If X is a square, the situation is like one of the cases shown:



We then construct ψ_j by stratifying as follows: Let $x_1^j, x_2^j, \dots, x_{n_j}^j$ be the finite sequence obtained by stepping into the spiral figure orthogonally from x_j . Define $\pi_1^j = x_1^j$ so that C_1^j will contain all the squares of length 1 on a side that are "stopped" by x_j . But there is only one such square, call it S_1^j . So to complete the double stratification we need only provide predicates ψ_1^j to recognize the squares S_1^j . But this can be done by

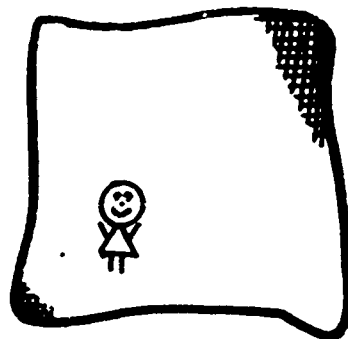
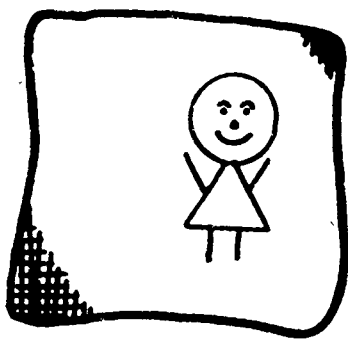
$$\psi_1^j = \left[\bigwedge_{x_k \in S_1^j} x_k - i^2 \bigwedge_{x_k \notin S_1^j} x_k \geq i^2 \right]$$

which is of order=1. $\{\psi\}$ has order ≤ 3 !

Q.E.D.

7.8 Application 6. Figures equivalent under Translation and Dilatation

Can a system of finite order recognize equivalence of two arbitrary figures under translation and size change?



Some reflection about the result and methods used in § 7.6 and 7.7 will suggest that we have all the ingredients, for 7.6 shows how to handle translation, and § 7.7 shows how to recognize all the translations and dilatations of a particular figure. Now dilatation involves serious complications with tolerance and resolution limits, in so far as our theory is still based on a fixed, discrete retina, and we do not want to face this problem squarely. None the less, it is interesting that the desired property can at least be approximated with finite order, in an intuitively suggestive fashion. (We do not think that a similar approximation can be made in the case of rotation-invariance, because the problem there is of a different kind, that cannot be blamed on the discrete retina. Rather, it is because the transformations of a rotation group cannot be simply ordered, and this "blocks" stratification-like methods).

Our method begins with the technique used in § 7.6 to find predicates $\pi_{(jk)}$ that "catch" the two figures in boxes. Then, just as in § 7.6, the problem is reduced to finding predicates $\psi_{(jk)}$ that need only operate within the boxes of fig. 7.6-1. We construct the $\psi_{(jk)}$'s by a brutal method: within each box we use the simple enumeration of points described in Chapter 7.5. Then we stratify four times (!) in succession with respect to:

- (1) x = highest and leftmost point of A
- (2) y = highest and leftmost point of B
- (3) x' = lowest and rightmost point of A
- (4) y' = lowest and rightmost point of B

We will need to define predicates $\psi_{x,y,x',y'}^{(jk)}$ for this. If the two vectors $\overrightarrow{x-y}$ and $\overrightarrow{x'-y'}$ do not have the same direction we set $\psi = 0$; otherwise we need a ψ to test whether or not for every vector displacement $\vec{v}$

$$y + \vec{v} \equiv x + \frac{x - x'}{y - y'} \vec{v}$$

and this is an order-2 predicate, leading finally to total order $\leq 2 + 4 + 2 = 8$. Of course, on the discrete retina the indicated operations on vectors will be ill-defined, but it seems clear that the result is not vacuous: for example, we could ask for recognition of the case where X_B is a translate and an integer-multiple of X_A in size, with each black square of X_A mapping into a correspondingly larger square in X_B .

7.9 Application 7. Equivalents of a Particular Figure

In constructing ψ for application 5, we noted that one can always construct an order-1 predicate to detect precisely one particular figure X_0 (by using

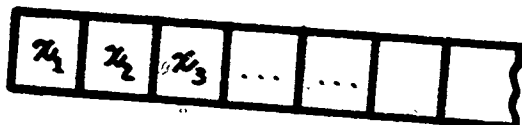
$\left[\sum_{x \in X_0} x + \sum_{x \notin X_0} x \geq 1 \right]$). It follows that if we can construct a stratification $\{\pi_i\}$ for a group G such that for all $g \in G$

$$x \in C_1 \wedge gx \in C_1 \rightarrow (gx = x) \vee (g = e)$$

then we can recognize exactly the G -equivalents of a given figure X_0 (with one order higher than the order used by the stratification π 's). This is

suggestive of a machine that "pre-processes" the figures by bringing them into a Normal Form. For this case our general construction method takes a very simple form:

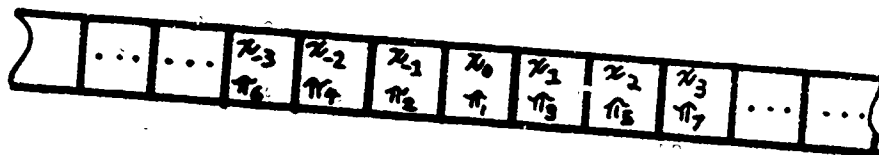
Consider a particular figure X_0 consisting of the ordered sequence of points $\{x_{i_1}, \dots, x_{i_p}\}$ on the half line



Let $\pi_j(X) = [x_j, \infty]$ and define

$$\pi_j(X) = \left[\sum_{x_k \in X_0} \overline{x_{j-i_p+k}} + \sum_{x_k \in X_0} x_{j-i_p+k} < 1 \right],$$

ignoring for the moment points with negative indices. Then, except for "edge-effects" we obtain a predicate of order-2 that recognizes precisely the translates of X_0 . Next we observe that there is really no difficulty in extending this to the two-way infinite line, for we can enumerate the π_i 's in the order



so that if a figure ends up in class C_{2j} we will have found its leftmost point x_{-j} and if it is in a C_{2j+1} we will have found its rightmost point x_j . In either

case we can construct an appropriate ψ . Hence, finally, we see that there exists for any given figure X_0 a predicate of order-2 that recognizes precisely the linear translations of X_0 , and there is no problem about boundedness because all ψ -supports are finite.

7.10 Apparent Paradox

Consider the case of $X_0 =$



We have just shown that there exists an order-2 ψ that accepts just the translates of this figure. Hence ψ must reject the non-equivalent figure,



But both of these figures have exactly the same n -tuple distribution spectrum (see Chapters 6.2 and 6.5) up to order-2! Each has 3 points, and each has 1 adjacent pair, 1 pair two units apart and 1 pair 3 units apart. Therefore, if all group-equivalent ϕ 's had the same weights, an order-3 perceptron would be needed to distinguish them. Thus if we could apply the group-invariance theorem we would in fact obtain a proof that no perceptron of order-2 can distinguish between these. This would be a contradiction! What is wrong? The answer is that the group-invariance theorem does not in general apply to predicates invariant under infinite groups. When a group is finite, e.g., cyclic, as in the toroidal spaces we have considered from time to time, one can always use the group-invariance theorem to make equal the coefficients of equivalent ϕ 's. But we cannot use it together with stratification, to construct the predicate on infinite groups.

With infinite groups we can use stratification for normalizing, but then we must face the possibility of getting unbounded coefficients within equivalent

ϕ 's; then the group-averaging operations do not, in general, converge. This will be shown as a theorem in Chapter 9.4.

7.11 Problems

A number of lines of investigation are intriguing: what is the relation between the possible stratifications, including repeated ones, and algebraic decompositions of the group into non-cyclic and cyclic factors? For what kinds of predicates can the group-invariance theorem be extended to infinite groups? What predicates have bounded coefficients in each equivalence class, or in each degree? Under what conditions do the "normal-form stratifications" of application 7 exist? For example, we conjecture that on circles or toroids, there is no bound on the order of predicates ϕ that select unique "normal form" figures under rotation groups:

$$\phi(X) \rightarrow (\psi(gX) \rightarrow gX = X).$$

We suspect that this may be the reason we were unable to extend the method of application 6 to the full Euclidean Similarity group, including rotation.

Remarks

1. For a long time we thought that equivalence problems, like that of § 7.5 and § 7.6 were not of finite order. Stratification was surprising.
2. Stratified predicates probably are physically unrealisable because of huge coefficients. We have no method for finding lower bounds, but it appears that the coefficients grow faster than exponentially in R , in general.
3. A stratification seems to correspond to a serial machine that operates sequentially upon the figure, with a sequence of group transformation elements, until some special event occurs, establishing its membership in C_j , and then

CHAPTER 8: THE DIAMETER-LIMITED PERCEPTRON

8.0

In this chapter 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 of diameter less than or equal to some length D . That is, $\text{Diameter}(S(\phi)) < D$. 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.

8.1 Positive Results

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

8.1.1 Uniform Picture

A diameter-limited perceptron can tell when a picture is entirely black, or entirely white: choose ϕ_i 's that cover the retina in regions (that may overlap) and define ϕ_i to be zero if and only if all the points it can see are white. Then

$$\sum \phi_i > 0$$

if and only if the picture has one or more black points, and not if the picture is blank.

Similarly, we could define the ϕ_1 's to distinguish the all-black picture from all others.

These patterns are recognizable because of their "conjunctively local" character (see Introduction): no ϕ -unit can really say that there is strong evidence that the figure is all-white (for there is only the slightest 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.

8.1.2 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_1 for each point to be 1 if that point is black, 0 otherwise. Then

$$\sum x_1 > S$$

is a recognizer for the class in question.

8.1.3 Non-intersecting Lines

One can say that a pattern is composed of non-intersecting lines if, in each small region, the pattern contains at most one line-segment. If we make each ϕ have value zero when this condition is met, unity when it is not,

$$\sum \phi_1 > 0$$

will reject all figures not in the class.

8.1.4 Triangles and Rectangles

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, $\hat{\phi}_i$, 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 non-overlapping* 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

$$\sum \phi_i - W \sum \hat{\phi}_i < 4$$

will be a specific recognizer for triangles. (It will, however, accept the blank picture, as well). Similarly, by setting the ϕ 's to recognize only right angles, we can discern the class of rectangles with

$$\sum \phi_i + W \sum \hat{\phi}_i < 5$$

A few other geometric classes can be captured by such tricks, but they depend on curious accidents. A rectangle is characterized by having four right angles, and none of the exceptions detected by the $\hat{\phi}_i$'s. In § 6.3.2 we did this for axis-parallel rectangles: for others there are obviously more serious resolution and tolerance problems. But there is no way to recognize the squares, even axis-parallel, with diameter-limited ϕ 's; the method of § 7.2.5 can't be so modified.

* Of course, this won't work when a vertex occurs at the edge of a ϕ -support. By suitable overlapping, and assignment of weights, the system can be improved, but it will always be an approximation of some sort. This applies to the definition of "line segment," etc., as well as to that of "vertex." See § 8.3.

8.1.5 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. In fact we show in the next section that even the simplest figure, that consists of just one point, cannot be recognized independently of position!

8.2.1 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.

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

$$f(X_1) \geq \theta$$

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,

$$f(X_1) - f(X_0) > 0. \quad (1)$$

Now choose another point x which is further than D away from x . 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

$$f(X_2) = \sum \alpha_i \phi_i \geq 0 \quad (2)$$

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),

$$f(X_{12}) > 0$$

but we require that

$$f(X_{12}) < 0$$

Remark: Of course, this is the same phenomenon noted in Chapter 0.3 and in Chapter 2.1. And it gives the method for proof of the last statement in Chapter 8.1.4.

8.2.2 Area Segments

The diameter-limited perceptron cannot recognize the class of figures whose areas A lie between two bounds $A_1 \leq A \leq A_2$.

Proof: this follows from the method of § 8.2.1, which is a special case of this, with $A_1 = 1$ and $A_2 = 1$. (But using the method of § 1.4), example (vii), this recognition is possible with order 2 if the diameter-limitation is relaxed).

8.2.3 Connectedness

The diameter-limited perceptron cannot decide when the picture is a single, connected whole, as distinguished from two or more disconnected pieces. At this point the reader will have no difficulty in seeing the formal correctness of the proof we gave of this informally in Chapter 0.3.

Proof: consider the four pictures

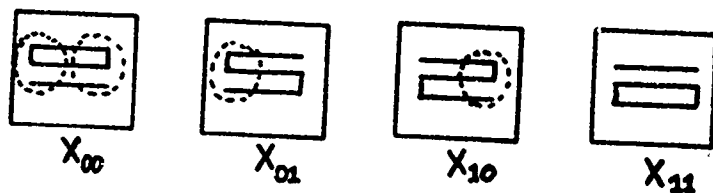


Fig. 8.2.3

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 $\psi \in L(\phi)$ for which

$$\sum \alpha_i \phi_i(X_{01}) \geq \theta$$

$$\sum \alpha_i \phi_i(X_{00}) < \theta$$

$$\sum \alpha_i \phi_i(X_{11}) < \theta$$

$$\sum \alpha_i \phi_i(X_{10}) \geq \theta$$

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

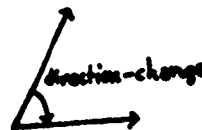
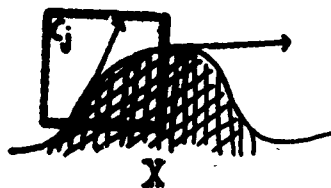
$$\sum \alpha_i \phi_i(X_{11}) \geq 0 + 0 - 0 = 0$$

contradicting the separation requirement.

8.3 Diameter-limited Integral Invariants

We observed in § 6.3.1 that convexity has order 3, but that the construction used there would not carry over to the diameter-limited case because it would not reject a figure with two widely separated convex components. On the other hand, § 8.1.4 shows how a diameter-limited predicate can capture some particular convex figures. The latter construction generalizes, but leads into serious problems about tolerance and, really, into questions about differentials.

Suppose that we define a diameter-limited family of predicates ϕ_c using the following idea: Choose an $\epsilon > 0$. Cover R with a partition of small cells C_j . For each integer k define ϕ_{jk} to be 1 if $C_j \cap X$ contains an "edge" with change-in-direction $> k\epsilon$ and otherwise $\phi_{jk} = 0$.



Now consider the "integral"

$$\sum_{jk} \epsilon \cdot \phi_{jk}$$

The contribution to the sum, of each segment of curve will be $\epsilon \cdot \frac{c}{\epsilon} = c$ where c is the change in direction of the segment, hence the total sum is the total "curvature" or, rather, the total |curvature|. Finally we claim that we can "realize" ψ convex as

$$\psi_{\text{CONVEX}} \sim \left[\sum_{jk} \epsilon \cdot \phi_{jk} < 2\pi \right]$$

because the total |curvature| of any figure must be $> 2\pi$ and only (and all) convex figures achieve the equality. We ignore figures that reach the edge of the retina and such matters.

A similar argument can be used to construct a predicate that uses the signed curvature to realize

$$G(X) < n$$

the functions of the Euler characteristic, since that invariant is just the total signed curvature divided by 2π .

One could go on to describe more sophisticated predicates that classify figures by properties of their "differential spectra."

However, we do not pursue this because these observations already raise a number of serious questions about tolerances and approximations. There are problems about the uniformity of the coverings, the sizes of ϵ and the diameter-limited cells C_j , and problems

about the cumulative errors in summing small approximate quantities. Certainly within the $E^2 \rightarrow R$ square map described in Chapter 5, or anything like it, all such predicates will give peculiar results whenever the diameter cells are not large compared to the underlying mesh, or small compared to the relevant features of the X's.

For example, we can regard the recognition of rectangles, as done in Chapter 6.3.2, as a pure artifact in this context, because it so depends on the mesh. The description in Chapter 8.1.4 of another form of the same predicate is worded in such a way that one could make reasonable approximations, within reasonable size ranges.

CHAPTER 9: MAGNITUDE OF THE COEFFICIENTS

9.1 Coefficients of the Parity Function Ψ_{PAR}

In §3.1 we discussed the predicate $\Psi_{\text{PAR}}(X) = \lceil |X| \text{ is an odd number} \rceil$ and showed that if Φ is the set of masks then all the masks must appear in any $L(\Phi)$ expression. One such expression is

$$\Psi_{\text{PAR}}(X) = \lceil \sum (-2)^{|S(\mu_i)|} \mu_i(X) < -1 \rceil$$

which contains each mask μ_i with coefficients that grow exponentially with the support-size of the masks. We will now show that the coefficients must necessarily grow at this rate, because the sign-changing character of parity requires that each coefficient be large enough to drown out the effects of the many coefficients of its submasks. In effect, we show that Ψ_{PAR} can be realized over the masks only by a stratification-like* technique! So suppose that we have $\Psi_{\text{PAR}} = \lceil \sum \alpha_i \mu_i > 0 \rceil$. Suppose also that the group-invariance theorem has been applied to make equal all α 's for μ 's of the same support-size, and suppose finally that the discrimination of Ψ_{PAR} is "reliable," e.g., that $\sum \alpha_i \mu_i \geq 2$ for odd parity and $\sum \alpha_i \mu_i \leq 0$ for even parity. (We use "2" instead of "1" to make the proof slightly neater.) Then we obtain the inequalities

$$\alpha_1 \geq 2$$

$$\alpha_2 + 2\alpha_1 \leq 0$$

$$\alpha_3 + 3\alpha_2 + 3\alpha_1 \geq 2$$

...

* But not by stratification itself, because the order cannot be bounded. In this Chapter we return to finite $|R|$ spaces.

or

$$\sum_{i=1}^n \binom{n}{i} \alpha_i \begin{cases} \geq 2 & \text{if } n \text{ is odd} \\ \leq 0 & \text{if } n \text{ is even} \end{cases}$$

Subtracting successive inequalities, we define

$$\begin{aligned} D_n &= \sum_{i=1}^{n+1} \binom{n+1}{i} \alpha_i - \sum_{i=1}^n \binom{n}{i} \alpha_i \\ &= \alpha_{n+1} + \sum_{i=1}^n \left[\binom{n+1}{i} - \binom{n}{i} \right] \alpha_i = \alpha_{n+1} + \sum_{i=1}^n \binom{n}{i-1} \alpha_i \\ &= \sum_{i=0}^n \binom{n}{i} \alpha_{i+1} \end{aligned}$$

so that for all n ,

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

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

$$\sum_{i=0}^M \binom{M}{i} [(-1)^i D_i - 2] \geq 0$$

Then
$$\sum_{i=0}^M \binom{M}{i} (-1)^i D_i \geq 2 \sum_{i=0}^M \binom{M}{i} = 2^{M+1}$$

The left-hand side is

$$\begin{aligned}
 & \sum_{i=0}^M \sum_{k=0}^i (-1)^i a_{k+1} \binom{i}{k} \binom{M}{i} = \sum_{k=0}^M \sum_{i=k}^M (-1)^i a_{k+1} \binom{i}{k} \binom{M}{i} \\
 & = \sum_{k=0}^M \sum_{i=k}^M (-1)^i a_{k+1} \left(\frac{i!}{k!(i-k)!} \right) \left(\frac{M!}{i!(M-i)!} \right) \\
 & = \sum_{k=0}^M \sum_{i=k}^M (-1)^i a_{k+1} \left(\frac{M!}{k!(M-k)!} \right) \left(\frac{(M-k)!}{(i-k)!(M-i)!} \right) \\
 & = \sum_{k=0}^M a_{k+1} \binom{M}{k} (-1)^k \sum_{j=0}^{M-k} \left(\frac{(M-k)!}{j!(M-k-j)!} \right) (-1)^j \\
 & = \sum_{k=0}^M a_{k+1} \binom{M}{k} (-1)^k (1-1)^{M-k} \\
 & = a_{M+1} (-1)^M
 \end{aligned}$$

so we have;

Theorem: the ratio of the largest coefficient to the smallest coefficient of any mask must exceed:

$$\frac{(-1)^M a_M}{a_1} \gg \frac{2^M}{2} = 2^{M-1}$$

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 $2^{|R|}$ boolean functions are linear threshold functions on the set of masks. So, on the average, specification of a function will require $2^{|R|}$ bits of coefficient-information, and non-uniformity of coefficient sizes would be expected to raise this by a substantial factor.

9.2 Coefficients Can Grow Even Faster than Exponentially in $|R|$

It might be suspected that ψ_{PAR} is a sort of worst case both because (a) parity is a worst function and (b) masks make a worst ϕ . In fact the masks make rather a good base because coefficients over masks never have to be larger than $|\alpha_i| = 2^{|S(\mu_i)|}$, as can be seen by expanding an arbitrary predicate into positive normal form. We now present a new predicate ψ_{EQ} together with a rather horrible ϕ , that leads to worse coefficients.

Let R be a set of points, $y_1, \dots, y_n, z_1, \dots, z_n$ and let $\{Y_i\}$ and $\{Z_i\}$ each be enumerations of the i subsets of the y 's and z 's respectively. Then any figure $X \subset R$ has a unique decomposition $C = Y_j \cup Z_k$.

We will consider the simple predicate ψ_{EQ} :

$$\psi_{EQ}(Y_j \cup Z_k) = [j = k]$$

which simply tests, for any figure X , whether its Y and Z parts have the same positions in the enumeration. The straightforward geometric example is that in which the two halves of R have the same form, and Y_i and Z_i are corresponding sets of y and z points.

We will construct a very special set Φ of predicates, for which $\psi_{EQ} \in L(\Phi)$ and show that any such realization must involve incredibly large coefficients! We want to point out at the start that the Φ we will use was designed for exactly this purpose. In the case of ψ_{PAR} we saw that coefficients can grow exponentially with the size of $|R|$; in that case the Φ was the set of masks, a natural set, whose interest exists independently of this problem. To show that there are even worse situations we construct a Φ with no other interest than that it gives bad coefficients.

We will define Φ to contain two types of predicates:

$$\psi_i(Y_j \cup Z_k) = [i = k]$$

$$\chi_i(Y_j \cup Z_k) = [(j = k \wedge i = k) \vee (j = k-1 \wedge i < k)]$$

each defined for $i = 1, \dots, 2^n$. Note that $|S(\psi_i)| = n$ and $|S(\chi_i)| = 2n$

First we must show that $\psi_{EQ} \in L(\Phi)$. But consider the proposed form:

$$\psi_{EQ} = \left\lceil \sum 2^i (\psi_i - x_i) < 1 \right\rceil$$

Case I: $j = k$

Then $\psi_k = 1$ and $x_k = 1$ hence $\psi_{EQ} = \left\lceil 2^k (1 - 1) < 1 \right\rceil = 1$

Case II: $j \neq k$ $j \leq k-1$

Then only $\psi_k = 1$ and $\psi_{EQ} = \left\lceil 2^k < 1 \right\rceil = 0$

Case III: $j = k-1$

Then $\psi_{k-1} = 1$ and $x_i = 1$ for $i = 1, \dots, k-1$.

So

$$\psi_{EQ} = \left\lceil 2^k - \sum_{i=1}^{k-1} 2^i < 1 \right\rceil = \left\lceil 2 < 1 \right\rceil = 0$$

and the predicate holds only for the $j = k$ case, as it should. So ψ_{EQ} is indeed in $L(\Phi)$.

Now we establish bounds on the coefficients. Consider any expression

$$\psi_{EQ} = \left\lceil \sum \alpha_i x_i + \sum \beta_i \psi_i > \theta \right\rceil$$

Then for sets $Y_{k+1} \cup Z_k$ we get $\beta_k < \theta$,

for sets $Y_k \cup Z_k$ we get $\alpha_k + \beta_k > \theta + 1$, where we assume strong association.

and for sets $Y_{k-1} \cup Z_k$ we get

$$\alpha_1 + \dots + \alpha_{k-1} + \beta_k < \theta$$

We can set $\theta = 0$ by subtracting it from every β , since just one β appears in each inequality.

Then $\beta_1 \leq 0$ and $\alpha_1 \geq 1$. Then, since

$$\alpha_k \geq 1 + \alpha_1 + \dots + \alpha_{k-1}$$

we have immediately $\alpha_2 \geq 2$, $\alpha_3 \geq 4$, ..., $\alpha_j \geq 2^{j-1}$

Since the index j runs from 1 to 2^n , the highest α must be at least 2^{2^n-1} times as large as the initial separation term $(\alpha_1 + \beta_1) - \beta_1 = \alpha_1$. This incredible growth rate is based in part on a mathematical joke: we note that an expression " $j = k$ " equivalent to that for ψ_{EQ} appears already within the definitions of the χ_i 's, and it is there precisely to not-quite-fatally weaken their usefulness in $L(\phi)$. Thus, one can not conclude that the ψ_{PAR} result is just due to the poor choice of the set of masks for its ϕ -base. (Problem: find a ϕ that makes the coefficients of ψ_{PAR} grow like $2^{2^{|R|}}$ -constant. Solution in Chapter 9.3).

Ironically, if we write ψ_{EQ} in terms of masks we have

$$\psi_{EQ} = \left[\sum My_i + Mz_i - 2My_1z_1 < 1 \right]$$

and the coefficients are very small indeed!

Problems: In 9.1, ϕ has $2^{|R|}$ elements and ψ_{PAR} requires coefficients like $2^{|R|}$. In 9.2 ϕ has $2^{\frac{1}{2}|R|}$ elements but the coefficients are like $2^{2^{|R|}}$. It is possible to make ϕ 's with up to $2^{2^{|R|}}$ elements. Does this mean there are ψ 's and ϕ 's with coefficients like $2^{2^{2^{|R|}}}$? (We think not. See § 9.3).

Can it be shown that one never needs coefficient ratios larger than $2^{|\phi|}$ for any ϕ ? Make more precise the relations between coefficient sizes and ratios. Can it be shown that the bounds obtained by assuming integer coefficients give bounds on the precisions required of arbitrary real coefficients? Can you establish linear bounds for coefficients for the predicates in Chapter 7?

The linear threshold predicate

$$\psi_{EQ} = \left[\sum 2^i (\psi_i - x_i) > 0 \right]$$

is very much like those obtained by the stratification-theorem method, in that at each level, i , the coefficient is chosen to dominate the worst case of summation of the coefficients of preceding levels. The result of theorems of §9.1 and §9.2 is that for those predicates there do not exist any linear forms with smaller coefficients, and this suggests to us that (with respect to given ϕ 's) perhaps there is a sense in which some predicates are inherently stratified. We don't have any strong ideas about this, except to point out that there is a serious shortage of computer-oriented concepts for classification of patterns. We do not know, for most of the cases in Chapter 7, which of them really require the stratification-like coefficient growth: that is to say, we don't have any general method to detect "inherent stratification."

9.3 Predicate With Possibly Maximal Coefficients

Define $||X||$ to be the index of X in an ordering of all the subsets of R . We will consider the simple predicate $\psi_{||PAR||} = \lceil ||X|| \text{ is odd} \rceil$ with respect to the following set $\phi = \{\phi_{||X||}\}$ of predicates:

$$\alpha_i(X) = \begin{cases} 0 & \text{if } ||X|| < i \\ 1 & \text{if } ||X|| = i \\ (||X|| - i) \bmod 2 & \text{if } ||X|| > i \end{cases}$$

Then $\psi_{||PAR||}$ is in $L(\phi)$ and is in fact realized by

$$\psi_{||PAR||} = \left\lceil \sum (-1)^i F_i \alpha_i < 0 \right\rceil$$

where F_i is the i -th Fibonacci number:

$$\{F_i\} = \{1, 1, 2, 3, 5, 8, 13, \dots\}.$$

Theorem: any form in $L(\phi)$ for $\psi_{||PAR||}$ must have coefficients at least this large; the F_n grow approximately as

$$\frac{1}{\sqrt{5}} \left(\frac{\sqrt{5} + 1}{2} \right)^n \sim 2^{0.7n}$$

so that the largest coefficient is then like

$$\sim 2^{\frac{3}{2}(\sqrt{5} + 1)} \cdot 2^n$$

The proof of the theorem can be inferred by studying the array below:

		$ X_1 =$	1	2	3	4	5	6	7	8	9	...
$i =$	$\alpha_i =$											
1	-1	1	1	0	1	0	1	0	1	0	...	
2	+1		1	1	0	1	0	1	0	1	...	
3	-2			1	1	0	1	0	1	0	...	
4	+3	$\vdots$			1	1	0	1	0	1	...	
5	-5					1	1	0	1	0	...	
6	+8	...					1	1	0	1	...	
7	-13	0	0	...				1	1	0	...	
		0	0	0	...				1	1	0	...
		0	0	0	0	...				1	1	...

and it can be seen if $\alpha_1 < 0$ and the coefficients are integers then

$$\alpha_{2i+1} < - \sum_{j=1}^i \alpha_{2j}$$

and

$$\alpha_{2i} \geq - \sum_{j=1}^i \alpha_{2j-1}$$

and the reader can verify that this implies that for all α_i ,

$$|\alpha_i + 1| \geq |\alpha_i| + |\alpha_{i-1}|$$

Discussion and conjecture

This predicate and its Φ has the same quality as that in Chapter 9.2 - that the Φ 's themselves are each almost the desired predicate. Note also that by properly ordering the subsets, we can still choose

$$\psi ||\text{PAR}|| = \psi_{\text{PAR}}$$

We conjecture that this example is a worst case: to be precise, if Φ contains $|\Phi|$ elements, the maximal coefficient growth cannot be faster than

$$\frac{(\frac{\sqrt{5}+1}{2})^{|\Phi|}}{2}$$

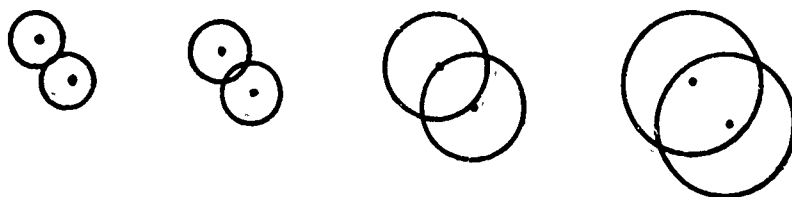
where the exponent constant is the Fibonacci, or golden rectangle ratio. Our conjecture is based only on arguments too flimsy to write down.*

9.4 The Group Invariance Theorem and Bounded Coefficients On The Infinite Plane

In § 7.10 we noted a counterexample to extending the group invariance theorem (§2.3) to infinite retinas. The difficulty came through using an infinite stratification that leads to unbounded coefficients. This in turn raises convergence problems for the symmetric summations used to prove equal the coefficients within an equivalence-class. If the coefficients are bounded, and the group contains a translation group, we can prove the corresponding theorem. We do not know stronger results: presumably there is a better theorem with

* Such as the fact that $\sqrt{5}$ usually occurs in upper bounds in the theories of rational approximations and geometry of numbers.

(a) a summability-type condition on the coefficients and (b) a weaker structure condition on the group). The proof depends on the geometric fact that for increasing concentric circles about two fixed centers



the proportion of area in common approaches unity.

9.4.1 Bounded Coefficients and Group Invariance

Let ψ be a predicate invariant under translation of the infinite plane.

Theorem: if the coefficients of the ϕ 's are bounded in each equivalence class, then there exists an equivalent perception with coefficients equal in each equivalence class.

Proof: Let T_c be the set of translations with displacements less than some distance C . Let $\psi = \left\lceil \sum \alpha_\phi \phi - \theta \right\rceil$.

Then define

$$\begin{aligned} \psi_c(X) &= \left\lceil \sum_{g \in T_c} \left(\sum \alpha_\phi \phi(gx) - \theta \right) \right\rceil = \left\lceil \sum_\phi \phi(X) \sum_g \alpha_\phi g^{-1} \phi - \sum_g \theta \right\rceil \\ &= \left\lceil \sum_\phi \phi(X) \sum_g \alpha_\phi \frac{g\phi}{g} - \sum_g \theta \right\rceil \end{aligned}$$

because T_C is closed under the group inverse. By the argument of § 2.3 each ψ_C is equivalent to ψ as a predicate. The following lemma shows that we can select an increasing sequence $C_1, C_2, \dots$, for which the limit of the average

$$\lim_{i \rightarrow \infty} \frac{1}{|C_i|} \sum \alpha_{\phi g}$$

has the same value independent of ϕ within every equivalence class.

LEMMA: Suppose some function $f(x)$ is bounded, i.e., $|f(x)| < M$, in E^2 .

Then there exists a sequence of increasing radii R_i such that

$$\lim_{R_i \rightarrow \infty} \left| \frac{1}{2\pi R_i} \int_{C_i} f(y) dA \right| < M$$

has value independent of the selection of the common center C_i , if the limit exists for any center at all.

Proof: Choose as center the origin and any sequence of R_i 's increasing without bound. Then for each i we have

$$\left| \frac{1}{2\pi R_i} \int_{C_i} f(y) dA \right| < M$$

so by compactness we can choose a convergent subsequence; call this $\{R_i'\}$, now. Now given any other center x' for the circles, note that

$$\left(\frac{1}{2\pi R_i'} \left| \int_{C_i'} f(y) dA - \int_{C_i'} f(x' + y) dA \right| < 2 \cdot M \cdot \frac{\Delta_i(x')}{2\pi R_i'} \right)$$

where $\Delta_i(x')$ is the area of overlap between the original C_i' and the new C_i' centered around x' . But as the radius grows, for any x'

$$\lim_{i \rightarrow \infty} \left(\frac{\Delta_i(x')}{R_i'} \right) = 0$$

Q.E.D.

The limit will exist except under the most peculiar conditions on the behavior of $f(x)$ at infinity.

To prove the main theorem, we simply choose a representative ϕ from some equivalence class, and set

$$f(g) = \alpha_{g\phi}$$

regarding g as a translation from the origin.

It follows that the perceptron obtained in § 7.4 must have unbounded coefficients, and that there is no equivalent member of $L(\phi)$ with bounded coefficients.

Note: The methods of § 9.2 and § 9.3 are similar to those used by Mghill and Kautz [1961]* to find maximal coefficients for the order-1 case. They show that with integer coefficients there is an order 1 predicate for which some coefficient exceeds $\frac{2}{e} \cdot \frac{1}{n} \cdot 2^n$.

* Mghill, J. and Kautz, W.H., "On the size of weights required for Linear-Input Switching Functions," IRE Trans. on Electronic Computers, June 1961, pp. 288 - 290.

CHAPTER 10: LEARNING

10.0 Introduction

Suppose one wants a machine that "discriminates" between two sets

$$\{P\} = P_1, \dots, P_p \text{ and } \{Q\} = Q_1, \dots, Q_q$$

of figures. Assuming that a set Φ of predicates is available, one wants to find the coefficients of a function ψ_{PQ} in $L(\Phi)$ with the property that for every k ,

$$\psi_{PQ}(P_k) = 1 \text{ and } \psi_{PQ}(Q_k) = 0.$$

That is, we would like to find a set $\{\alpha_\varphi\}$ of coefficients such that

$$[\sum \alpha_\varphi \varphi(P_k)] > 0 \text{ but } [\sum \beta_\varphi \varphi(Q_k)] < 0.$$

But suppose further that for some reason we don't want to design the machine especially for this job, perhaps

- i) because we have to build the machine before we are told what $\{P\}$ and $\{Q\}$ are, or
- ii) because the job may be changed at a later time, or even
- iii) because we don't have a good enough analytical definition of $\{P\}$ and $\{Q\}$, hence can't think of a theoretical way to calculate the $\{\alpha_\varphi\}$.

Then it becomes tempting to consider building a machine that itself can accept information and calculate an appropriate set of coefficients--in short,

a machine that "learns."

In the first few sections of this chapter we will develop the theory of a particularly simple and elegant learning machine that calculates coefficients when it is given a sequence of P's and Q's and told which class each is in. It is just about the simplest machine that might be said to be able to "learn," and further on we will discuss its efficiency, and range of capability in relation to some more sophisticated concepts of "learning machines."

Because we are now concerned more with the sets of coefficients $\{\alpha_\varphi\}$ than with the nature of Φ itself, it will be convenient to think of the functions in $L(\Phi)$ as associated with the sets $\{\alpha_\varphi\}$ regarded as vectors, and we will make heavy use of the geometry of the vector-space whose base vectors are the φ 's in Φ , and with coefficients usually the integers.

Warning: the vector-space base is the set of φ 's, and not the points of $\mathbb{R}^1$. Also, in this chapter we will think of the forms $\sum \alpha_i \varphi_i$ as elements of a vector space; one should remember that the set $L(\Phi)$ of φ 's isn't a vector space, and that for each $\varphi \in L(\Phi)$ there are many α -vectors*. (In fact, though it is not

* It may be observed that vector geometry occurs only in this chapter of this book. In the general perceptron literature, vector geometry is the chief mathematical tool (followed closely by statistics--which also plays a small role in our development.) If we were to volunteer one chief reason why so little was learned about perceptrons in the decade that they have been studied, we would point toward the use of this vector geometry. For in thinking about the $\sum \alpha_i \varphi_i$'s as vectors, the relations between the patterns $\{X\}$ and the predicates in $L(\Phi)$ have become very obscure. The α -vectors are not linear operators on the patterns themselves; they are co-operators, that is, they operate on spaces of functional operators on the patterns. Since the bases-- Φ -classes--of their vector spaces are arbitrary, one can't hope to use them to discover much about the kinds of predicates that will lie in an $L(\Phi)$. The important questions aren't about the linear properties of the $L(\Phi)$'s, but about the orders of complexities in computing pattern qualities from the information in the $\{\varphi(X)\}$ set itself.

important here, the set of sets $\{\alpha\}$ that define a given ψ forms a convex "linear hypercone," i.e., a convex set closed under scalar multiplication.)

With these warnings in mind, we can regard any figure (= subset) X of R as determining a vector V_X with components $(\varphi_1(X), \varphi_2(X), \dots, \varphi_n(X))$.

And any predicate ψ in $L(\varphi)$ is determined by at least some vector A_ψ with components

$$(\alpha_1, \alpha_2, \dots, \alpha_n)$$

so that we can write an inner-product expression for $\psi(X)$:

$$\psi(X) \iff A_\psi \cdot V_X > 0.$$

(It is convenient to assume that φ contains the identity function $\varphi(X) \equiv 1$ so that we won't need an explicit threshold θ in our formulae.)

Our goal is to find a discriminating function ψ_{PQ} , or, equivalently an A_{PQ} , with the property

$$A_{PQ} \cdot V_{P_k} > 0 \text{ and } A_{PQ} \cdot V_{Q_k} < 0.$$

For want of a better idea, it occurs to us to try to find A_{PQ} by a "learning program," as follows:

- Step 0: Set $A = (1, 1, \dots, 1)$, or to any other initial value you please!
- Step 1: Choose an element of $\{P_k\}$ or $\{Q_k\}$, call it V .
- Step 2: Compute the sign of $A \cdot V$. If it is correct (i.e., has the proper sign) go back to step 1. Otherwise replace A by $A \pm V$ where the sign is the one $A \cdot V$ should have, and go back to step 1.

The idea is simple: if $A \cdot V$ is too large, the change will result in $(A - V) \cdot V = A \cdot V - |V|^2$ next time, and this has a better chance to be negative. Conversely, if $A \cdot V$ is too negative, $(A + V) \cdot V = A \cdot V + |V|^2$ is more likely to be positive. Thus we have a simple kind of "feedback"--whenever the system makes an error, then A is "reinforced"--that is, slightly modified--in a direction designed to correct the error!

Will it work? It seems terribly simple-minded, because each correction is performed with just one P_k or Q_k in view. Why would one expect general improvement when a correction designed to correct for one P or Q may make A wrong for many others? Indeed, this will happen, especially at the beginning of the process. The remarkable fact is that this procedure will ultimately work: if there exists any A_{PQ} at all then the procedure will eventually find one*. (Then it will continue to be correct, so will remain in Step (1) above.)

And there is no constraint on the choice in Step (1) on the V 's, say that

* That it finds its own solution rather than the one we had in mind might make one think, according to one's attitude, that perceptrons are either more random or more original than they seemed. A better statement might be: If there is a solution cone (and there is at most one) then the perceptron will find some point near its boundary.

every P_k and Q_k will eventually recur--sufficiently often--a very weak condition.

This is the celebrated "Perceptron Convergence Theorem" apparently first conjectured by Rosenblatt.

We will prove the theorem shortly. First we must make a few stipulations about the geometry. According to our conventions, the components of the A-vectors are real numbers--the same as our α_i 's. (They could be assumed to be integers without loss of generality.) The V-vectors as defined above had only the values 0 and 1 of the ϕ_i 's, but there is no reason to so restrict them, and we will allow arbitrary coefficients for them as well. This makes it convenient to dispose of the differences between the $\{P_k\}$ and $\{Q_k\}$ sets, and we can simplify our formulation as follows:

Define, for $1 \leq k \leq p$

$$f_k = (\phi_1(P_k), \phi_2(P_k), \dots, \phi_n(P_k))$$

and, for

$$p+1 \leq k \leq p+q$$

$$f_k = (-\phi_1(Q_{k-p}), -\phi_2(Q_{k-p}), \dots, -\phi_n(Q_{k-p})).$$

The negatives on the Q-vectors reduce the problem to the simplest form: given a set $f_1, \dots, f_k, \dots, f_{p+q}$ of vectors find a A-vector A_ψ for which

$$A_\psi \cdot f_k > 0$$

for all k . The difference between the P 's and Q 's has been removed.

The convergence theorem will follow as an easy consequence of three simple Lemmas:

10.1 Geometric Lemmas

Lemma 1: Let $V_1, \dots, V_n, \dots$ be a sequence of vectors such that there is a bound b on their lengths:

$$(i) \quad |V_i| < b$$

and for all i

$$(ii) \quad V_i \cdot S_i \leq 0$$

where S_i is defined to be the vector sum

$$S_i = V_1 + V_2 + \dots + V_{i-1}.$$

Then

$$|S_n| < b\sqrt{n}$$

for every n .

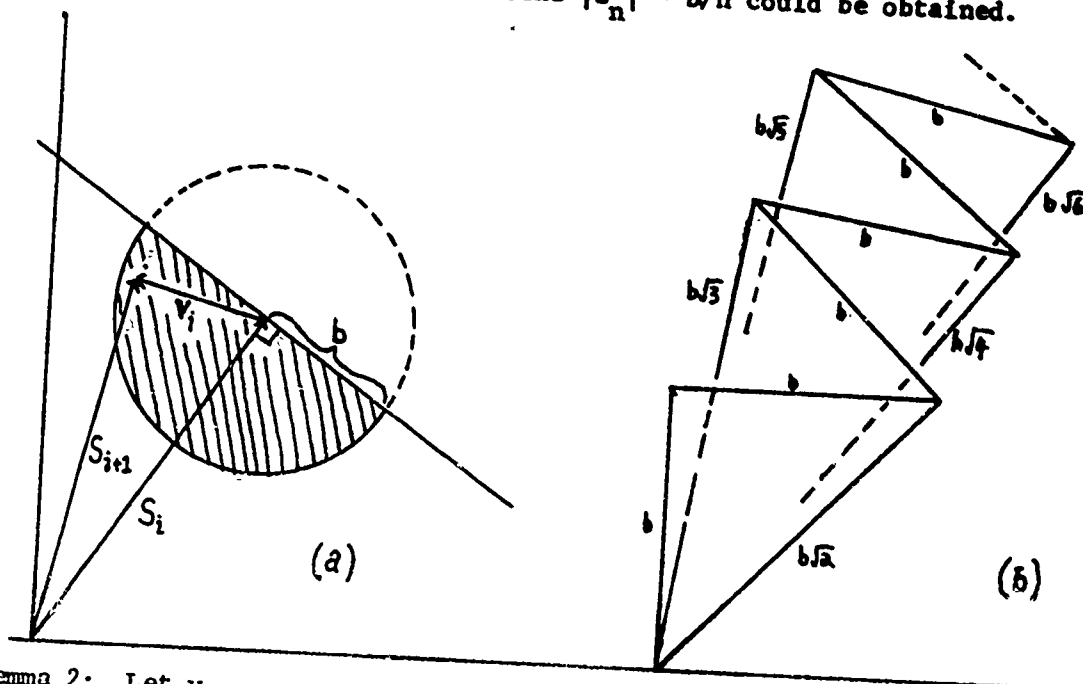
Proof:

$$\begin{aligned} |S_{i+1}|^2 &= |S_i + V_i|^2 \\ &= |S_i|^2 + 2S_i \cdot V_i + |V_i|^2 \\ &< |S_i|^2 + 2 \cdot 0 + b^2 \end{aligned}$$

by i) and ii). Hence, by induction, $|S_n|^2 < n b^2$.

Q.E.D.

Geometrically this is very simple. The conditions i) and ii) state that S_{i+1} must lie in the shaded semicircle in Fig. (a). Then Fig. (b) shows the extreme case in which the bound $|S_n| = b/\sqrt{n}$ could be obtained.



Lemma 2: Let $v_1, \dots, v_n \dots$ be an infinite sequence of vectors such that for some fixed vector A the inner products are bounded away from 0:

$$(i) \quad 0 < d < v_i \cdot A.$$

Then there is a constant c such that, for all n

$$|v_1 + \dots + v_n| > c \cdot n,$$

i.e., the sum grows at least linearly with n .

Proof: $(v_1 + \dots + v_n) \cdot A = v_1 \cdot A + \dots + v_n \cdot A \geq n \cdot d$, by i)

and

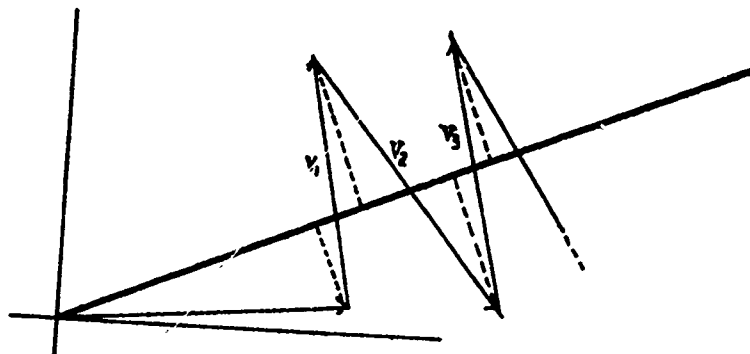
$$(v_1 + \dots + v_n) \cdot A \leq |v_1 + \dots + v_n| \cdot |A|$$

by the Cauchy-Schwartz inequality. Hence

$$|v_1 + \dots + v_n| > n \cdot \frac{d}{|A|}.$$

Q.E.D.

Here the geometry is trivial: the projection of each v_i in the A -direction must exceed $\frac{d}{|A|}$, so their sum must grow at least linearly in n .

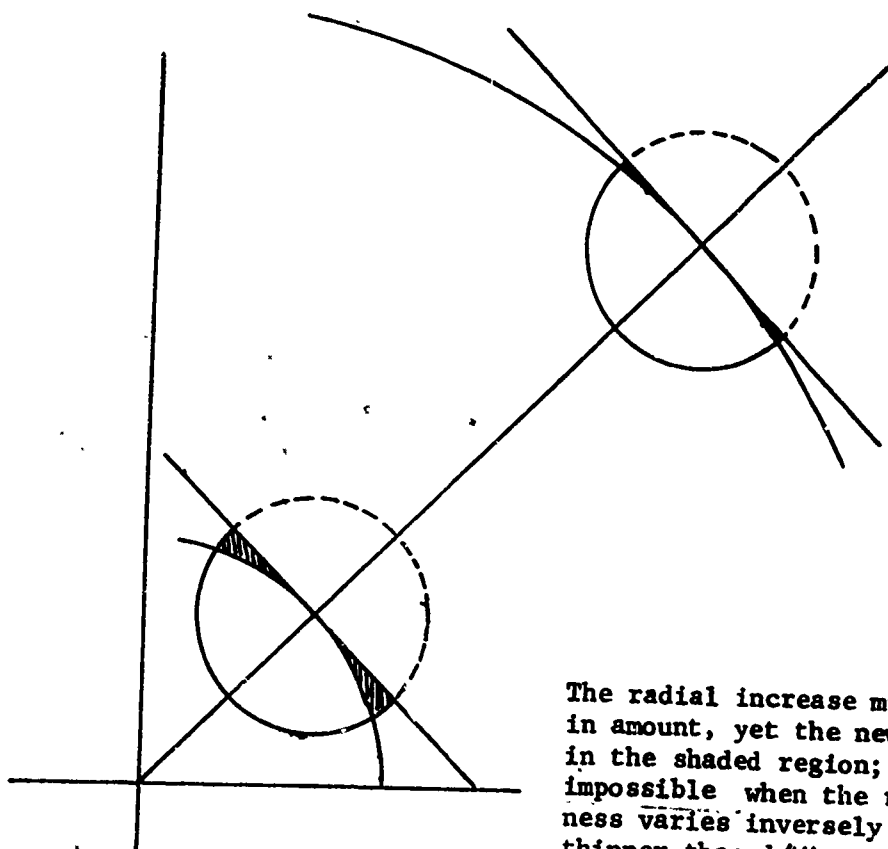


Lemma 3: No infinite sequence $v_1, \dots, v_n \dots$ can satisfy for all n the conditions of both Lemma 1 and Lemma 2.

Proof: For this would require positive constants b and c for which

$$c \cdot n < b \sqrt{n}$$

for arbitrarily large n .



The radial increase must be at least $d/|A|$ in amount, yet the new vector must remain in the shaded region; this becomes impossible when the region, whose thickness varies inversely with R , becomes thinner than $d/|A|$.

10.2 The Convergence Theorem

Let $F = f_1, \dots, f_n$ be a finite collection of vectors and that there exists a vector A such that for all i

$$f_i \cdot A > 0.$$

Since F is finite there must exist a number d such that for all i

$$0 < d < f_i \cdot A.$$

The learning process is described by the following Program:

- 0: Set j to 1. Set S_k to $\vec{0}$.
- 1: Choose an element of F and call it f .
- 2: If $f \cdot S_j > 0$, go back to Step 1. (Otherwise $f \cdot S_j \leq 0$.)
- 3: Define v_j to be f , and define S_{j+1} to be $S_j + v_j$.
Replace j by $j + 1$ and go to Step 1.

Theorem 10.2: The sequence $v_1, \dots, v_j, \dots$ produced by the Program cannot be infinite.

Proof: The $\{v_j\}$ satisfy the conditions of Lemmas 1 and 2, so Lemma 3 applies.

Corollary: The number of errors the Program can ever make is bounded above

by (Lemma 3): $c \cdot n < b\sqrt{n}$ and by (Lemma 1): $\frac{d}{|A|} n < \max_i |f_i| \sqrt{n}$, so

$$n \leq \frac{|A|^2 \max_i |f_i|^2}{\min_i |f_i \cdot A|^2}.$$

The point is that once the program can make no more errors, it must have "converged" to a solution vector. (This might seem to be a peculiar argument--like trusting a surgeon because he has attained his full year's mortality in the first month--but some reflection should show the difference.)

10.3 Application: Learning the Parity Function ψ_{PAR}

Consider the problem of reinforcement-learning the coefficients (over the set Φ of masks) for ψ_{PAR} as described in §9.1. The coefficient vector must grow to length

$$|s_n|^2 = 2^{2n} + \binom{n}{1} 2^{2(n-1)} + \dots + \binom{n}{0} 1^2 = \sum_{i=0}^n \binom{n}{i} 2^{2i}$$

$$= (1 + 2^2)^n = 5^n$$

and the maximal $\hat{x}$ vector has length

$$|x|^2 = 2^n$$

because it can have 2^n non-zero terms. Now Lemma 1 tells us that if T is the minimal number of learning reinforcements, it is bounded below by

$$|s_n|^2 < T \cdot |x|^2$$

or

$$T > \left(\frac{5}{2}\right)^n = (\text{Lower bound}).$$

Thus Lemma 1 can be used to give a lower estimate on learning time (provided one knows a minimal length of a solution coefficient vector).

Unfortunately, we know little about the real learning times. For $\hat{x}_{PAR}$, since the largest coefficient is $\geq 2^n$ we already know it requires $\geq 2^n = \left(\frac{4}{2}\right)^n$ reinforcements, i.e., 2^n errors in training, for terms that make the largest mask have value 1. This occurs for only one pattern $X = R$. If we imagined a training sequence that cycled through all possible X 's, or selected them uniformly at random, this would average one of every 2^n trials, so the training sequence should have to persist for 2^{2n} trials. Even this is

hopelessly optimistic, for the $n-1$ degree terms all have to be forced to $\leq -2^{n-1}$ in the negative direction, and we believe that the learning time for ψ_{PAR} on $|R| = n$ is at least 2^{3n} , but have no good method for finding out. In any case, one is not encouraged to try to "learn" parity with perceptrons! Finally, the corollary 11.2 tells us that

$$T < \frac{5^n \cdot 2^n}{1} \quad (\text{Upper bound})$$

so we have the number of errors and corrections made while learning ψ_{PAR} pinched* between

$$\left(2\frac{1}{2}\right)^n < T < (10)^n.$$

* Suppose parity is learned somehow. Then the largest mask will get 2^n reinforcements. This will cause $n \cdot 2^n$ false positive increments at level $n-1$. These can be corrected only by reinforcing them, each $\frac{3}{2} 2^n$ times negatively. At level $n-2$ we will have accumulated a score of $2^n - 2 \cdot \frac{3}{2} 2^n = -2 \cdot 2^n$ and each will need $\frac{9}{4} \cdot 2^n$ reinforcements to bring it up to 2^{n-2} . There are $\binom{n}{2}$ of these. This estimate leads to

$$\binom{n}{0} \left(\frac{3}{2}\right)^0 2^n + \binom{n}{1} \left(\frac{3}{2}\right)^1 2^n + \binom{n}{2} \left(\frac{3}{2}\right)^2 2^n + \dots = \left(\frac{5}{2}\right)^n 2^n$$

= 5^n reinforcements.

But this may be an underestimate because it does not include provision for forcing the machine to be wrong enough to get the reinforcements.

Chapter 11 GEOMETRIC PREDICATES AND SERIAL ALGORITHMS

11.0 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-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 (§9) and the following "best possible" result for the serial computation of parity. Let $x_1, x_2, \dots, x_n$ 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.

11.1 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 $x_1, \dots, x_{|R|}$ of the points of R is: X is connected when each point x_i in X , except the first point in X has 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 .

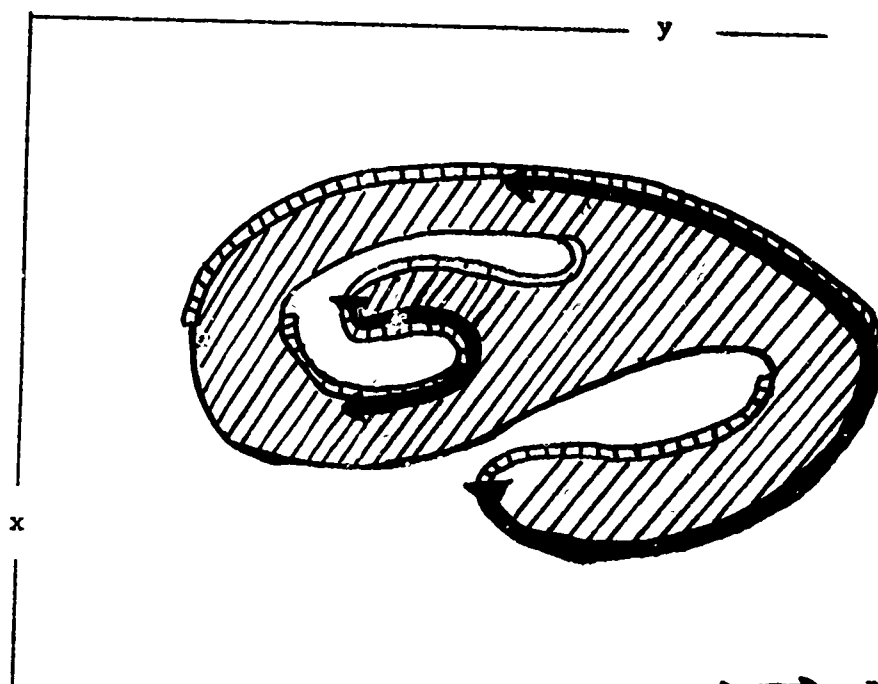
Analysis

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

TRACE will return to x_1 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 non-empty 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 are examined more than three times.



Boundary points read
by TRACE



Boundary points read
by SCAN

11.2 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 $x_1, \dots, x_{2^n \cdot 2^n}$ be an enumeration of the points of R in the order

$$\begin{array}{cccc} 1 & 2^{n+1} & \dots & (2^n+1)2^{n+1} \\ 2 & 2^{n+2} & \dots & (2^n+2)2^{n+2} \\ \vdots & \vdots & \ddots & \vdots \\ 2^n & 2 \cdot 2^n & \dots & 2^n \cdot 2^n \end{array}$$

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	$\dots n \dots$	I_x	$\dots n \dots$	J_y	$\dots n \dots$	J_x	$\dots n \dots$	K
-------	-----------------	-------	-----------------	-------	-----------------	-------	-----------------	-----

where " $\dots n \dots$ " 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 the "...n..." intervals. For example, " $i = |R|$ " if and only if there are all zeros in the "...n..."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 non-trivial 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 "...n..." 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 co. xity 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 $\lceil |X| \text{ is prime} \rceil$ 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 are known to have straightforward geometric interpretations.