Marvin Minsky and Seymour Papert

Perceptrons

An Introduction to Computational Geometry

Expanded Edition

0

0.0 Readers

In writing this we had in mind three kinds of readers. First, there are many new results that will interest specialists concerned with "pattern recognition," "learning machines," and "threshold logic." Second, some people will enjoy reading it as an essay in abstract mathematics; it may appeal especially to those who would like to see geometry return to topology and algebra. We ourselves share both these interests. But we would not have carried the work as far as we have, nor presented it in the way we shall, if it were not for a different, less clearly defined, set of interests.

The goal of this study is to reach a deeper understanding of some concepts we believe are crucial to the general theory of computation. We will study in great detail a class of computations that make decisions by weighing evidence. Certainly, this problem is of great interest in itself, but our real hope is that understanding of its mathematical structure will prepare us eventually to go further into the almost unexplored theory of parallel computers.

The people we want most to speak to are interested in that general theory of computation. We hope this includes psychologists and biologists who would like to know how the brain computes thoughts and how the genetic program computes organisms. We do not pretend to give answers to such questions—nor even to propose that the simple structures we shall use should be taken as "models" for such processes. Our aim—we are not sure whether it is more modest or more ambitious—is to illustrate how such a theory might begin, and what strategies of research could lead to it.

It is for this third class of readers that we have written this introduction. It may help those who do not have an immediate involvement with it to see that the theory of pattern recognition might be worth studying for other reasons. At the same time we will set out a simplified version of the theory to help readers who have not had the mathematical training that would make the later chapters easy to read. The rest of the book is self-contained and anyone who hates introductions may go directly to Chapter 1.

0.1 Real, Abstract, and Mythological Computers

We know shamefully little about our computers and their computations. This seems paradoxical because, physically and logically,

computers are so lucidly transparent in their principles of operation. Yet even a school boy can ask questions about them that today's "computer science" cannot answer. We know very little, for instance, about how much computation a job should require.

As an example, consider one of the most frequently performed computations: *solving a set of linear equations*. This is important in virtually every kind of scientific work. There are a variety of standard programs for it, which are composed of additions, multiplications, and divisions. One would suppose that such a simple and important subject, long studied by mathematicians, would by now be thoroughly understood. But we ask, How many arithmetic steps are absolutely required? How does this depend on the amount of computer memory? How much time can we save if we have *two* (or n) identical computers? Every computer scientist "knows" that this computation requires something of the order of n^3 multiplications for n equations, but even if this be true no one knows—at this writing—how to begin to prove it.

Neither the outsider nor the computation specialist seems to recognize how primitive and how empirical is our present state of understanding of such matters. We do not know how much the speed of computations can be increased, in general, by using "parallel" as opposed to "serial"—or "analog" as opposed to "digital"—machines. We have no theory of the situations in which "associative" memories will justify their higher cost as compared to "addressed" memories. There is a great deal of folklore about this sort of contrast, but much of this folklore is mere superstition; in the cases we have studied carefully, the common beliefs turn out to be not merely "unproved"; they are often drastically wrong.

The immaturity shown by our inability to answer questions of this kind is exhibited even in the language used to formulate the questions. Word pairs such as "parallel" vs. "serial;" "local" vs. "global," and "digital" vs. "analog" are used as if they referred to well-defined technical concepts. Even when this is true, the technical meaning varies from user to user and context to context. But usually they are treated so loosely that the species of computing machine defined by them belongs to mythology rather than science.

Now we do not mean to suggest that these are mere pseudo-problems that arise from sloppy use of language. This is not a book of "therapeutic semantics"! For there *is* much content in these intuitive ideas and distinctions. The problem is how to capture it in a clear, sharp theory.

0.2 Mathematical Strategy

We are not convinced that the time is ripe to attempt a very general theory broad enough to encompass the concepts we have mentioned and others like them. Good theories rarely develop outside the context of a background of well-understood real problems and special cases. Without such a foundation, one gets either the vacuous generality of a theory with more definitions than theorems—or a mathematically elegant theory with no application to reality.

Accordingly, our best course would seem to be to strive for a *very thorough* understanding of well-chosen particular situations in which these concepts are involved.

We have chosen in fact to explore the properties of the simplest machines we could find that have a clear claim to be "parallel"—for they have no loops or feedback paths—yet can perform computations that are nontrivial, both in practical and in mathematical respects.

Before we proceed into details, we would like to reassure non-mathematicians who might be frightened by what they have glimpsed in the pages ahead. The mathematical methods used are rather diverse, but they seldom require advanced knowledge. We explain most of that which goes beyond elementary algebra and geometry. Where this was not practical, we have marked as *optional* those sections we feel might demand from most readers more mathematical effort than is warranted by the topic's role in the whole structure. Our theory is more like a tree with many branches than like a narrow high tower of blocks; in many cases one can skip, if trouble is encountered, to the beginning of the following chapter.

The reader of most modern mathematical texts is made to work unduly hard by the authors' tendency to cover over the intellectual tracks that lead to the discovery of the theorems. We have

tried to leave visible the lines of progress. We should have liked to go further and leave traces of all the false tracks we followed; unfortunately there were too many! Nevertheless we have occasionally left an earlier proof even when we later found a "better" one. Our aim is not so much to prove theorems as to give insight into methods and to encourage research. We hope this will be read not as a chain of logical deductions but as a mathematical novel where characters appear, reappear, and develop.

0.3 Cybernetics and Romanticism

The machines we will study are abstract versions of a class of devices known under various names; we have agreed to use the name "perceptron" in recognition of the pioneer work of Frank Rosenblatt. Perceptrons make decisions—determine whether or not an event fits a certain "pattern"—by adding up evidence obtained from many small experiments. This clear and simple concept is important because most, and perhaps all, more complicated machines for making decisions share a little of this character. Until we understand it very thoroughly, we can expect to have trouble with more advanced ideas. In fact, we feel that the critical advances in many branches of science and mathematics began with good formulations of the "linear" systems, and these machines are our candidate for beginning the study of "parallel machines" in general.

Our discussion will include some rather sharp criticisms of earlier work in this area. Perceptrons have been widely publicized as "pattern recognition" or "learning" machines and as such have been discussed in a large number of books, journal articles, and voluminous "reports." Most of this writing (some exceptions are mentioned in our bibliography) is without scientific value and we will not usually refer by name to the works we criticize. The sciences of computation and cybernetics began, and it seems quite rightly so, with a certain flourish of romanticism. They were laden with attractive and exciting new ideas which have already borne rich fruit. Heavy demands of rigor and caution could have held this development to a much slower pace; only the future could tell which directions were to be the best. We feel, in fact, that the solemn experts who most complained about the "exaggerated claims" of the cybernetic enthusiasts were, in the balance, much more in the wrong. But now the time has come for maturity, and this requires us to match our speculative enterprise with equally imaginative standards of criticism.

0.4 Parallel Computation

The simplest concept of parallel computation is represented by the diagram in Figure 0.1. The figure shows how one might compute a function $\psi(X)$ in two stages. First we compute *independently* of one another a set of functions $\varphi_1(X)$, $\varphi_2(X)$, ..., $\varphi_n(X)$ and then combine the results by means of a function Ω of n arguments to obtain the value of ψ.

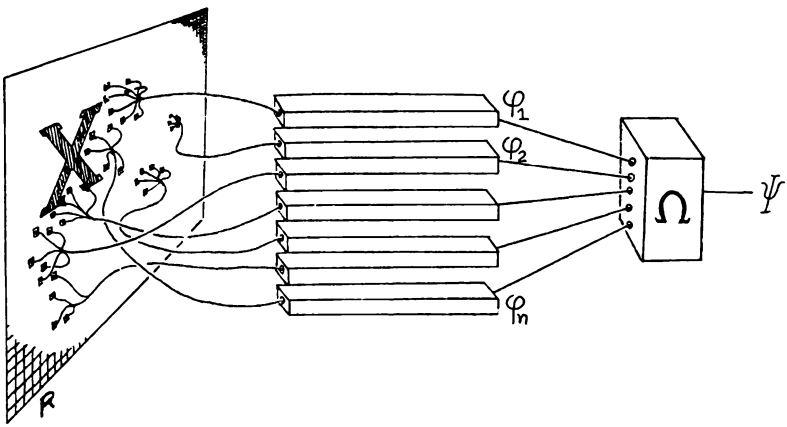


Figure 0.1

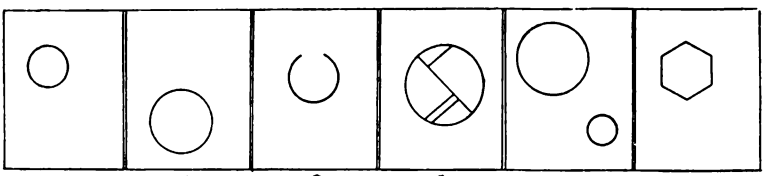
To make the definition meaningful—or, rather, productive—one needs to place some restrictions on the function Ω and the set Φ of functions $\varphi_1, \varphi_2, \ldots$. If we do not make restrictions,we do not get a theory: any computation ψ could be represented as a parallel computation in various trivial ways, for example, by making one of the φ's be ψ and letting Ω do nothing but transmit its result. We will consider a variety of restrictions, but first we will give a few concrete examples of the kinds of functions we might want ψ to be.

0.5 Some Geometric Patterns; Predicates

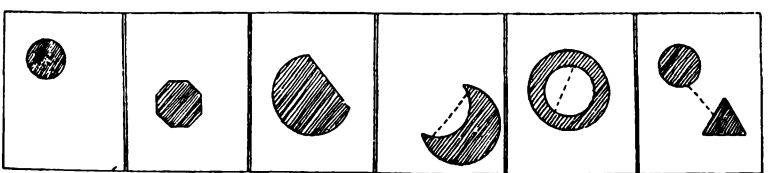
Let R be the ordinary two-dimensional Euclidean plane and let X be a geometric figure drawn on R. X could be a circle, or a pair of circles, or a black-and-white sketch of a face. In general we will think of a figure X as simply a subset of the points of R (that is, the black points).

Let $\psi(X)$ be a function (of figures X on R) that can have but two values. We usually think of the two values of ψ as 0 and 1. But by taking them to be FALSE and TRUE we can think of $\psi(X)$ as a predicate, that is, a variable statement whose truth or falsity depends on the choice of X. We now give a few examples of predicates that will be of particular interest in the sequel.

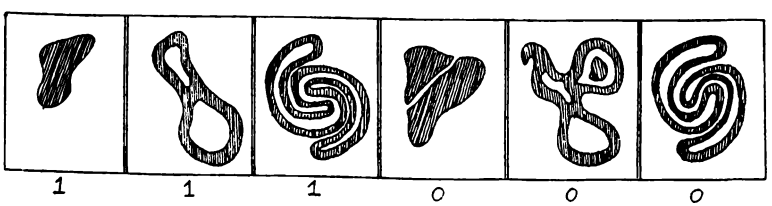
$$\psi_{\text{CIRCLE}}(X) = \begin{cases} 1 \text{ if the figure } X \text{ is a circle,} \\ 0 \text{ if the figure is not a circle;} \end{cases}$$



$$\psi_{\text{CONVEX}}(X) = \begin{cases} 1 \text{ if } X \text{ is a convex figure,} \\ 0 \text{ if } X \text{ is not a convex figure;} \end{cases}$$



$$\psi_{\text{CONNECTED}}(X) = \begin{cases} 1 \text{ if } X \text{ is a connected figure,} \\ 0 \text{ otherwise.} \end{cases}$$

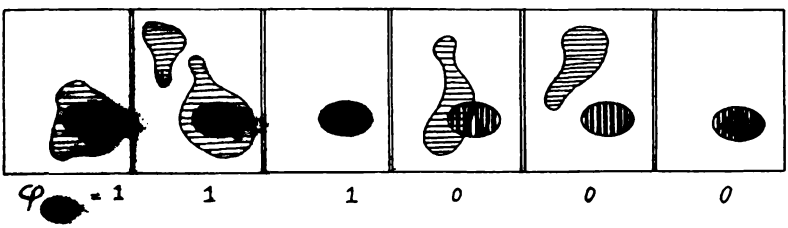


We will also use some very much simpler predicates.* The very simplest predicate "recognizes" when a particular single point is in X: let p be a point in the plane and define

$$\varphi_p(X) = \begin{cases} 1 \text{ if } p \text{ is in } X, \\ 0 \text{ otherwise.} \end{cases}$$

Finally we will need the kind of predicate that tells when a particular set A is a subset of X:

$$\varphi_A(X) = \begin{cases} 1 \text{ if } A \subset X, \\ 0 \text{ otherwise.} \end{cases}$$



0.6 One simple concept of "Local"

We start by observing an important difference between $\psi_{\text{CONNECTED}}$ and ψ_{CONVEX}. To bring it out we state a fact about convexity:

Definition: A set X fails to be convex if and only if there exist three points such that q is in the line segment joining p and r, and

$$\begin{cases} p \text{ is in } X, \\ q \text{ is not in } X, \\ r \text{ is in } X. \end{cases}$$

Thus we can test for convexity by examining triplets of points. If all the triplets pass the test then X is convex; if any triplet fails (that is, meets all conditions above) then X is not convex. Because all the tests can be done independently, and the final decision made by such a logically simple procedure—unanimity of all the tests—we propose this as a first draft of our definition of "local."

*We will use "φ" instead of "ψ" for those very simple predicates that will be combined later to make more complicated predicates. No absolute logical distinction is implied.

Definition: A predicate ψ is conjunctively local of order k if it can be computed, as in §0.4, by a set Φ of predicates φ such that

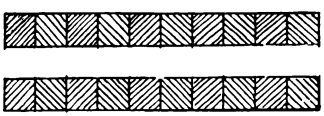
$$\begin{cases} \text{Each } \varphi \text{ depends upon no more than } k \text{ points of } R; \\ \psi(X) = \begin{cases} 1 \text{ if } \varphi(X) = 1 \text{ for every } \varphi \text{ in } \Phi, \\ 0 \text{ otherwise.} \end{cases} \end{cases}$$

Example: ψ_{CONVEX} is conjunctively local of order 3.

The property of a figure being *connected* might not seem at first to be very different in kind from the property of being convex. Yet we can show that:

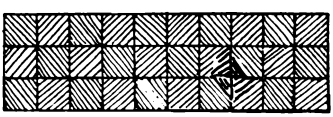
Theorem 0.6.1: $\psi_{\text{CONNECTED}}$ is not conjunctively local of any order.

PROOF: Suppose that $\psi_{\text{CONNECTED}}$ has order k. Then to distinguish between ~~the two figures~~ these two $k+1$-wide figures —



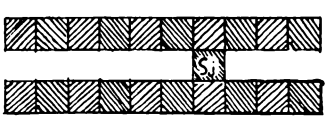
X_0

and



X_1

there must be some φ_0 such that $\varphi_0(X_0) = 0$, because X_0 ~~which has value 0 on X_0 which~~ is not connected. All φ's have value 1 on X_1, which is connected. Now, φ_0 can depend on at most k points, so there must be at least one middle square, say S_j, that does not contain one of these points. But then, on the figure X_2,



X_2

which is connected, φ_0 must have the same value, 0, that it has on X_0. But this cannot be, for all φ's must have value 1 on X_2.

Of course, if some φ is allowed to look at *all* the points of R then $\psi_{\text{CONNECTED}}$ can be computed, but this would go against any concept of the φ's as "local" functions.

0.7 Some Other Concepts of Local

We have accumulated some evidence in favor of "conjunctively local" as a geometrical and computationally meaningful property of predicates. But a closer look raises doubts about whether it is broad enough to lead to a rich enough theory.

Readers acquainted with the mathematical methods of topology will have observed that "conjunctively local" is similar to the notion of "local property" in topology. However, if we were to pursue the analogy, we would restrict the φ's to depend upon all the points inside small circles rather than upon fixed numbers of points. Accordingly, we will follow two parallel paths. One is based on *restrictions on numbers of points* and in this case we shall talk of predicates of *limited order*. The other is based on restrictions of distances between the points, and here we shall talk of *diameter-limited predicates*. Despite the analogy with other important situations, the concept of local based on diameter limitations seems to be less interesting in our theory—although one might have expected quite the opposite.

More serious doubts arise from the narrowness of the "conjunctive" or "unanimity" requirement. As a next step toward extending our concept of *local*, let us now try to separate essential from arbitrary features of the definition of *conjunctive localness*. The intention of the definition was to divide the computation of a predicate ψ into two stages:

Stage I:
The computation of many properties or features φ_α which are each easy to compute, either because each depends only on a small part of the whole input space R, or because they are very simple in some other interesting way.
Stage II:
A decision algorithm Ω that defines ψ by combining the results of the Stage I computations. For the division into two stages to be meaningful, this decision function must also be distinctively homogeneous, or easy to program, or easy to compute.

The particular way this intention was realized in our example ψ_{CONVEX} was rather arbitrary. In Stage I we made sure that the φ_α's were easy to compute by requiring each to depend only upon a few points of R. In Stage II we used just about the simplest im-

aginable decision rule; if the φ's are *unaminous* we accept the figure; we reject it if even a single φ disagrees.

We would prefer to be able to present a perfectly precise definition of our intuitive local-vs.-global concept. One trouble is that phrases like "easy-to-compute" keep recurring in our attempt to formulate it. To make this precise would require some scheme for comparing the complexity of different computation procedures. Until we find an intuitively satisfactory scheme for this, and it doesn't seem to be around the corner, the requirements of both Stage I and Stage II will retain the heuristic character that makes formal definition difficult.

From this point on, we will concentrate our attention on a particular scheme for Stage II—"weighted voting," or "linear combination" of the predicates of Stage I. This is the so-called perceptron scheme, and we proceed next to give our final definition.

0.8 Perceptrons

Let $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$ be a family of predicates. We will say that

ψ is linear with respect to Φ

if there exists a number θ and a set of numbers $\{\alpha_{\varphi_1}, \alpha_{\varphi_2}, \ldots, \alpha_{\varphi_n}\}$ such that $\psi(X) = 1$ if and only if $\alpha_{\varphi_1}\varphi_1(X) + \cdots + \alpha_{\varphi_n}\varphi_n(X) > \theta$. The number θ is called the threshold and the α's are called the coefficients or weights. (See Figure 0.2). We usually write more compactly

$$\psi(X) = 1 \text{ if and only if } \sum_{\varphi \in \Phi} \alpha_\varphi \varphi(X) > \theta.$$

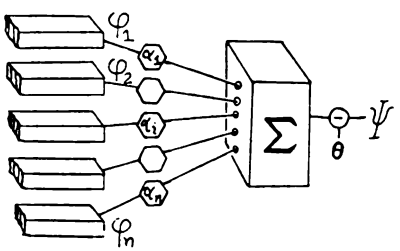


Figure 0.2

The intuitive idea is that each predicate of Φ is supposed to provide some evidence about whether ψ is true for any figure X. If, on the whole, $\psi(X)$ is strongly correlated with $\varphi(X)$ one expects α_φ to be positive, while if the correlation is negative so would be α_φ. The idea of correlation should not be taken literally here, but only as a suggestive analogy.

Example: Any conjunctively local predicate can be expressed in this form by choosing $\theta = -1$ and $\alpha_\varphi = -1$ for every φ. For then

$$\sum (-1)\,\varphi(X) > -1$$

Or one could write (See §1.2.1)
$\sum \varphi(x) = 0$, or $\sum \varphi(X) < 1$.

exactly when $\varphi(X) = 0$ for every φ in Φ. (The senses of TRUE and FALSE thus have to be reversed for the φ's, but this isn't important.)

Example: Consider the seesaw of Figure 0.3 and let X be an arrangement of pebbles placed at *some* of the equally spaced points $\{p_1, \ldots, p_7\}$. Then R has seven points. Define $\varphi_i(X) = 1$ if and only if X contains a pebble at the ith point. Then we can express the predicate

"The seesaw will tip to the right"

by the formula

$$\sum (i - 4)\,\varphi_i(X) > 0,$$

where $\theta = 0$ and $\alpha_i = (i - 4)$.

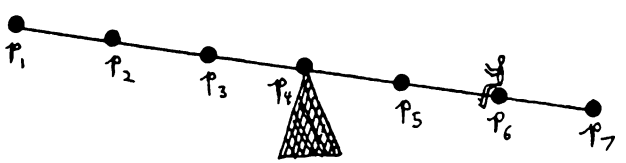


Figure 0.3

There are a number of problems concerning the possibility of infinite sums and such matters when we apply this concept to recognizing patterns in the Euclidean plane. These issues are discussed extensively in the text, and we want here only to reassure the mathematician that the problem will be faced. Except when there is a good technical reason to use infinite sums (and this is sometimes the case) we will make the problem finite by two general methods. One is to treat the retina R as

made up of discrete little squares (instead of points) and treat as equivalent figures that intersect the same squares. The other is to consider only bounded X's and choose Φ so that for any bounded X only a finite number of φ's will be nonzero.

Definition: A perceptron is a device capable of computing all predicates which are linear in some given set Φ of partial predicates.

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. For reasons that will become clear as we proceed, there is little to say about all perceptrons in general. But, by imposing certain conditions and restrictions we will find much to say about certain particularly interesting *families* of perceptrons. Among these families are

1. Diameter-limited perceptrons: for each φ in Φ, the set of points upon which φ depends is restricted not to exceed a certain *fixed diameter* in the plane.

2. Order-restricted perceptrons: we say that a perceptron has order $\leq n$ if no member of Φ depends on more than n points.

3. Gamba perceptrons: each member of Φ may depend on all the points but must be a "linear threshold function" (that is, each member of Φ is itself computed by a perceptron of order 1, as defined in 2 above).

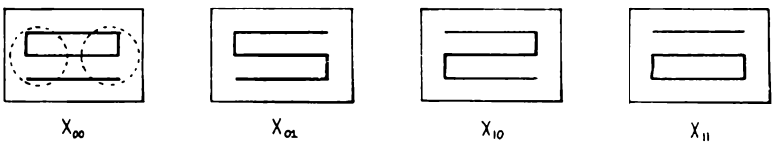
4. Random perceptrons: These are the form most extensively studied by Rosenblatt's group: the φ's are random Boolean functions. That is to say, they are order-restricted and Φ is generated by a stochastic process according to an assigned distribution function.

5. Bounded perceptrons: Φ contains an infinite number of φ's, but all the α_φ lie in a finite set of numbers.

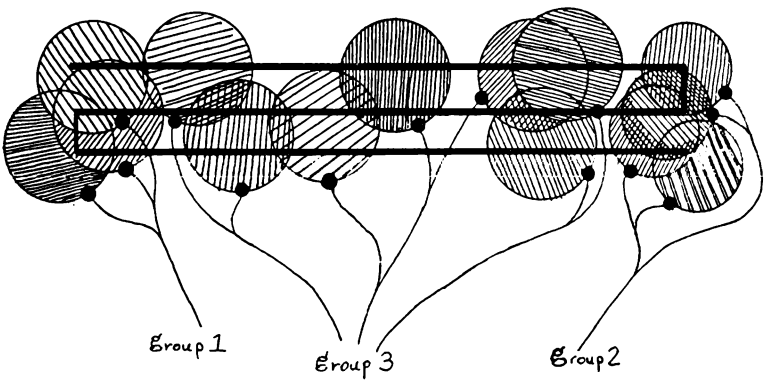
To give a preview of the kind of results we will obtain, we present here a simple example of a theorem about diameter-restricted perceptrons.

Theorem 0.8: No diameter-limited perceptron can determine whether or not all the parts of any geometric figure are connected to one another! That is, no such perceptron computes $\psi_{\text{CONNECTED}}$.

The proof requires us to consider just four figures



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



It is understood that the diameter in question is given at the start, and we *then* choose the X_{ij}'s to be several diameters in length. Suppose that such a perceptron could distinguish disconnected figures (like X_{00} and X_{11}) from connected figures (like X_{10} and X_{01}), according to whether or not

$$\sum \alpha_\varphi \varphi > \theta$$

that is, according to whether or not

$$\left[\sum_{\text{group 1}} \alpha_\varphi \varphi(X) + \sum_{\text{group 2}} \alpha_\varphi \varphi(X) + \sum_{\text{group 3}} \alpha_\varphi \varphi(X) - \theta\right] > 0$$

where we have grouped the φ's according to whether their support sets lie near the left, right, or neither end of the figures. Then for X_{00} the total sum must be negative. In changing X_{00} to X_{10} only $\Sigma_{\text{group 1}}$ is affected, and its value must *increase* enough to make the

total sum become positive. If we were instead to change X_{00} to X_{01} then $\Sigma_{\text{group 2}}$ would have to increase. But if we were to change X_{00} to X_{11}, both $\Sigma_{\text{group 1}}$ and $\Sigma_{\text{group 2}}$ will have to increase by these same amounts since (locally!) the same changes are seen by the group 1 and group 2 predicates, while $\Sigma_{\text{group 3}}$ is unchanged in every case. Hence, net change in the $X_{00} \rightarrow X_{11}$ case must be even more positive, so that if the perceptron is to make the correct decision for X_{00}, X_{01}, and X_{10}, it is forced to accept X_{11} as connected, and this is an error! So no such perceptron can exist.

Readers already familiar with perceptrons will note that this proof—which shows that diameter-limited perceptrons cannot recognize connectedness—is concerned neither with "learning" nor with probability theory (or even with the geometry of hyperplanes in n-dimensional hyperspace). It is entirely a matter of relating the geometry of the patterns to the algebra of weighted predicates. Readers concerned with physiology will note that—insofar as the presently identified functions of receptor cells are all diameter-limited—this suggests that an animal will require more than neurosynaptic "summation" effects to make these cells compute connectedness. Indeed, only the most advanced animals can apprehend this complicated visual concept. In Chapter 5 this theorem is shown to extend also to order-limited perceptrons.

0.9 Seductive Aspects of Perceptrons

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 coefficients α_φ to suitable values. Thus "programming" takes on a pleasingly homogeneous form. Moreover since "programs" are representable as points $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ 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 coefficients to change in the right direction when the machine makes an inappropriate decision. The *perceptron convergence theorems* (see Chapter 11) define conditions under which this procedure is guaranteed to find, eventually, a correct set of values.

0.9.1 Homogeneous Programming and Learning

To separate reality from wishful thinking, we begin by making a number of observations. Let Φ be the set of partial predicates of a perceptron and $L(\Phi)$ the set of all predicates linear in Φ. Thus

$L(\Phi)$ is the repertoire of the perceptron—the set of predicates it can compute when its coefficients α_φ and threshold θ range over all possible values. Of course $L(\Phi)$ could in principle be the set of *all* predicates but this is impossible in practice, since Φ would have to be astronomically large. So any physically 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 did not realistically measure this cost. In particular they neglect the following crucial points:

1. The idea of thinking of classes of geometrical objects (or programs that define or recognize them) as classes of n-dimensional vectors $(\alpha_1, \ldots, \alpha_n)$ loses the geometric individuality of the patterns and leads only to a theory that can do little more than *count* the number of predicates in $L(\Phi)$! This kind of imagery has become traditional among those who think about pattern recognition along lines suggested by classical statistical theories. As a result not many people seem to have observed or suspected that there might be *particular* meaningful and intuitively simple predicates that belong to *no* practically realizable set $L(\Phi)$. We will extend our analysis of $\psi_{\text{CONNECTED}}$ to show how deep this problem can be. At the same time we will show that certain predicates which might intuitively seem to be difficult for these devices *can*, in fact, be recognized by low-order perceptrons: ψ_{CONVEX} already illustrates this possibility.

2. Little attention has been paid to the size, or more precisely, the information content, of the parameters $\alpha_1, \ldots, \alpha_n$. We will give examples (which we believe are typical rather than exceptional) where the ratio of the largest to the smallest of the coefficients is meaninglessly big. Under such conditions it is of no (practical) avail that a predicate be in $L(\Phi)$. In some cases the information capacity needed to store $\alpha_1, \ldots, \alpha_n$ is even greater than that needed to store the whole class of figures defined by the pattern!

3. Closely related to the previous point is the problem of *time of convergence* in a "learning" process. Practical perceptrons are essentially finite-state devices (as shown in Chapter 11). It is therefore vacuous to cite a "perceptron convergence theorem" as assurance that a learning process will eventually find a correct

setting of its parameters (if one exists). For it could do so trivially by cycling through all its states, that is, by trying all coefficient assignments. The significant question is how fast the perceptron learns relative to the time taken by a completely random procedure, or a completely exhaustive procedure. It will be seen that there are situations of some geometric interest for which the convergence time can be shown to increase even faster 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-organizing" systems will provide a generous sample of discussions of "learning" or "adaptive" machines that lack even the degree of rigor and formal definition to be found in the literature on perceptrons. The proponents of these schemes seldom provide any analysis of the range of behavior which can be learned nor do they show much awareness of the price usually paid to make some kinds of learning easy: they unwittingly restrict the device's total range of behavior with hidden assumptions about the environment in which it is to operate.

These critical remarks must not be read as suggestions that we are opposed to making machines that 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 reasonably matched to the task, as they are in Samuel's checker player. A perceptron whose φ's are properly designed for a discrimination known to be of suitably low order will have a good chance to improve its performance adaptively. Our purpose is to explain why there is little chance of much good coming from giving a high-order problem to a quasi-universal perceptron whose partial functions have not been chosen with any particular task in mind.

It may be argued that *people* are universal learning machines and so a counterexample 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 that govern the course of whatever the process of self-programming may be. Moreover, it takes time for us to become universal learners: the sequence of transitions from infancy to intellectual maturity seems rather a confirmation of the

thesis that the rate of acquisition of new cognitive structure (that is, learning) is a sensitive function of the level of existing cognitive structure.

0.9.2 Parallel Computation

The perceptron was conceived as a parallel-operation 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 of them may in fact be relevant to any particular final decision. The *total amount* of computation may become vastly greater than that which would have to be carried out in a well planned sequential process (using the same φ's) whose decisions about what next to compute are conditional on the outcome of earlier computation. Thus the choice between parallel and serial methods in any particular situation must be based on balancing the increased value of reducing the (total elapsed) time against the cost of the additional computation involved.

Even low-order predicates may require large amounts of wasteful computation of information which would be irrelevant to a serial process. This cost may sometimes remain within physically realizable bounds, especially if a large tolerance (or "blur") is acceptable. High-order predicates usually create a completely different situation. An instructive example is provided by $\psi_{\text{CONNECTED}}$. As shown in Chapter 5, *any* perceptron for this predicate on a 100 × 100 toroidal retina *needs* partial functions that *each* look at many hundreds of points! In this case the concept of "local" function is almost irrelevant: the partial functions are themselves global. Moreover, the fantastic number of possible partial functions with such large supports sheds gloom on any hope that a modestly sized, randomly generated set of them would be sufficiently dense to span the appropriate space of functions. To make this point sharper we shall show that for certain predicates and classes of partial functions, the *number* of partial functions that have to be used (to say nothing of the size of their coefficients) would exceed physically realizable limits.

The conclusion to be drawn is that the appraisal of any particular scheme of parallel computation cannot be undertaken rationally without tools to determine the extent to which the problems to be solved can be analyzed into local and global components. The lack of a *general* theory of what is global and what is local is no excuse for avoiding the problem in particular cases. This study will show that it is not impossibly difficult to develop such a theory for a limited but important class of problems.

0.9.3 The Use of Simple Analogue Devices

Part of the attraction of the perceptron lies in the possibility of using very simple physical devices—"analogue computers"—to evaluate the linear threshold functions. It is perhaps generally appreciated that the utility of this 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 (in Chapter 10) the range and sizes of the coefficients in the linear representations of certain predicates. It will be seen that certain ratios can increase faster than exponentially with the number of distinguishable points in R. It follows that for "big" input sets—say, R's with more than 20 points—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 qualifications we 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 limiting factors that will assert themselves as soon as the small model is scaled up to a usable size.*

0.9.4 Models for Brain Function and Gestalt Psychology

The popularity of the perceptron as a model for an intelligent, general-purpose learning machine has roots, we think, in an image of the brain itself as a rather loosely organized, randomly interconnected network of relatively simple devices. This impression in turn derives in part from our first impressions of the bewildering structures seen in the microscopic anatomy of the brain (and probably also derives from our still-chaotic ideas about psychological mechanisms).

In any case the image is that of a network of relatively simple elements, randomly connected to one another, with provision for making adjustments of the ease with which signals can go across the connections. When the machine does something bad, we will "teach" it not to do it again by weakening the connections that were involved; perhaps we will do the opposite to reward it when it does something we like.

The "perceptron" type of machine is one particularly simple version of this broader concept; several others have also been studied in experiments.

The mystique surrounding such machines is based in part on the idea that when such a machine learns the information stored is not localized in any particular spot but is, instead, "distributed throughout" the structure of the machine's network. It was a great disappointment, in the first half of the twentieth century, that experiments did not support nineteenth century concepts of the localization of memories (or most other "faculties") in highly local brain areas. Whatever the precise interpretation of those not particularly conclusive experiments should be, there is no question but that they did lead to a search for nonlocal machine-function concepts. This search was not notably successful. Several schemes were proposed, based upon large-scale fields, or upon "interference patterns" in global oscillatory waves, but these never led to plausible theories. (Toward the end of that era a more intricate and substantially less global concept of "cell-assembly" —proposed by D. O. Hebb [1949]—lent itself to more productive theorizing; though it has not yet led to any conclusive model, its popularity is today very much on the increase.) However, it is not our goal here to evaluate these theories, but only to sketch a picture of the intellectual stage that was set for the perceptron concept. In this setting, Rosenblatt's [1958] schemes quickly took root, and soon there were perhaps as many as a hundred groups, large and small, experimenting with the model either as a "learning machine" or in the guise of "adaptive" or "self-organizing" networks or "automatic control" systems.

The results of these hundreds of projects and experiments were generally disappointing, and the explanations inconclusive. The machines usually work quite well on very simple problems but deteriorate very rapidly as the tasks assigned to them get harder. The situation isn't usually improved much by increasing the size and running time of the system. It was our suspicion that even in those instances where some success was apparent, it was usually due more to some relatively small part of the network, and not really to a global, distributed activity. Both of the present authors (first independently and later together) became involved with a somewhat therapeutic compulsion: to dispel what we feared to be

the first shadows of a "holistic" or "Gestalt" misconception that would threaten to haunt the fields of engineering and artificial intelligence as it had earlier haunted biology and psychology. For this, and for a variety of more practical and theoretical goals, we set out to find something about the range and limitations of perceptrons.

It was only later, as the theory developed, that we realized that understanding this kind of machine was important whether or not the system has practical applications in particular situations! For the same kinds of problems were becoming serious obstacles to the progress of computer science itself. As we have already remarked, we do not know enough about what makes some algorithmic procedures "essentially" serial, and to what extent—or rather, at what cost—can computations be speeded up by using multiple, overlapping computations on larger more active memories.

0.10 General Plan of the Book

The theory divides naturally into three parts. In Part I we explore some very general properties of linear predicate families. The theorems in Part I apply usually to all perceptrons, independently of the kinds of patterns considered; therefore the theory has the quality of algebra rather than geometry. In Part II we look more narrowly at interesting geometric patterns, and get sharper but, of course, less general, theorems about the geometric abilities of our machines. In Part III we examine a variety of questions centered around the potentialities of perceptrons as *practical* devices for pattern recognition and learning. The final chapter traces some of the history of these ideas and proposes some plausible directions for further exploration.

To read this book, one does not have to "know" all the mathematics mentioned in it. Most of the "harder" mathematical sections are "terminal"— they can be skipped without losing the sense of later chapters. The best chapters to skip are §4, §5, §7, and §10.