MODELS   OF   LEARNING   AND   MEMORY

by

Oscar Peter Buneman

Submitted for the degree of Doctor of Philosophy at the

University of Warwick.   November, 1970.

Abstract

The perceptron and the hologram are two dissimilar devices
which have been advanced as neurological models.   It is shown
that there are other and perhaps more plausible models which have
properties common to both of these devices.   The performance of
these intermediate models which are termed Associative Nets is
described and analysed statistically.   The main similarities
and differences between perceptron theory and holography can
also be demonstrated;   it is possible to demonstrate formal links
between the translation invariance in certain types of holography
and group invariance in perception-theory.   Some single proofs
of certain theorems in the latter are also given and some other
learning procedures are formulated.

It is shown that the important difference between these
various models resides in the method used to accomplish a mod-
ification.   If this modification is an analogue of a neurological
change at a synaptic level, then it should be possible to qualify
the relevance of those models by determining what types of synaptic
change can take place in different parts of the nervous system.
Although the evidence is far from complete, it suggests that the
neocortex is limited to having one type of synaptic change.
Finally, each model is discussed in respect of its neurological
plausibility.

Acknowledgements

CHAPTER 1 Introduction

Two theories have been put forward in recent years which, as models of the brain, have enjoyed some popularity. These are holography and the theory of perceptrons and similar learning machines. It is intended to show that between these two superficially dissimilar models there lies a series of alternative theories, some of which are proposed as serious neurological models. An attempt will also be made to examine these theories on the basis of existing neurological evidence and as a result of this examination a proposal is advanced concerning a structural and functional difference between the neocortex and other parts of the nervous system.

In order to demonstrate the formal links between these theories, the first chapters of this dissertation will be concerned with a largely mathematical treatment of these theories. The physiological or psychological rationale for their construction will be stated briefly in this chapter and as a short introduction to some of the theoretical chapters. The detailed investigation of this will be reserved for the later chapters.

Though a great deal of work has recently been directed at discovering a neurological basis for memory or learning (no attempt is made at the moment to define those terms) our understanding is still not very far advanced. This is perhaps partly

due to the limited techniques of experimental neurology; perhaps it is even more due to the fact that we do not know which questions to ask of experimental neurology: that is, we lack good hypotheses about the nervous system. Nevertheless broad neurological considerations allow us to infer for example that the organisation of the brain does not very much resemble the organisation of the hardware of a digital computer.

No less important to our study of behaviour is our intuition about it. The recent advances made in our knowledge of the structure of English are founded more or less wholly on the intuition of native speakers. It is hard to believe that this study will not throw some light on the organisation of the human brain. We might even hope for the reverse if some sound neuro-logical principles of the right kind were to be discovered.

Such considerations have led some neurophysiologists to take holography seriously as a model for memory. While holography, as a brain model, is somewhat condemned later on, many of its features are obviously attractive. Among these are the non local storage of information, the survival of the image after damage to the hologram, the ability to 'recognise' displaced patterns and, in certain cases the recovery of a 'ghost' image. These same considerations led to the development of associative nets and group invariant associative nets which have many of these features in common with holography and other desirable

properties as well. It is felt that such systems could be realised in the nervous system and some evidence is given for this.

Group invariant associative nets were initially developed as an attempt to solve some problems about our use of language, which may involve our ability to recognise the deep structure of one sentence as a part of the deep structure of another. It was felt that such an operation might be accomplished by a parallel system rather than a serial search - which in this case is a very slow procedure. It is far too early to say whether or not such networks are embodied in the nervous system. It is however suggested that they may be relevant to the psychological phenomenon of generalisation in which for example one recognises an object in some position even though one has never seen it in precisely that position before.

One of the most gratifying results of this study was the close relationship it has to some recent results in the theory of perceptrons in which it is possible to say a certain amount about the size and structure of a perceptron that evaluates a predicate which is invariant under certain transformations of its arguments. It also served to underline the fact that the fundamental difference between perceptrons and the other models lies in the way a change is effected in these devices. This is the only real distinction which will be made between models

of memory and models of learning.   Such a distinction will be central to the neurological investigation of the later chapters in which it is suggested that certain parts of the nervous system are more suited to 'remember' and others are better able to 'learn' according to this distinction.   The evidence for this is, of course, rather scanty but it does result in some hypotheses about the nervous system which could be tested by present neurophysiological and neuroanatomical techniques.

A final chapter will be devoted to speculation about how the brain can use a system which remembers or learns.   It is unsatisfactory to produce a theory of memory unless one has some notion of a schema of behaviour into which it might fit. The relevance of group invariant associative nets to language and·forms of simple generalisation will be discussed along with the relation between these·and other brain models.

CHAPTER 2   Holography

2.1.   General Principle

Before embarking on a description of those aspects of holography which are considered to be of particular interest as models of memory, we outline the general physical set-up. Holography was invented by Gabor[1,2] as a method of getting round the limits of resolution of electron microscopes of that time. Gabor's proposal was to expose a photographic plate to the diffraction pattern produced by a small object in the coherent illumination of the electron lens system of an electron microscope. The developed plate is then put into a scaled up system of coherent optical illumination in such a way that the ratio of the linear dimensions of the two systems is the ratio of the wavelengths of the two sources. An enlarged image of the original object, which may be three dimensional, occurs in the optical system and this may be viewed or photographed by normal optical means.

According to Gabor, the principle underlying the reconstruction is the following. Suppose that two coherent monochromatic sources of the same frequency illuminate a photographic plate. Then the wave at that plate may be described by the functions:

$$A_0 \, e^{i(\phi_0 + \frac{2\pi ct}{\lambda})} \quad \text{and} \quad A_1 \, e^{i(\phi_1 + \frac{2\pi ct}{\lambda})}$$

Where A_0, A_1, ϕ_0, ϕ_1, are real numbers and functions only of position on the plate. The intensity of the superposition of these two waves is then given by

$$(A_0 e^{i\phi_0} + A_1 e^{i\phi_1})^{\frac{1}{2}} (A_0 e^{-i\phi_0} + A_1 e^{-i\phi_1})^{\frac{1}{2}}$$

$$= (A_0^2 + A_1^2 + 2A_0 A_1 \cos(\phi_i - \phi_1))^{\frac{1}{2}}$$

The plate can then be developed in such a way that the photographic density is the square of this intensity, so that when this plate is illuminated in the same way but with just one of these sources $A_0 e^{i(\phi_0 + \frac{2\pi i ct}{\lambda})}$ say, we get a transmitted wave:

$$e^{\frac{2\pi i ct}{\lambda}} (A_0^3 e^{i\phi_0} + A_1^2 A_0 e^{i\phi_0} + A_0^2 A_1 e^{i\phi_1} + A_0^2 A_1 e^{i(2\phi_0 - \phi_1)}) \qquad 2.1$$

If A_0 is constant over the plate, one part of this expression will be $A_0^2 A_1 e^{i(\phi_1 + \frac{2\pi ct}{\lambda})}$ which is the second wave multiplied by the constant A_0^2. There are clearly other transmitted waves and it is only in certain circumstances that these other waves can be separated or neglected as small in comparison with the reconstruction of the second wave.

The most straightforward physical device to demonstrate this property is illustrated in figure 1. This has been used by Leith and Upatnieks[3]. A laser produces a coherent beam of illumination which is reflected from both an object and a mirror to a photographic plate. The plate is exposed and developed and the object removed. On replacing the plate exactly in its original position a virtual image of the object appears at the position of the object. In addition a real image is produced at the reflected position of the object in the photo-

graphic plate. There is also a plane wave component from the mirror. These are three of the four terms in the expression 2.1. and they are easily separated in the physical system.

Figure 1

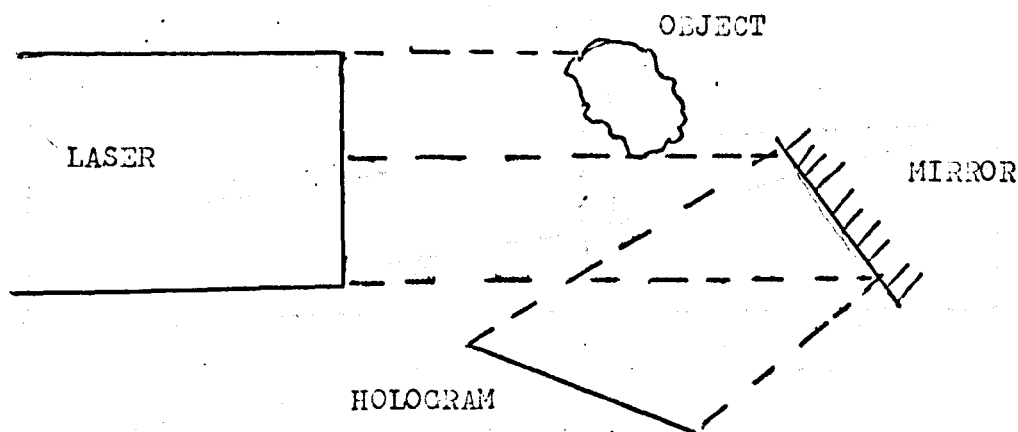Pribram[11] has proposed this device as a model of memory, among his reasons are first that the hologram is a 'non local' store, there is not a point to point correspondence between the plate and the object (as there would be in normal photography) and second that a fragment of the plate can be used to reconstruct the whole object, though with some blurring. But we will examine this analogy in detail later.

2.2. Fourier Transform Holography

In Fourier Transform Holography the hologram plate records the power spectrum of a photographic transparency. The system shown in figure 2, which was proposed by Van Heerden[4] does just this operation.
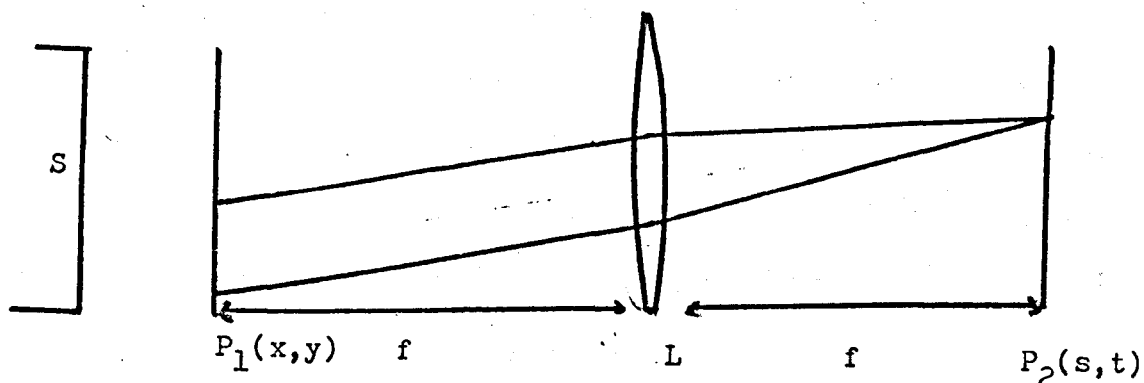
<u>Fig. 2</u>

A source of coherent light illuminates a photographic transparency at P_1 which is in the focal plane of the lens L. P_2 is a photographic plate in the other focal plane. If the lens is reasonably large then the amplitude of the wave arriving at P_2 is proportional to

$$A(s,t) = \frac{1}{\sqrt{2\pi}} \iint_{P_1} e^{2\pi i(sx + ty)} f(x,y)\ dx\ dy$$

where (x,y) and (s,t) are coordinates for P_1 and P_2 measured from the optic axis respectively and $f(x,y)$ describes the density of the transparency at P_1. If we suppose that $f(x,y)$ is zero off P_1 then $A(s,t)$ - where it is defined, is the two dimensional fourier transform of f.

To understand the usefulness of this, we first list some of the elementary properties of fourier transforms. · The fourier transform mf of a complex valued function f of the real time is defined by

$$mf(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{2\pi isx} f(x) \, dx$$

mf is only well defined for certain functions f, but the operation can be extended to take account of generalised functions so that, for example, a constant function has a well defined fourier transform, namely a delta function [5] . The properties of fourier transforms that we shall need here are:

1) $m^2 f = m\bar{f}$ where $\bar{f}(x) = f(-x)$

2) $mf^* = (\overline{mf})$

3) $m(fg) = (mf) \underset{*}{*} (mg)$ where $f * g$ is the convolution
    of f and g i.e. $\int_{-\infty}^{+\infty} f(t)g(x-t)dt$

We also define the correlation of f and g to be the function:

$$(f \circledast g)(x) = \int_{-\infty}^{+\infty} f(t)g(t-x)dt = f * \bar{g}.$$

And while convolution is commutative and associative, for correlation:

$$f \circledast g = \bar{g} \circledast \bar{f}$$

$$(f \circledast g) \circledast h = f \circledast (g \circledast \bar{h})$$

So that $(mf)(mf)^* = m(f \underset{*}{*} \bar{f}) = m(f \circledast f)$

This is the power spectrum of f which is also the fourier transform of the autocorrelation of f. Finally, if δ is

the generalised function with the property that

$$\int_{-\infty}^{+\infty} \delta f = f(0) \text{ then } f *_{} \delta = f \text{ and } \delta \circledast f = f$$

These remarks apply equally to complex valued functions of R^n. Now if we expose the plate in Van Heerden's apparatus and develop it appropriately the density of this plate will be proportional to the square intensity so that we shall have recorded (mf) (mf)*. The exposed plate can now be replaced in its former position and in the position of the photographic transparency at P_1 we place a new transparency described by $g(x,y)$. Immediately in front of P_2 we will then get a wave described by mg and immediately after it a wave described by

$$(mg) \ (mf) \ (mg)*$$

Now by means of a second lens (see figure 3), we can perform the same optical operation to arrive at the fourier transform of this expression:

$$\bar{g} *_{} \hat{f} \quad f = \bar{g} *_{} (f \circledast f) \qquad\qquad 2.2.$$
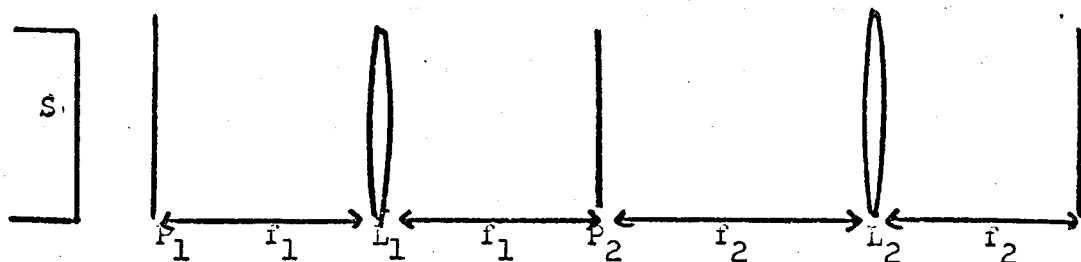

Fig. 3

Now $f \circledast f$ reaches a maximum at 0 and if f is suitably noisy or random (in a sense we shall examine shortly) we may approximate it by a delta function at 0 so that the reconstructed image at P_3 is:

$$\bar{g} * (f \circledast f) \sim \bar{g}$$

so that we recover, first of all an inverted image of g. It may happen that, if g is a fragment of f then $g \circledast f$ also resembles a delta function so that the recovered image at P_3 can also be approximated:

$$(\bar{g} \circledast \bar{f}) * \bar{f} \sim \bar{f}$$

The latter is the 'ghost image' of f, that is to say we have recovered the whole of the original function f by putting into the system a fragment g of that function. Of course, these two approximations for the recovered image will only be consistent when $f = g$, in practice the image of g will be stronger than the image of f to an extent which depends on the size of the fragment g. Moreover - had the fragment g been displaced from f by a, a two-vector, then $g \circledast f$ would be a delta function displaced by a so that the ghost image of f would be correspondingly displaced, and in the output, the ghost image of f will match with the strong reconstruction of g.

2.3. Applications

By taking various forms of f we can put the fourier transform holograph to a variety of uses.   In particular if $f = f_1 + f_2$ where f_1 and f_2 are non zero in adjacent regions then, having recorded the hologram of f we may use either of f_1 or f_2 to evoke an output of the other.   So that putting in f_1 will produce an output of f_1 together with a ghost image of $f_1 + f_2$.

A special case of this is where f_2 consists of a single bright spot, at the origin say, so the function whose hologram we construct is of the form $f_1 + \delta$, where δ is the delta function representing this bright spot.   On inputting g to the system we reconstruct:

$$\bar{g} \ast (\bar{f}_1 + \delta) \circledast (f_1 \circledast \delta)$$
$$= g \ast (\bar{f}_1 \circledast f_1) + \bar{g} \ast \bar{f}_1 + \bar{g} \ast \bar{f}_1 + \bar{g} \qquad 2.3$$

There are two choices we could make for g, the first is a bright spot, again suppose it is at the origin.   Then the reconstruction of the image at P_3 is

$$f_1 \circledast f_1 + f_1 + \bar{f}_1 + \delta \qquad\qquad 2.4$$

Suppose f_1 had been chosen so that it occupied a limited region of space whose greatest distance was less than the distance of this region from the bright spot (figure 4).   Then of the

terms of 2.3 $f_1 \circledast f_1$ and δ would occupy a central region

and the terms f_1 and $\overline{f_1}$ would be non zero in diametrically

opposite regions which are both adjacent from the central

region.

reconstruction
of $\overline{f_1}$

b>a

bright spot

a

←extent of
f_1
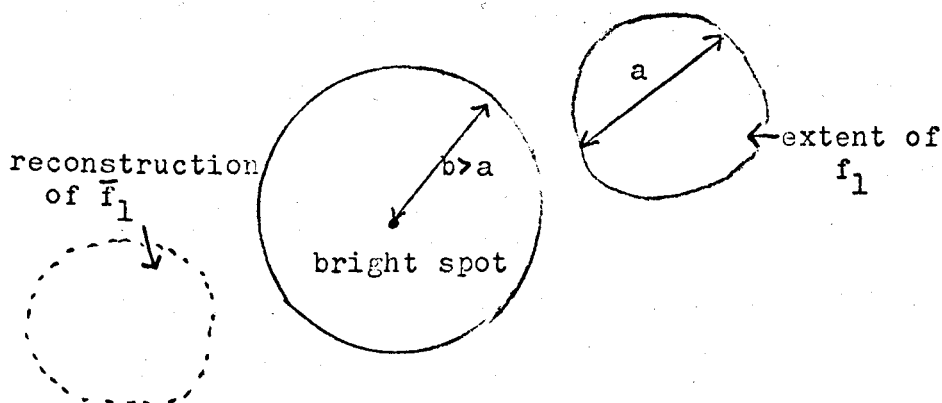
Fig. 4.

In consequence, we would get <u>accurate</u> recall of f_1 and this is

the case that most closely relates to the apparatus shown in

figure 1.

The second choice we could make for g is to make it a

fragment of f_1 so that from 1.4 the recalled pattern has now a term

$\overline{g} \underset{*}{} f_1$ which is $(\overline{g \circledast f_1})$ so that we create a bright spot in the

(inverted) image at the origin. Moreover if this fragment is

displaced the bright spot will be displaced by a corresponding

amount, so that the fragment may be located in the original.

The latter is Van Heerden's suggestion for locating a fragment

of a page in a stored library.

The final application of this system is to make a 'multiple hologram'. We expose the hologram plate to a number of separate transparencies and record

$(mf_1)(mf_1)* + (mf_2)(mf_2)* +....+(mf_k)(mf_k)*$

If we now input g_i, a fragment of f_i one of the output terms is

$$(\bar{g}_i \circledast \bar{f}_i) * \bar{f}_i \sim \bar{f}_i$$

Hopefully, there will not be much correlation between g_i and any of the other stored f_i, so that the reconstruction of $\bar{f}_i$ will predominate though we should expect there to be an increase in the noise level.

As to the noise, suppose now that in a one dimensional case, the signal f is repeated by a discrete set of values $(\phi_1 \phi_2 .... \phi_n)$ and that g is represented by $(\gamma_1 ....\gamma_m)$. Then we may express the recall by

$$P_k = \sum_{ij} \gamma_{-i} \phi_{-j+i} \phi_{k-j}$$
$$= \sum_{ij} \gamma_i \phi_{j-i} \phi_{k+j}$$

where $\phi < i \leq m$, $0 < j-i \leq n$ $0 < k+j \leq n$
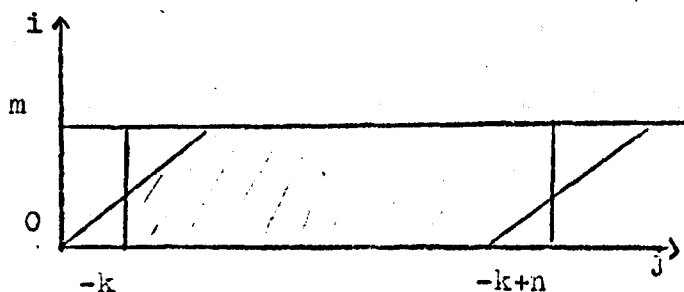
The number of terms which contribute to each term ρ_k is represented by the shaded area of the diagram above. We now make the assumption that the $\dot{\phi}_i$ are independent with a gaussian distribution, each with a variance α. We also assume that g is a fragment of f so that $Y_i = \dot{\phi}_i$ for i=1...m Then, in general the variance of ρ_k will be $\alpha^3 \sqrt{N_k}$ where N_k is the number of terms in the expression for ρ_k or the area of the shaded region above. Those terms which contribute to the ghost image ρ'_k are given by

$$\rho'_k = \sum_{i=k+j} Y_i \ \phi_{j-i} \ \phi_{k+j}$$

$$= \sum_j Y_{k+j} \ \dot{\phi}_{k+j} \ \dot{\phi}_{-k} \qquad\qquad 2.5$$

whose variance will be $m \alpha^3$ since there are m non-zero y terms. So for ρ_k $1 \leq k \leq n$ the signal to noise ratio is $\sqrt{\dfrac{m}{N_k}}$ Within these limits N_k varies between m^2 and $nm-m^2$ so that the signal to noise ratio is never worse than $\sqrt{\left(\dfrac{m}{n-1}\right)}$ nor better than 1. This is not very satisfactory, unless there is some redundancy in the signals we cannot expect good recall of a ghost image from a small fragment.

The expression 2.5 above for ρ'_k gives us a clue for a system which is fundamentally simpler than holography and

possesses, or can be modified to possess all the desirable
aspects of holography.   These are associative nets to which
we examine in some detail.

CHAPTER 3   Associative Nets

3.1.   Linear Associative Nets

In this chapter we shall discuss a class of models for memory which are conceptually simpler than holographic models yet retain many of the properties which make holography an interesting model. The likelihood that these associative nets are realised in the nervous system will be discussed in detail later, but it is worth pointing out now that such nets require only those properties of cells and synapses·which are almost universally acknowledged to exist in the nervous system. One property of fourier transform holography will be lost - the ability of such holographs to produce a displaced output from a correspondingly displaced input. Later we shall see how to recover this property and gain others by modifying these nets.

Rather than store continuous functions we shall think of the signals to be stored and recorded as vectors. If we have a set of n-vectors $\{\underline{x}^i\}$ where $\underline{x}^i = (x_1^i, \ldots x_n^i)$ i=1, 2, ...k then we may store the numbers $W_{pq} = \sum_{i=1}^{k} x_p^i \; x_q^i$ and on inputting a vector $\underline{y} = (y, \ldots y_n)$ we may recover by matrix multiplication a signal

$$\sum_{p=1}^{n} y_p \sum_{i=1}^{k} x_p^i \; x_q^i \quad \text{or}$$

$$\sum_{i=1}^{k} (\underline{y} \cdot \underline{x}^i) \; \underline{x}^i$$

Now as in holography if $\underline{y}$ is a suitable fragment of $\underline{x}^i$ then the term $(\underline{y}\cdot\underline{x}^i)\underline{x}^i$ will predominate and we again recover a ghost image of $\underline{x}^i$. Figure 1 shows a system which will do just this operation.
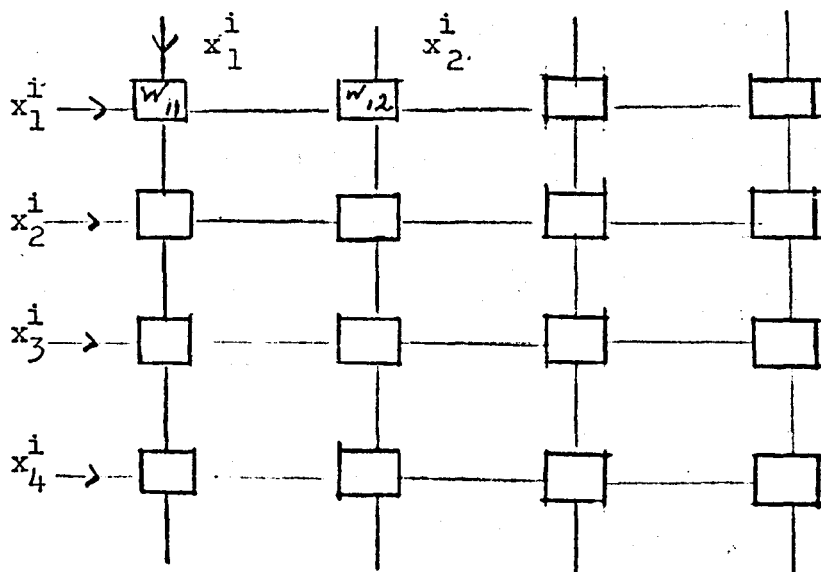
Figure 1

Each box in Figure 1 represents a real number or "weight". When a vector $\underline{x}^i$ is stored, the weight at the pth row and qth column is increased by $x_p x_q$. To recall we input a vector $\underline{y}$ along the rows and form the weighted sum of its components in each column - these constitute the outputs. Such a system is linear and we note that if we have only <u>one</u> stored vector $\underline{x}$ the output from any input vector which is not orthogonal to $\underline{x}$ will be a multiple, the recovery of a 'ghost' image in this case is not at all remarkable, but unlike the hologram it is noise free.

In the same way that an estimate for the signal to noise ratio was found for the holograph we can estimate the signal to noise ratio in this linear associative net. Suppose there are k stored n-vectors and suppose $\underline{y}$ is a fragment of one of them in that it matches $\underline{x}^i$ say, in m places and is O elsewhere. Then, as before if the components of the stored vectors are taken independently from a gaussian distribution of variance α, the variance of the 'signal' term $(\underline{y}.\underline{x}^i)\underline{x}^i$ will be $m\alpha^3$ while the noise term will have variance $(k-1)n\alpha^3$ and the signal to noise ratio will be $\sqrt{\frac{m}{(k-1)n}}$ which is of course infinite when k=1.

There is nothing to stop us storing pairs of signals in an associative net of this kind, for example, let $(\underline{x}^1,\underline{y}^1)\ (\underline{x}^2,\underline{y}^2)\ \ldots\ (\underline{x}^k,\underline{y}^k)$ be _pairs_ of n vectors. The weight (p,q) is now $\sum_{i=1}^{k} x^i_p\, y^i_q$ and inputting $\underline{z}$, a fragment of $\underline{x}^i$ we get $\sum_{i}(\underline{z}.\underline{x}^i)\underline{y}^i$ which should result in an output of $\underline{y}^i$, such a system then 'associates' the pairs $\underline{x}^i$ and $\underline{y}^i$. The recall of $\underline{y}^i$ is again subject to exactly the same noise as the case discussed previously. It is also worth noting that if the system is suitably constructed we may use it reversibly to associate $\underline{y}$ patterns with $\underline{x}$ patterns, or their fragments. This is not a very plausible biological state of affairs.

The Linear Associative Net with Thresholds

It is possible in part to get round the noise problem in an associative net by endowing the signals with some sort of redundancy and using this redundancy to find the signal in the noise. As an example we consider the case in which each x_p^i and y_q^i is independently chosen to be +1 or -1 with probability $\frac{1}{2}$. The output from such a net will not, in general consist of 1's and -1's so we call the output on one line +1 if the output is in fact positive and -1 otherwise. This could be readily realised by some threshold device on each output line.

We now determine the probability of error in such a system. Suppose that $\underline{z}$ is a fragment of $\underline{x}^1$ and agrees with it in exactly m places (it is still ± 1 on the other places) so that $\underline{z}.\underline{x}^1 = 2m-n$. This will be the strength of the signal on each output line. If the noise on that line i.e. $\sum_{i=2}^{k} (\underline{z},\underline{x}^i)\, y_q^i$ is less than this number we get no error - if it is greater we get an error with probability $\frac{1}{2}$.

The probability that $(\underline{z}.\underline{x}^i)$ takes on a given value s is

$$P(\underline{z}.\underline{x}^i = s) = \frac{1}{2^n} \binom{n}{\frac{n+s}{2}}$$

where the binomial coefficient is taken to be 0 if n+s is odd. The probability that $\sum_{i=2}^{k} (\underline{z}.\underline{x}^i)\, y_q^i$ takes on the value s, given $\underline{y}$ is:

$$\sum_{y_q^2 s_2 + \ldots y_q^k s_k = s} \left(\prod_{i=2}^{k} \left(\frac{n}{\frac{n+s_i}{2}} \right) \cdot \frac{1}{2^n} \right)$$

by examining coefficients in $(x+\frac{1}{x})^{n(k-1)}$ this is seen to be

$$\left(\begin{array}{c} n(k-1) \\ \frac{n(k-1)+s}{2} \end{array} \right) \cdot \frac{1}{2^{n(k-1)}}$$

and this expression is, by the normal approximation

$$\frac{1}{\pi \sqrt{n(k-1)}} \, \ell^{-\frac{s^2}{4\pi n(k-1)}}$$

and the probability of error then becomes

$$\frac{1}{4\pi\sqrt{n(k-1)}} \int_{t > s} e^{-\frac{t^2}{4\pi n(k-1)}}$$

This can be evaluated numerically, for example, to get an error rate of less than 1% we require that $t^2 > 10 \, n(k-1)$ approximately. Certainly if k is of order n and t=n i.e. we use the whole of an $\underline{x}^i$ as an input, the recall should become better as n increases. While this is not totally satisfactory as a store it is very much better than a linear associative net or the fourier transform holograph. It can also be shown that choosing equal numbers (on average) of +1's and -1's for the components of each vector is a "best" case: we might, for example have chosen the vectors from a different essemble in the following way.

Suppose the vectors ($\underline{x}^i$, $\underline{y}^i$ and $\underline{z}$) are chosen to have components $\frac{+1}{2p}$ with probability p and $\frac{-1}{2q}$ with probability q. Then we note that $\underline{x}^i . \underline{x}^i = n$ and the expectation of $\underline{x}^i . \underline{z}$ is 0. The variance however is

$$\frac{1}{16}\left(\frac{p^2}{p^4} + \frac{q^2}{q^4} + \frac{2p^q}{p^2 q^2}\right) = \frac{1}{16}\left(\frac{1}{p} + \frac{1}{q}\right)^2$$

which is minimum when $p=q=\frac{1}{2}$.

By the Central Limit Theorem, the probability that $\sum_{i=2}^{k} \underline{z}.\underline{x}^i > s$ will, for large n and k be least when $p=q=\frac{1}{2}$ - the case we have already considered.

The Binary Associative Net

In a recent paper by Willahaw, Buneman and Longuet-Higgins[6] a rather different type of associative net was investigated. In this the input and output vectors consisted of ones and zeros and the weights, unlike the linear associative nets were restricted to being 1 or 0. The input vectors $\underline{x}^i$ now with N_A components are randomly selected from a set of such vectors, each with exactly M_A ones and $N_A - M_A$ zeros. The output vectors $\underline{y}^i$ with N_B components were similarly chosen from a set with exactly M_B zeros.

The weight w_{pq} is set to one if $x_p^i \, y_q^i = 1$ for at least one i. We can think of these weights as switches, any switch is turned on by simultaneously seeing a one in its input and output lines and is never turned off. If we have stored R pairs of vectors and we then input a vector $\underline{x}^1$ say, we get an output

$$\sum_{i=1}^{R} (\underline{x}^1 \cdot \underline{x}^i) \, \underline{y}^i = M_A \underline{y}^1 + \sum_{i=2}^{R} (\underline{x}^1 \cdot \underline{x}^i) \underline{y}^i$$

So that if we put a threshold of just under M_A, we will get a vector with ones whenever $\underline{y}^1$ has a one and possibly other ones as well which we regard as spurious.

If $M_A \ll N_A$ and $M_B \ll N_B$ we can calculate the probability that any line gives out a spurious one. The probability p_c that any weight is one is (since the patterns are chosen at random) given by

$$1 - p_c = \left(1 - \frac{M_A M_B}{N_A N_B} \right)^R$$

and since $M_A M_B$ is very small compared with $N_A N_B$ we may write

$$R M_A M_B = N_A N_B \, \log_e(1 - p_c)$$

The probability that one line spuriously exceeds threshold is $p_c^{M_A}$ and, if we demand that there be on average one spurious output for each output pattern:

$$\cdot (N_B - M_B) \, p_c^{M_A} = 1$$

which may again be approximated by

$$M_A \log_e p_c = -\log_e N_B$$

If we regard the information stored in such a network as the information necessary to store the R output patterns this is

$$I = R \log_2 \binom{N_B}{M_B} \text{ bits}$$

and this may be written, using the previous expressions as

$$I = N_A N_B \log_e 2 \log_2 p_c \log_2(1-p_c)$$

I then, reaches its maximum when $p_i = \frac{1}{2}$, that is when half the weights are one. In this case $I = N_A N_B \log_e 2$, so that the information density is about 0.69. Even if we were to use the $N_A N_B$ binary weights in some other manner, as in the core store of a computer, the information density could not be higher than 1. It is surprising that the information stored can be so high while we retain the association with arbitrary input patterns.

CHAPTER 4 Group Invariant Associative Nets

Although the associative nets of the previous chapter had many properties in common with the holograph, the ability to recognise displaced patterns was lost.   Even if such an ability has no direct relevance to human perception, it is possible that the ability to recognise some transformations of an input may account for the psychological phenomenon of generalisation.   For example, an object may be recognised by humans and some animals when it is viewed in some position in which it has never been seen before.   Some transformation of the input (even if it is not displacement) has taken place. In this chapter we shall show how an associative net may be modified so that when a transformation of its inputs takes place, we can ensure that a previously specified transformation of its outputs also takes place.   It will emerge later that from the theory of such nets that we can prove, as a special case, Minsky and Papert's group invariance theorem[7] for perceptrons.   Before dealing with the general theory, two examples are given, the first is an associative net which is functionally similar to a holograph in that it will give displaced outputs from displaced inputs.   The second will recognise one graph as a subgraph of another.   This is a problem which may bear upon the manipulation of language, but the discussion of this is reserved for the last chapter.

These two examples will be extensions of the binary
associative net, so that the connections will be referred to
as switches.   Later, in the general theory we shall take account
of the other types of associative net.

4.1. The Correlograph

In the same paper in which they put forward the binary
associative net, Willshaw, Buneman and Longuet-Higgins[6] also
proposed an optical device the 'correlograph' which would record
directly the cross correlation of two photographic transparencies.
Such a device could, though it was limited by diffraction, be
used to display all the properties of the holograph without the
use of coherent light.. In detail the correlograph functioned
as a binary associative net with its switches 'tied'.   This
means that whenever one switch in such a net is turned on (by
recording a pair of patterns) a whole subset of the switches is
also turned on.   Suppose that, in a binary associative net
with N input and N output lines we turn on the switch (i,j)
which connects the ith input line to the jth output line.   We
then turn on any other switch (p,q) if

$$p = i+r \bmod N$$

$$q = j+r \bmod N$$

for some value of r.   (We work modulo N to avoid boundary
conditions which would produce, in some cases, incomplete outputs).

As before, we store pairs of patterns in which the input patterns have M_A active lines and the threshold on the output lines is just under M_A. Suppose that such an input pattern is given by the active input lines $p(1)$, $p(2)$ ... $p(M_A)$ and this evokes an output of active lines $q(1)$, $q(2)$ ... $q(k)$. Then the input of the displaced pattern:

$$p(1)+r, \ p(2)+r, \ \ldots\ldots \ p(M_A)+r$$

evokes the correspondingly displaced output:

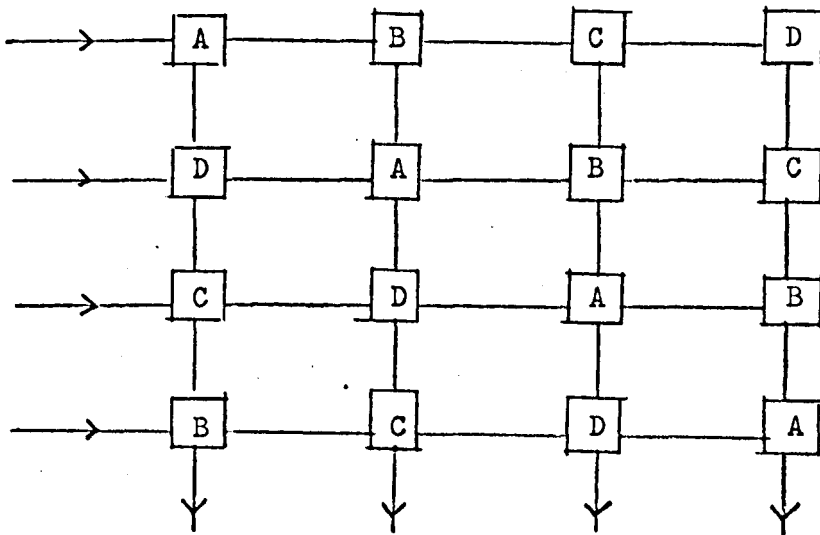$$q(1)+r, \ q(2)+r, \ \ldots\ldots \ q(k)+r$$

Fig. 4.1.   The sets of switches tied for a simple correlograph.

Tying switches in this way reduces by a factor $\frac{1}{N}$ the number of effective switches and consequently reduces by the same factor the amount of information that can be stored in such a net.   We shall show later that some improvement can be gained

by suitably coding the inputs so that more input lines and
more sets of tied switches are provided. But the net as it
stands can be used to store cross correlations or auto-
correlations and function in a similar fashion to a holograph.

4.2. The Graph Isomorphism Problem

The problem is as follows: given two graphs F and G,
how do we recognise that F is a part of G and, given that F
is a part of G, what is the function that carries the nodes of
F to the nodes of G if such a function is unique? Formally
let T_F and T_G be two sets (the sets of nodes). A graph F is
a relation T_F that is a subset R_F of $T_F \times T_F$; similarly a
relation R_G defines a graph G. We say there is a graph mono-
morphism from F to G if there is an injection T_F to T_G, ϕ say,
such that

$$(\phi \times \phi) \ (R_F) \subset R_G$$

F and G are isomorphic if there is a monomorphism F to G and
a monomorphism G to F. Clearly any graph is isomorphic to the
graph produced by permuting its nodes. What we shall do now
is to demonstrate the existence of associative nets with tied
switches which will, up to a point, recognise one graph as part
of another. It will be recalled that a tied associative net
can be made to function as a device which stores correlations and
that if the autocorrelation of a function or pattern f is stored,
then a sufficiently large fragment of f will produce a 'ghost

image' of f and this image will be displaced by an amount corresponding to the displacement of the fragment.   The following procedure for tying switches in an associative net, will in certain circumstances produce a 'ghost' graph with its nodes permuted to match with the input of the fragment.

We start by supposing that the given graph G is defined on a node set T_G containing N elements.   G is then specified by a subset of the N^2 pairs in $T_G \times T_G$.   Now let us suppose that these N^2 pairs are in fact the input lines to an associative net.   We have observed that permuting the nodes of a graph leaves us with an isomorphic graph so, for example the following figures will represent isomorphic graphs:

In fact the set of all such graphs will be determined by corresponding permutations of the rows and columns in such a figure.   If the inputs are ordered row by row 1 ... N^2 then the following permutations of these inputs will give isomorphic

inputs

$$pN+q - \rho(p).N + \rho(q)$$

where ρ is an element of the symmetric group S_N and p,q are chosen so that q < N. Let us call this permutation group Σ_N, it is isomorphic to S_N. As we did in the discrete correlograph, we tie the switches together in the sense that if we turn on (p,q) we also turn on the switch $(\sigma(p), \sigma(q))$ for all $\sigma \in \Sigma_N$. Figure 4 2.1 shows Σ and the switches on the $3^2 \times 3^2$ associative net that we would tie together in consequence. Each set of tied switches is denoted by the same letter.

A graph on three nodes X Y Z, say is represented by a choice of a subset of the numbers in the following diagram

	X	Y	Z
X	1	2	3
Y	4	5	6
Z	7	8	9

This would be rather a simple graph, but graphs of this size are adequate to illustrate the function of such associative nets.

$$
\begin{array}{ccc}
(1\ 2\ 3 & 4\ 5\ 6 & 7\ 8\ 9) \\
(5\ 6\ 4 & 8\ 7\ 9 & 2\ 3\ 1) \\
(9\ 7\ 8 & 3\ 1\ 2 & 6\ 4\ 5) \\
(5\ 4\ 6 & 2\ 1\ 3 & 8\ 7\ 9) \\
(9\ 8\ 7 & 6\ 5\ 4 & 3\ 2\ 1) \\
(1\ 3\ 2 & 7\ 9\ 8 & 4\ 6\ 5)
\end{array}
$$

	1	2	3	4	5	6	7	8	9
1	A	B	B	C	D	E	C	E	D
2	F	I	J	K	G	L	M	N	H
3	F	J	I	M	H	N	K	L	G
4	G	K	L	I	F	J	N	M	H
5	D	C	E	B	A	B	E	C	D
6	H	M	N	J	F	I	L	K	G
7	G	L	K	N	H	M	I	J	F
8	H	N	M	L	G	K	J	I	F
9	D	E	C	E	D	C	B	B	A

Figure 4.2   The permutation group $\sum$ and an associative net with its switches tied by $\sum$.

The graph:
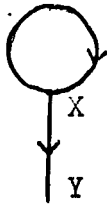
would then be represented by (X,X) and (X,Y) or 1 and 2 in
this diagram.   Now we could form the equivalent of an auto-
correlation function in this diagram;   that is to say we store
the pair both with signals in lines 1 and 2 and in consequence
turn on all the switches denoted by A B F I.   We could now put
into this net a part of this graph, not necessarily on the same
nodes;   for example we could put in the graph:

which is represented by 7 in the diagram and, if the threshold
of the output lines is set at just under 1 we get an output on
lines 7 and 9, that is the graph:


We not only get out the little one-link graph we put in but
also a 'ghost image' of the original!   This ghost image is
again suitably altered to match with the input.   Before embarking
on extensions of this idea it is as well to examine the limitations
of such a device in this form.

Our first observation is that if we put in a graph which is ambiguously monomorphic with the stored graph then we cannot expect a unique representation of the stored graph in the output. This corresponds to correlating a fraction of a picture with the autocorrelation of that picture when the fraction is an ambiguous part of the whole, we would expect an output of more than one 'ghost image'. This ambiguity may take on rather more serious forms; in the example above inputting the simple graph,

produces the output:
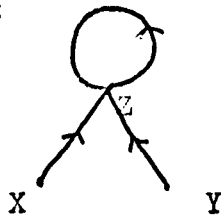
which is the superposition of the two possible continuations of this simple graph which match the stored graph. To get round this in any practical application of this device, we would have to have some a priori knowledge of how we wanted the graph continued and then to limit the number of output lines in an appropriate fashion.

A more serious limitation occurs because the number of groups of tied switches in such a device does not continue to grow as the size of the net increases. In fact if N (the number of nodes) is four or more then the number of such groups is 15.

Compare this with the correlograph where the number of these
groups is n for an n × n net.

To see this*, we can label each switch in the net by
(i,j,k,l) where i,j,k,l run from 1 to N.   (i,j) defines the
node pair which gives an input line and (k,l) similarly defines
an output line.   Two switches (i,j,k,l) and (i',j',k',l') are
tied if and only if there is a permutation of S_N which carries
i to i', j to j', k to k', and l to l'.   For example if
i,j,k,l are all different and so are i',j',k',l' if N 4
then we can form such a permutation, so that such switches are
tied.   In general there will be a partition of these four elements
such that each member of the partition contains equal elements
and any elements from two distinct members are unequal.   Again,
if N 4 then any other (i',j',k',l') with a similar partition
will represent a switch tied to the first.   It is easy to see
that there are only 15 such partitions.

We have already remarked that the capacity of such a net
as an information store is limited by the number of distinct
(untied) switches so that this upper limit would prove a serious
drawback if this device were used to store much information.
However, there is a method of increasing this capacity by the
use of masks.   A mask can be envisaged as a new input line

---

* I am very grateful to Mr. Stephen Isard for pointing out
  this solution.

which is active only when a certain subset of the old input
lines are active, in the example just given the pair of input
lines (1,7) would define a mask.   The order of a mask is the
order of the subset which defines it.   If we were prepared to
add to the $N^2 \times N^2$ associative net all the $\frac{1}{2}N^2(N^2-1)$ masks of
order 2 then we have a further set of input lines and switches
and we can again give a procedure for tying together the new
switches (we will do this in the next section).   It can be
shown at some length that the number of groups of tied switches
on these masks is not greater than 114, equality being reached
when N=6.   For N=3 the number is 56.

4.3.   The Theory of G.I. Associative Nets

In this section we formalise what was done in the last
section for graph recognition.   To start with we shall
generalise the definition of an associative net so that the
cases we considered in Chapter 3 will be special cases of this.
Suppose, firstly that the inputs $\{a_i\}$ are not necessarily binary
but can assume any real value.   We also assume that the connections
may also be specified by real numbers W_{ji}, likewise the thresholds
θ_j on the output lines.   We adopt Minsky and Papert's notation[7]
$\lceil P \rceil$ which has the value   1 when the predicate P is true and 0
if it is false.   (For example $\lceil 3 > 2 \rceil = 1$ $\lceil P \& Q \rceil = \lceil P \rceil \lceil Q \rceil$)

$$b_j = \left\lceil \sum_{i=1}^{n} w_{ji} \, a_i > \theta_j \right\rceil \qquad \qquad 4.3.1.$$

then defines the jth output, b_j.

We now assume that there is a group action which permutes
both the input and the output lines.   Let G be a group and
let ρ and ρ' be homomorphisms from G to the permutation groups
S_n and S_m on the input lines and output lines respectively.   For
example in the correlograph n=m and G was the cyclic group of
order n;   in the case of the graph recogniser, G was S_N and this
was mapped onto the group $\sum_N$ on both the input and output lines.
In general, however we need not assume that ρ G and ρ'G are
isomorphic.   Let $g \in G'$, in this section, by an abuse of
notation $a_{\rho g(i)}$ will be written $a_{g(i)}$ and $b_{\rho' g(j)}$ will be
written $b_{g(j)}$.   This should not produce any confusion especially
since we shall reserve the suffices i,n for the input and j,m
for the output.   Our first result is that if certain of the
weights and thresholds are equal then the action of G on an input
pattern will result in the corresponding action of G on an output
pattern.   Let $\underline{a}$ be the vector $(a_1, a_2 \ldots a_n)$ and let $\underline{a}_g$ be the
vector $(a_{g(1)}, \ldots a_{g(n)})$ define $\underline{b}$ and $\underline{b}_g$ similarly, finally let
$\tilde{\psi}$ be the function defined by 4.3.1. so that $\underline{b} = \tilde{\psi}(\underline{a})$.

Prop. 4.3.2.

If for each g in G, $w_{ji} = w_{g(j)\ g(i)}$ and $\Theta_j = \Theta_{g(j)}$

then if $\underline{b} = \underline{\Psi}(\underline{a}), \underline{b}_g = \underline{\Psi}(\underline{a}_g)$

Proof:
$$b_{g(j)} = \left\lceil \sum_{i=1}^{n} w_{g(j)i}\ a_i > \Theta_{g(j)} \right\rceil$$

$$= \left\lceil \sum_{i=1}^{n} w_{g(j)\ g(g^{-1}(i)}\ a_i > \Theta_{g(j)} \right\rceil$$

$$= \left\lceil \sum_{i=1}^{n} w_{j\ g^{-1}(i)}\ a_i > \Theta; \right\rceil \text{, by the hypotheses}$$

$$= \left\lceil \sum_{i=1}^{n} w_{ji}\ a_{g(i)} > \Theta; \right\rceil \quad \text{since the sum proceeds over all i}$$

which is the jth component of $\underline{\Psi}(\underline{a}_g)$

This shows that we get, as we expected from the examples, from inputs altered by an element of G correspondingly altered outputs. However, we would also expect that if ρ were not an endomorphism then it is possible that we only get this group invariant property because the outputs are rather simple. That is to say that several output lines will always produce the same signal. The following corollary makes this precise:

Let H be the group $\rho^{-1}(1)$ that is, the null space of ρ. Then $b_{g(j)} = b_j$ for each g in H.

Since g is in $\rho^{-1}(1)$ $\underline{a}_g = \underline{a}$ so that

$$\underline{b} = \underline{\Psi}(\underline{a}) = \underline{\Psi}(\underline{a}_g) = \underline{b}_g$$

In order to avoid this situation of the output patterns being redundant we need to be able to identify G with a group action on the input lines. However we shall need the more general result of when we deal with the possibility of encoding the inputs. Before this, there is an interesting converse to the last proposition.   It is clear that tying the connections of an associative net will reduce the amount we can store in it. Is it possible that there are nets which exhibit this group invariant property but whose performance would be altered by tying their switches?   The answer is no:

Prop. 4.3.3

Let $\{\underline{a}^t, \underline{b}^t\}$ be a set of input-output pairs of an associative net defined by 4.3.1. so that $\underline{b}^t = \underline{\Psi}(\underline{a}^t)$ for each t.   If also there is a group G and homomorphisms ρ and ρ' as before such that $\underline{\Psi}_g(\underline{a}^t_g) = \underline{b}^t_g$ for each t and for each $g \in G$.

Then there is an associative net with weights $w'_{ij} = w'_{g(i)\ g(j)}$ and $\Theta'_i = \Theta'_{g(j)}$ for all $g \in G$;   and which defines a function $\underline{\Psi}'$ such that $\underline{\Psi}(\underline{a}^t_g) = \underline{\Psi}'(\underline{a}^t_g)$ for each t and each $g \in G$.

<u>Proof</u>   First of all we observe that if

$$\lceil x > y \rceil = \lceil f_1(x) > h_1(y) \rceil = \lceil f_2(x) > h_2(y) \rceil \ldots$$

for some functions $f_1 \; f_2 \ldots h_1 \; h_2 \ldots$ then :

$$\lceil x > y \rceil = \lceil \sum_i f_i(x) > \sum_i h_i(x) \rceil$$

Now by our assumption that $\bar{\psi}$ is invariant under G, for each

t we have

$$b^t_{g^{-1}(j)} = \lceil \sum_{i=1}^{n} w_{ji} \; a^t_{g^{-1}(i)} > \theta_j \rceil$$

$$= \lceil \sum_{i=1}^{n} w_{jg(i)} \; a^t_i > \theta_j \rceil$$

So that $b^t_j = b^t_{g^{-1}(j)}$

$$= \lceil \sum_{i=1}^{n} w_{g(j) \; g(i)} \; a^t_i > \theta_{g(j)} \rceil$$

But b^t_j is also $\lceil \sum_{i=1}^{n} w_{ji} \; a^t_i > \theta_{g(j)} \rceil$

Summing over all $g \in G$, we have by our first observation

that $b^t_j = \lceil \sum_{i=1}^{n} \left(\sum_{g \in G} w_{g(j) \; g(j)} \right) a^t_i > \sum \theta_{g(j)} \rceil$

So that we may set $w'_{ji} = \left(\sum_{g \in G} w_{g(j) \; g(i)} \right)$

and $\theta'_j = \sum_G \theta_{g(j)}$

So that θ'_j and w'_{ji} have the required property.

This result, then shows that if a net has a group invariant property, an associative net with tied connections can be built which has the same power in that it not only realises the invariance under the same group, but that the input output pairs will be the same.

Encoding

Since tying switches will always reduce the amount of information that can be stored in an associative net, it would be advantageous to increase the number of connections in the net, and possibly in consequence, the number of sets of tied connections. We can do this by a set of <u>encoding functions</u> Φ which are simply a set of real valued functions on the input space. The set of encoding functions will map the space of inputs into a new and generally higher dimensional space. The set Φ is said to be closed under G if for all $g \in G$ and $\phi \in \Phi$:

$$\phi^g(\underline{x}) = \phi(\underline{x}_g) \in \Phi$$

where we have used $\underline{x} = (x_1, x_2, \ldots x_1)$ to denote an element in the input space, and supposed that G is a permutation group on these inputs.

Now it is easy to see that if Φ is closed under G then there is a homomorphism from G to S_Φ, the permutation group on Φ. We are now in a position to let the values of the functions ϕ serve as the inputs to an associative net so that

$a_i = \phi_i$ for some ordering of the encoding functions.   We can further let $a_{g(i)} = \phi_i^g$ and tie the connections of this new associative net as before.   We now get the result that the input output pairs of this net with encoding functions will be invariant under G.   That is if:

$$b_j(\underline{x}) = \left\lceil \sum_{\phi} W_{ji} \, \phi_i(\underline{x}) > \theta_j \right\rceil$$

then

$$b_{g(j)} \, (\underline{x}_g) = b_j(\underline{x})$$

In a similar way the outputs of an associative net can also be encoded, but in order to store anything in such a net the encoding functions would have to be reversible.   Moreover it is useful only in certain cases to have any coding on the output wires.   An associative net whose connections are thresholds are described by arbitrary real numbers and whose encoding functions are masks can be thought of as a battery of two layer perceptrons.   It will be shown shortly that such a device can realise any map from the space of all binary inputs to the space of all binary outputs, so that any coding of the outputs is unnecessary.   In cases where the limitations of the net produce incorrect outputs it may be possible to correct for these by encoding functions which exploit some redundancy in the uncorrected outputs.

Returning to the graph recogniser, we can see that this

higher. Likewise with the correlograph, we can use as extra inputs masks of order two or more and retain the group invariant property of the network. As an example there are $\binom{n}{2}$ masks of order 2, and for the correlograph we tie them with the output lines under the group Z_n with the appropriate homomorphisms. Under Z_n the masks split into $\frac{n-1}{2}$ transitive subsets of n elements each if n is odd. If n is even there are $n/2 - 1$ transitive sets of n elements and one of n/2 elements. Thus in the first case we will have $\frac{n(n-1)}{2}$ groups of tied switches and in the second $n(n/2-1) + n/2 = \frac{n(n-1)}{2}$ groups. In general there will be $\binom{n}{r}$ masks of order r and these will give rise to $\binom{n}{r}$ groups of switches in the correlograph.

The most interesting theoretical aspect of group invariant associative nets is the close relation they bear to Minsky and Papert's two layer group invariant perceptrons. The result which gives, as a special case, their group invariance theorem is the following extension of 4.3.3. to the case where we have encoding functions.

Prop. 4.3.4.

Suppose an associative net has a set of encoding functions $\mathcal{J}$ closed under G a permutation group on the arguments of these functions, and that G acts on the output lines as before

If for a set of inputs $\underline{x}^1$, $\underline{x}^2$ ... $\underline{x}^n$

$$b_j^t (\underline{x}^t) = b_{g(j)}^t (\underline{x}_g^t) \text{ for all } g, t \text{ and } j$$

Then the same outputs $\{ \underline{b}_g^t \}$ can be realised by an associative net in which $W_{ji} = W_{g(j)\ g(i)}$ and $\Theta_j = \Theta_{g(j)}$.

The proof is immediate from 4.3.3.

4.4. Generalised Correlation

We have seen in the previous sections how associative nets can exhibit group invariant properties. But associative nets - unlike holographs - operate on discrete input spaces. We now turn to the problem of whether we can generalise the definition of correlation (for functions defined on the real line or R^n) so that we can get group invariant properties for groups other than the translation group. We shall also see that when we get such group invariant properties (they are only obtained for certain groups) we can sometimes obtain the equivalent of a fourier transform which, as before, maps the generalised correlation of two functions into their (pointwise) product. To begin with we recapitulate on the uses of correlation:

Suppose that f and g are two functions on the real line (we shall extend to higher dimensions later). Their cross correlation $f \circledast g$ was defined to be:

$$(f \circledast g)(t) = \int_{-\infty}^{+\infty} f(x) g(x - t) dx$$

Correlations may be used in the following ways if the functions used satisfy certain conditions. These statements are not mathematically precise; they need the statistical qualification

1. Recognition

If g resembles a fragment of f for example if

$$g(x) = f(x) \text{ on } a \leq x \leq b$$
$$0 \text{ elsewhere}$$

then $f \circledast g$ will have a peak at t=0 whose width and sharpness will depend on f and g. In general if f and g are 'noisy' and b-a is large, this peak will be sharp.

If g' resembles a displaced fragment of f, for example if $g'(x) = g(x+\tau)$ then

$$(f \circledast g')(t) = (f \circledast g)(t-\tau)$$

so that $f \circledast g'$ will have a displaced but otherwise similar peak at $t = \tau$. Thus we could use this peak to locate a fragment of a signal in that signal.

2. Recall

Consider $(g \circledast f) \oplus g = (g \circledast g) \oplus f$; if as before g is such as to make $g \circledast g$ a sharp peak at 0 then $(g \circledast g) \circledast f$ will approximate f. So that if we can store $f \circledast g$, a subsequent correlation of this with g will produce an output or 'recall' of f. Moreover if we had stored several pairs $f_1 g_1, f_2 g_2, \ldots$ as well as f g so that the contents of the store were $g \circledast f + g_1 \circledast f_1 + g_2 \circledast f_2 \ldots$, then on correlating this with g would give the output

$$(g \circledast g) \circledast f + (g \circledast g_1) \circledast f + (g \circledast g_2) \circledast f + \ldots$$

Then, provided g does not 'resemble' g_1 g_2 ... the output
should again approximate f.   As before, a displaced input of
g will produce a correspondingly displaced output of f.   Another
interesting feature of this type of storage is that if we put in
an inverted copy of f i.e. f(-x) we get an inverted output of g
so that the store can be used reversibly.

3.   Ghost Images

A related, but not identical situation to that in the previous
section is where we store the autocorrelation of a function say
f $\circledast$ f.   If we form the correlation of this with another function
g we expect to get from (f $\circledast$ f) $\circledast$ g a copy of g.   But since this
previous expression can also be written as (f $\circledast$ g) $\circledast$ f, if g, as in s
section 1 resembles a part of f, then we should also get a weaker
output of f.   This is the 'ghost image' of holography and is a
special case of the more general holographic reconstruction
which does not employ fourier transforms.   The same remarks
about storing several patterns and displacements apply.

It has already been said that these remarks lack any kind
of mathematical rigour and that the signal to noise ratio in
the output must be examined as before, before these claims can
be made with any statistical justification.   However our immediate
purpose is to examine the possibility of extending these remarks
to take account of other operations of the input space apart from
displacements.   As a start we will examine the possibility of

recognising one function as a dilated version of another.

Suppose that we have the problem of recognising a function on a line as a dilated (expanded or contracted) version of another. Let f be such a function, and let g be a dilation of it so that $f(x) = g(\lambda x)$. Now consider the function

$$\phi(t) = \int_{-\infty}^{\infty} f(x) \; g(tx) \; \frac{dx}{|x|}$$

it is clear that this integral will not exist unless we place some restriction on f and g such as being $O(x^a)$ for some positive a) in a neighbourhood of O. We can first show that ϕ reaches its maximum value at $t = \lambda$; for:

$$\phi(t) = -\tfrac{1}{2} \int_{-\infty}^{+\infty} (g(\lambda x) - g(tx))^2 \; \frac{dx}{|x|}$$
$$+\tfrac{1}{2} \int_{-\infty}^{+\infty} (g(\lambda x))^2 \; \frac{dx}{|x|} \; +\tfrac{1}{2} \int_{-\infty}^{+\infty} g((tx))^2 \; \frac{dx}{|x|} \qquad 4.4.1.$$

The second two terms are constant and the first term cannot be greater than O. Then ϕ must reach its maximum value at $t=O$ where the first term is zero. Moreover we can make the same remarks as before about storing and retrieving as before except that now, the retrieved patterns are dilated to an extent which corresponds to the dilations of the inputs.

We might now wonder if there is a generalisation of this procedure to other operations on the real line. The following

proposition shows when and when we cannot hope to extend it to other operations.

Let $\mathcal{F}$ be a set of functions defined on S and taking values in S, where S is a subset of the real line, which is a semigroup under composition i.e.

$\mu_1 \mu_2$ is a function in $\mathcal{F}$ whenever μ_1, μ_2 are in $\mathcal{F}$

Prop. 4.4.2.

If: 1) There is a non-trivial homomorphism E carrying the semigroup into the additive reals.

   2) There is a point a of S with the property that: for any x in S there is an f in F and $\mu(a) = x$

   3) $\mu_1(a) = \mu_2(a)$ implies $\mu_1 = \mu_2$.

Then we can find a function Y: $S \to R$ such that:

$$Y(x) = Y(\mu(x)) - E(\mu)$$

Proof: Let $Y(x) = E(\mu_1)$ where $\mu_1(a) = x$. μ_1 exists
       and is unique by 2 and 3.

       Now $Y(\mu(x)) = E(\nu)$ where $\mu(x) = \nu(a)$ so that
       $\nu(a) = \mu(\mu_1(a))$ and hence $\nu = \mu \mu_1$

       We can therefore write
       $$Y(\mu(x)) = E(\mu \cdot \mu_1) = E(\mu_1) + E(\mu)$$
       $$= Y(x) + E(\mu).$$

The importance of this is that we can find integrals which are invariant under transformations by members of S of the underlying space. Suppose that, in addition that the functions are differentiable and that the function Y turns out to be differentiable, then from the preceding proposition we find that.

$$\phi(x) = \mu'(x) \quad \phi(\mu(x))$$

where $\phi = Y'(x)$

so that $\int_S f(\mu(x)) \phi(x) \, dx = \int_S f(\mu(x) \quad \phi(\mu(x)) \cdot \mu(x) \, dx$
$$= \int_S f(x) \quad \phi(x) \, dx$$

Now in the example 4.4.1. we took $\mu_t(x)$ to be tx so that the conditions of the previous proposition are satisfied by taking a=1 and $E(\mu) = \log t$. We then find that $Y(x) = \log x$ and that $\phi = 1/|x|$ as we had beforehand. (Taking a different a would not have altered this latter result).

It is worth perhaps, examining other functions on the real line to see how this method applies in other cases for example:

μ	$E(\mu_t)$	ϕ
$x+t+tx$	$\log(1+t)$	$\dfrac{1}{1+x}$
x^t	$\log t$	$\dfrac{1}{x \log x}$
tx	$\log t$	$\dfrac{1}{x}$
$e^{(e^{(t \log(\log x))})}$	$\log t$	$\dfrac{1}{x . \log x . \log(\log x)}$

In all these it is easy to find an appropriate range for the integration, though perhaps, in this one dimensional case, the only interesting operators are translation, x+t and dilation tx.

It is possible to use these ideas to show that we can meaningly define a correlation function over the set of transformations. We define

$$(f \oplus_{\not{\mathcal{X}}} g)\ (t) = \int_s f(x)\ g(\mu_t(x))\ \phi(x)\ dx$$

It is then easy to show that if $g(\mu_{t_1}(x)) = f(x)$ for some t_1 then $(f \oplus_{\not{\mathcal{X}}} g)(t)$ reaches its maximum value where $x=t_1$, by using the same procedure that we used in 4.4.2.

We can now show that this generalisation of a correlation function is in some cases equivalent to the standard definition. We assume, in addition to the conditions of prop. 4.4.2. that E is an epimorphism from $\mathcal{X}$ to R. It may happen that the functions f, g which we wish to correlate can be written in the form

$$f(x) = F(Y(x)), g(x) = G(Y(x)) \hspace{2cm} 4.4.3.$$

then:

$$(f \oplus_{\mu} g)(t) = \int_s F(Y(x))\ G(Y(\mu_t(x)))\ \phi(x)\ dx$$

$$= \int_{Y(s)} F(Y(\ G(Y+E(\mu_t))\ dY$$

From the conditions of prop. 3.2., if E is an epimorphism $\mathcal{X} \to R$
then Y maps S onto R so that the range of this integral is R and
that it takes on a unique value for each value of $E(\mathcal{X})$.   So
that if we can effect the transformation 4.4.3. then we can, by
mapping R to $E(\mathcal{X})$ transform this generalised correlation into the
standard (additive) definition.

The point of 3.2. is that it enables us to construct, in
certain cases the correlation ov r a given group of transformations,
it is a sufficient, rather than a necessary set of conditions.
There are certainly generalisations of it; for example the second
condition, a sort of transitivity, can be replaced by a condition
which divides S into a set of transitive subsets.

4.4.2. will also generalise to higher dimensions: $\mathcal{X}$ is
now a semigroup of operations $S \subset R^n \to R^n$, E carries S into the
additive group R^n Y is then a function $S \to R$ and, if it is differ-
entiable is the Jacobian of Y.

We cannot, in general hope to define correlation functions
in this manner for non-commutative groups, for example the group
of all translations and dilations of the real line is not
commutative, an element of it can be written

$$\mu_{a,b}(x) = ax + b$$

it is easily seen that condition 3 of 4.4.2. is not satisfied
by such a set of functions and that the integral

$$\int_{-\infty}^{+\infty} f(ax + b) \, \phi(x) \, dx$$

for any choice of f for which the integral exists will only be

invariant if it is 0.   It is worth noting that if f and g

are functions which only take on the value 0 or 1 and if

$f(x) = g(ax + b)$ then the integral

$$\int_{-\infty}^{+\infty} f(x)\, g(sx + t)\, dx$$

takes on its maximum value where s=a, t=b because this

integral is always less than the integral of $f(x)$ or $f^2(x)$.

We now turn to the theory of two layer perceptrons and

demonstrate the relation between these and group invariant

associative nets.

CHAPTER 5   Perceptrons

In this chapter we shall examine the correspondence between
group invariant associative nets and Minsky and Papert's work on
perceptrons.[7]   Some of their main results will be proved here by
rather different methods, which seem more direct.   Their work
is based on the fact that any Boolean function of n Boolean
variables can be realised by a 'two layer perceptron' the first
layer consists of a set of masks whose outputs are weighted and
summed.   According as this sum is greater or less than a certain
threshold we get an output of 1 or 0.

To make this precise, an input $\underline{x}$ will be an n-tuple
$(x_1, x_2, \dots x_n)$ where each x_i is either 0 or 1.   A mask, as
it was defined previously, is a function of the form:

$$\phi_S(\underline{x}) = \prod_{i \in S} x_i$$

where S is a subset of the integers 1 ... n.   Also as before,
$\lceil P \rceil$ is defined to have the value 1 whenever P is true otherwise
it is 0.   A Boolean function ψ is linear in a set Φ of Boolean
functions if

$$\psi(\underline{x}) = \left\lceil \sum_{\phi \in \Phi} w_\phi \phi(\underline{x}) > \theta \right\rceil$$

for real numbers w_ϕ and θ, then:

Prop. 5.1.   Any Boolean function $\psi(\underline{x})$ is linear in the set of
all masks $\{\phi_S(\underline{x})\}$

Proof: by induction on n, the number of variables. If we let the empty set define the constant mask, ϕ_0 which always has the value 1, then the Boolean constants (functions of no variables) are given by $\lceil W_0 \phi_0 (\underline{\lambda}) > 0 \rceil$, if we take W_0 as +1 or -1. This is the inductive base. Now we assume that for n-1 variables: $(x_1, x_2, \ldots x_{n-1}) = \underline{x}'$, say

$$\psi'(\underline{x}') = \lceil \sum_{S'} W_{S'} \phi_{S'}(\underline{x}') > 0 \rceil$$

where the S' are subsets of the integers 1, 2 ... n-1.

If $x_n = 1$, we may write in consequence:

$$\psi(\underline{x}) = \lceil \sum_{S'} a_{S'} \phi_{S'}(\underline{x}') > 0 \rceil$$

and if $x_n = 0$

$$\psi(\underline{x}) = \lceil \sum_{S'} b_{S'} \phi_{S'}(\underline{x}') > 0 \rceil$$

this is equivalent to writing

$$\psi(x) = \lceil x_n \sum_{S'} a_S \phi_{S'}(\underline{x}') + (1-x_n) \sum_{S'} b_{S'} \phi_{S'}(\underline{x}') > 0 \rceil$$

now the function $x_n \phi_{S'}(\underline{x}')$ is simply the mask $\phi_S(\underline{x})$ where $S = S' \cup \{x_n\}$ so that the last expression, which is

$$\psi(\underline{x}) = \lceil \sum_{S'} [(a_{S'} - b_{S'}) x_n \phi_{S'}(\underline{x}) + b_{S'} \phi_{S'}(\underline{x}')] > 0 \rceil$$

is linear in a set of masks, and the induction is complete.

The parity predicate

As an example of putting predicates in this form, we examine the nature of the predicate that the number of ones in the input is an odd number, or:

$$\psi_P(\underline{x}) = \lceil \sum_{i=1}^{n} x_i \text{ is odd} \rceil$$

We define the order of a mask ϕ_S to be the order of (or the (finite) number of points in) the set S. The orders of the masks necessary to represent certain predicates are of theoretical interest, and certainly of practical interest if one believes that perceptrons are useful models of parallel processing. From this point of view ψ_P is a particularly 'bad' case, as the following results show.

<u>Prop. 5.2.</u> If ψ_P is linear in a set ϕ of masks then ϕ contains the mask of order n.

Proof: If n=1 then clearly ψ_P requires the mask of order 1. Let N be the first n for which ψ_P is linear in a set of masks of order less than N. Then

$$\psi_P(\underline{x}) = \lceil \sum_{S \in \phi} a_S \psi_S(\underline{x}) > 0 \rceil$$

where ϕ contains masks of order less than N. Now let ϕ_1 be the subset of ϕ whose masks contain x_N and let $\phi_2 = \phi - \phi_1$.

Then

$$\psi'_P = \left\lceil \sum_{S \epsilon \Phi_1} a_S \phi_S + \sum_{S \epsilon \Phi_2} a_S \phi_S > 0 \right\rceil$$

Now if $S \epsilon \Phi_1$, define S' as $S - \{x_n\}$ and set $a_{S'} = a_S$ also

let Φ'_1 be the set of all such sets S'. Φ'_1 then x contains

masks of order x less than N-1. So that

$$\psi'_P (\underline{x}) = \left\lceil x_N \sum_{S' \epsilon \Phi'_1} a_{S'} \phi_{S'} (\underline{x}) + \sum_{S \epsilon \Phi_2} a_S \phi_S(\underline{x}) > 0 \right\rceil$$

In this expression we can set $x_N=1$ or $x_N=0$. In the first

case we will get a Boolean function of the N-1 variables

$\underline{x}' = (x_1, x_2, \ldots x_{N-1})$ which must be the <u>even</u> predicate::

$$\left\lceil \sum_{i=1}^{N-1} x_i \text{ is even} \right\rceil$$

In the second case we must get the odd predicate ψ'_P on these

N-1 variables. So that by taking the negation of the case

$x_N=1$, we get two expressions for the odd predicate ψ'_P ,

namely:

$$\psi'_P (\underline{x}') = \left\lceil - \sum_{S' \epsilon \Phi_1} a_{S'} \phi_{S'} (\underline{x}') - \sum_{S \epsilon \Phi_2} a_S \phi_S(\underline{x}') \geq 0 \right\rceil$$

$$= \left\lceil \sum_{S \epsilon \Phi_2} a_S \phi_S (\underline{x}') > 0 \right\rceil$$

We can now 'sum' these predicates to get an equivalent

expression for ψ'_P:

$$\left\lceil - \sum_{S' \epsilon \Phi_1} a_{S'} \phi_{S'} > 0 \right\rceil .$$

But we have already noted that ϕ_1' contains masks of order at most N-2, so that we have found an expression for ψ_P' which is linear in masks of order N-2; and this contradicts our initial assumption.

From the proof of this proposition we can also infer that ψ_P must contain any mask that ψ_P' always contains. Continuing this argument, if ψ_P^A is the parity predicate on the variables $\{x_i : i \in A\}$ where A is any subset of the integers 1 ... n then ψ_P contains any mask that ψ_P^A must contain, in particular it must contain the mask ϕ_A (by the previous proposition. Hence:

Prop: 5.3. ψ_P contains all masks.

Using similar methods we can also find a set of weights for ψ_P:

Prop. 5.4. An expression for ψ_P is:
$$\psi_P = \left\lceil -\sum (-2)^{|S|} \phi_S > 0 \right\rceil$$

Proof

Define the function $P_N(x_1, x_2, \ldots x_N)$ as
$$-\sum (-2)^{|S|} \phi_S$$

where the sum extends over all subsets (including the empty set) of the numbers 1, 2 ... N. We shall prove inductively that

$$P_N = 1 \quad \text{if } \sum_1^N x_i \text{ is odd}$$

$$P_N = -1 \quad \text{if } \sum_1^N x_i \text{ is even}$$

Certainly P_1 has this property; assume that P_{N-1} also has it, then the expression

$$-x_N P_{N-1} + (1-x_N) P_{N-1}$$

also has it. But this can be seen to be P_N. The point of doing this is the number of masks and their coefficients constitute a worst possible case, i.e. there are predicates where masks of high order and large weights are needed. By examining the proof of Prop. 5.1. it can be seen that if the weights are integers, then any Boolean function can still be realised and that the magnitude of the weights need not be higher than 2^N.

Group invariant perceptrons

There is a very close relation between the two layer perceptrons that have just been discussed and the associative nets with encoding functions that we discussed in the last chapter. In fact, a two layer perceptron is such a net whose inputs are binary, whose encoding functions are masks and which has precisely one output line. Let G be a group of permutations of the integers 1 ... n and let S be a subset of those integers. We can define the $g(S)$ by $\{i: g(i) \in S\}$ so that for each mask ϕ_S we can define a mask $\phi_{g(S)}$. As before, a set Φ of masks is closed under G if $\phi_{g(S)} \in \Phi$ whenever $\phi_S \in \Phi$, so that we have a

homomorphism ρ from G to the permutation group on Φ.   Also
we let ρ' be the trivial homomorphism from G to the trivial
group of permutations of the one output line.   Then the two
results that we derive from 4.3 of the previous chapter are:

Prop. 5.5.   If Φ is closed under G and if, for all masks ϕ_S
and all $g \in G \, \alpha_S = \alpha_{g(S)}$ then the predicate

$$\psi(\underline{x}) = \left\lceil \sum_S \alpha_S \, \phi_S(\underline{x}) > 0 \right\rceil$$

is invariant under G;   that is to say that

$$\psi(\underline{x}_g) = \psi(\underline{x}) \text{ for all } g \in G.$$

Prop. 5.6.   If Φ is closed under G and if ψ has a representation:

$$\psi = \left\lceil \sum a_S \, \phi_S > 0 \right\rceil$$

and is invariant under G

then it also has a representation

$$\psi = \left\lceil \sum b_S \, \phi_S > 0 \right\rceil$$

where $b_S = b_{g(S)}$ for all g.

The latter, then is Minsky and Papert's Group invariance theorem.
They use it to obtain bounds on the sizes of perceptrons needed
to recognise certain, mostly geometrical, predicates.   In
addition they use it to show that ψ_p, the parity predicate,
requires all masks, which has been done here without it.   If
the input to the perceptron is thought of as a two dimensional
array or retina on which is presented a black and white figure,
those input lines which see black register 1 and the others

register O.   Connectivity for example cannot be recognised

by a perceptron whose order (the maximum order of masks) is

bounded independently of the size of the retina.   For an

examples of partly serial methods for topological decomposition,

some are given in Perceptrons, another is given by Buneman[8].

CHAPTER 6 Learning algorithms

6.1. Perceptron learning

The methods of weight adjustment in the associative nets
described earlier were straightforward; connections were
established, or modified as some function of the immediate
activity on either side of that connection. If one output
line of an associative net is considered, the connections are
set by a simultaneous presentation of the input and output in
that line. Suppose that no such information is given, but
merely an indication whether or not the current response is
correct. Some modifying instruction is given to that connection
which is not identical with the input to that connection or the
response in the output line. The most famous example of this
type of modification is Perceptron convergence* which was first
demonstrated by Rosenblatt[9] and two proofs are given of the
Perceptron convergence theorem by Nillson[10]. The proof given
here has a certain amount in common with Nillson's geometric
proof (though it does not involve geometry); its main virtue
is to show that there are a variety of feedback procedures which
work for fundamentally similar reasons.

---

* The word 'perceptron' occurs in a variety of literature
and assumes different meanings. In this chapter, when it is
used, its reference will be to a 'one layer' device, which is
like the second layer of the two layer Perceptrons of the
previous chapter.

Let S be a finite set of vectors in R^n and let T be a subset of S. T is said to be linearly separable (with respect to S) if there is a vector $\underline{a}$ in R^n and a real number Θ such that the vector $\underline{x}$ of S is in T if and only if $\underline{a}.\underline{x} > \Theta$. ($\underline{a}.\underline{x}$ denotes the scalar product of $\underline{a}$ and $\underline{x}$). This condition is equivalent to saying that there is a n-1 hyperplane separating T and S-T so that T is linearly separable if and only if S-T is. The problem is to find a vector $\underline{a}$ and a threshold Θ given a sequence of elements in S and the knowledge of whether or not they are in T. For the purposes of learning, a perceptron can be regarded as a device which is presented with a sequence of $\underline{x}$ elements in S, at each input it outputs a 1 or 0. It is desired to have an output of 1 just for those elements of T. If then an output is incorrect, the weights are adjusted in a way which depends on what sort of error was made. Now a perceptron gives an output $\lceil \underline{a}.\underline{x} > \Theta \rceil$. In order to give the weight adjustment procedure, it will be easier to work with the set S' of <u>augmented</u> vectors. The augmentation of $\underline{x} = (x_1, x_2, \ldots x_n)$ is the vector $\underline{y} = (x_1, \ldots x_n, -1)$. T' is the set of augmentations of vectors in T. The Perceptron's behaviour is then described by $\lceil \underline{x}.\underline{y} > 0 \rceil$, where $\underline{w}$ is the vector in R^{n+1} $(a_1, \ldots a_n, \Theta)$. We shall write this function $\delta(\underline{w}.\underline{y})$ where δ is the obvious step function.

Now we give the procedure for adjusting the weights $\underline{w}$.

On the presentation of a vector $\underline{y}$:

1)   if $\delta(\underline{w}.\underline{y}) = 1$ and $\underline{y}$ is in T' or if

     $\delta(\underline{w}.\underline{y}) = 0$ and $\underline{y}$ is in S' - T' then

     $\underline{w}$ is left unchanged.

2)   if $\delta(\underline{w}.\underline{y}) = 0$ and $\underline{y}$ is in T' then $\underline{w}$

     is increased by $c\underline{y}$.

3)   if $\delta(\underline{w}.\underline{y}) = 1$ and $\underline{y}$ is in S' - T' then

     $\underline{w}$ is decreased by $c\underline{y}$

here c is a positive constant.

Let $\underline{y}^{i(r)}$ be a sequence of vectors in S' where $r = 1, 2, \ldots$
Then the perceptron convergence theorem states that if T is
linearly separable then the vector $\underline{w}$ only changes a finite number
of times. In particular, if the sequence $\underline{y}^{i(r)}$ is such that
each member of S' occurs in it infinitely often, then after a
finite number of presentations it will have 'learnt' T', that
is to say it will give correct outputs.

The condition that T, and consequently T' are linearly
separable says that there is a vector $\underline{w}^*$ such that $\delta(\underline{w}^*.\underline{y}) = 1$
if and only if $\underline{y}$ is in T'. We are now in a position to write
the procedure given above as:

     $\underline{w}$ is changed to $\underline{w} + c\underline{y}(\delta(\underline{w}^*.\underline{y}) - \delta(\underline{w}.\underline{y}))$.

at each presentation of a vector $\underline{y}$ in S'.

In order to prove the convergence theorem, we first need the lemma:

If T is linearly separable, then for any positive number K we can find a $\underline{w}^*$ such that:

$$\underline{y} \ \epsilon T \Rightarrow \underline{w}^* \cdot \underline{y} > K$$

$$\underline{y} \ \epsilon \ S'-T' \Rightarrow \underline{w}^* \cdot \underline{y} < -K$$

Proof. Since T' is finite, the condition that there is a $\underline{w}$ such that $\underline{w} \cdot \underline{y} > 0 \Leftrightarrow \underline{y} \ \epsilon \ T'$ means that we can find two positive numbers ϵ, δ such that

$$\underline{y} \ \epsilon \ T' \Rightarrow \underline{w} \cdot \underline{y} > + \delta 2 \epsilon$$

and $\quad \underline{y} \ \epsilon \ S'-T' \Rightarrow \underline{w} \cdot \underline{y} \leq 0$

Now recalling that the last component of $\underline{y}$ is -1 we can add $\delta + \epsilon$ to the last component of $\underline{w}$ to get $\underline{w}'$ such that

$$\underline{y} \ \epsilon \ T' \Rightarrow \underline{w}' \cdot \underline{y} > \epsilon$$

and $\quad \underline{y} \ \epsilon \ S'-T' \Rightarrow \underline{w}' \cdot \underline{y} < \epsilon$

Finally we can multiply the vector $\underline{w}'$ by the quantity K/ϵ to get the desired result.

To prove the main result we first examine the change in the squared distance between $\underline{w}$ and $\underline{w}^*$ at the presentation of $\underline{y}^{i(r)}$. To start with an arbitrary choice $\underline{w}^1$ is made for $\underline{w}$, at the presentation of $\underline{y}^{i(k)}$ it changes from $\underline{w}^k$ to $\underline{w}^{k+1}$. Writing $\underline{y}$ for $\underline{y}^{i(k)}$ the squared distance change is:

$$|| \underline{w}^* - \underline{w}^k ||^2 - || \underline{w}^* - \underline{w}^{k+1} ||^2$$

$$= -(\underline{w}^{k+1} - \underline{w}^k) \cdot (\underline{w}^{k+1} - \underline{w}^k) + (2(\underline{w}^{k+1} - \underline{w}^k) \cdot (\underline{w}^* - \underline{w}^k)$$

which is, by the expression for the weight change

$$-c^2 \underline{y} \cdot \underline{y} (6(\underline{w}^* \cdot \underline{y}) - 6(\underline{w}^k \cdot \underline{y}))^2$$

$$+ 2c(\underline{w}^* \cdot \underline{y} - \underline{w}^k \cdot \underline{y}) (6(\underline{w}^* \cdot \underline{y}) - 6(\underline{w}^k \cdot \underline{y}))$$

The first term here is non-positive and the second (since 6 is a non-decreasing monotine function) is non-negative. Moreover, whenever a weight change occurs, the first term will have magnitude $c^2 \underline{y} \cdot \underline{y}$ and the second will have magnitude at least $2c | \underline{w}^* \cdot \underline{y} |$, the second term is only non-zero when $\underline{w}^* \cdot \underline{y}$ and $\underline{w}^k \cdot \underline{y}$ have different signs. By the lemma, we can choose $\underline{w}^*$ so that

$$\underline{w}^* \cdot \underline{y} > \left(c^2 \max_{\underline{y} \in S'} (\underline{y} \cdot \underline{y}) + b \right) \Big/ 2c$$

for any positive b and for all y in S', again using the finiteness of S'.

In this way we can ensure that the squared distance change between $\underline{w}$ and $\underline{w}^*$ is greater than b every time $\underline{w}$ is changed, but this distance cannot decrease indefinitely so $\underline{w}$ can only change a finite number of times, which was the result required. (Prop. 6.1.1.).

6.2.. Continuous analogues of Perceptron Convergence

Perceptron convergence is a particular form of feedback
to a system which can only output a 0 or 1.   Suppose we have
a perceptron like system which realises a function $\delta\,(\underline{w}.\underline{y})$
where δ is now some more general function.   Moreover, suppose
that there is a 'desired' output from this system $\delta\,(\underline{w}^{*}.\underline{y})$ say,
we shall show how, in certain cases, the same or similar weight
adjustment procedures produce a convergence to this desired state.
Suppose, in fact that the weight adjustment is, as before, given
by:

$\underline{w}$ is increased by $c\underline{y}(\,\delta\,(\underline{w}^{*}.\underline{y}) - \delta\,(\underline{w}.\underline{y}))$

at the presentation of $\underline{y}$

<u>Prop. 6.2.1.</u>   if δ is a monotone non-decreasing function such
that

$$\delta\,(a) - \delta(b) < \frac{M}{2c}\,(a-b)$$

where $M = \max\limits_{\underline{y} \in S'}\,(\underline{y}.\underline{y})$.   Then $|\,\underline{w}^{*} - \underline{w}\,|$ never increases and if
$\delta\,(\underline{w}^{*}.\underline{y}) \neq \delta\,(\underline{w}.\underline{y})$, $|\,\underline{w}^{*} - \underline{w}\,|$ decreases.   The proof is immediate
from examination of the expression for the squared distance
change.   Note that the condition on δ above means that δ must
be continuous.   This ensures:

<u>Prop. 6.2.2.</u>   If $\underline{y}^{i(r)}$ is as before, a series in which each $\underline{y}$
in S' occurs infinitely often, then for any positive ϵ there is
an integer N_{ϵ} for which

$\delta\,(\underline{w}^{*}.\underline{y}^{i(r)}) - \delta(\underline{w}^{r}.\underline{y}^{i(r)}) < \epsilon$ for all $r > N_{\epsilon}$

Suppose the contrary, then there would be a subsequence of the positive integers $\sum$ and an ϵ such that

$$\delta(\underline{w}^* \cdot \underline{y}^{i(r)}) - \delta(\underline{w}^r \cdot \underline{y}^{i(r)}) > \epsilon \text{ for all r in } \sum$$

so that for each r in $\sum$, $\| \underline{w}^* - \underline{w} \|$ decreases. Also, $\sum$ must be such that some $\underline{y}$ in S' must occur infinitely often, for this $\underline{y}$ the decrease in $\| \underline{w}^* - \underline{w} \|$ is then a continuous positive function of $\underline{w}$ and hence $\| \underline{w}^* - \underline{w} \|$ must tend to zero. Therefore, $| \underline{y} \cdot \underline{w}^* - \underline{y} \cdot \underline{w} |$ must also tend to zero and, by the continuity of δ, this contradicts our initial supposition.

The condition for this convergence can be interpreted in the following way. If we are given δ, a continuous monotone increasing function, and we are given S' and in consequence $M = \max_{\underline{y} \in S'}(\underline{y} \cdot \underline{y})$ then c must be chosen so that

$$c < \frac{M}{2} \sup_{a \neq b} \frac{a - b}{\delta(a) - \delta(b)}$$

$$= \frac{M}{2} \sup \frac{1}{\delta'(x)} \quad \text{(if } \delta \text{ is differentiable)}$$

so that the weight adjustments cannot be too large. If c exceeds this amount the weight vector $\underline{w}$ may oscillate rather than converge, and it is easy to construct examples in which this happens. This means that the rate of convergence or learning can only be increased at the risk of some sort of oscillatory behaviour. It could be possible to have a c which was varied with the $\underline{y}$ in order to gain some increase in the convergence rate,

but we would have to compute or be given $\underline{y} \cdot \underline{y}$ at each presentation of $\underline{y}$.

6.3. Learning from a continuous input

Up to now we have thought of the presentations of the $\underline{y}$ as discrete events, and the time intervals between these presentations being of no importance. We get a rather nice variant of these learning procedures if we think of S' as a connected and now not generally finite subset of R^{n+1} (recall that S' is a set of augmented vectors in R^n). The function $\underline{y}(t)$ is then input to the system, and at the same time we adjust the weight vector which is now also a function of time, $\underline{w}(t)$. It is immediately obvious that it is worth trying weight adjustments of the form:

$$\frac{d\underline{w}}{dt} = c\underline{y}(\, \sigma(\underline{w}^* \cdot \underline{y}) - \sigma(\underline{w} \cdot \underline{y}))$$

Now consider the rate of change of $\| \, \underline{w}^* - \underline{w} \, \|^2$
this is $-c(\underline{w}^* \cdot \underline{y} - \underline{w} \cdot \underline{y}) \, (\, \sigma(\underline{w}^* \cdot \underline{y}) - \sigma(\underline{w} \cdot \underline{y}))$
which is always non-positive if σ is any monotone non-decreasing function. The conditions that all $\underline{y}$ in S' are 'learnt' are firstly that S' is compact, second, that $\underline{y}(t)$ fills S' that is for any open set U of S' and any time t_1 there is a time $t > t_1$ at which $\underline{y}(t) \in U$.

Prop. 6.3.1.   Under these conditions for any positive ϵ there
is a time t_ϵ for which

$$\delta(\underline{w}^* \cdot \underline{y}) - \delta(\underline{w} \cdot \underline{y}) < \epsilon \text{ for any time } t > t_\epsilon .$$

The proof follows in almost the same way to the corresponding
proof in the previous section.   It is not necessarily true
that this sort of convergence means that $\underline{w}$ need get close to
$\underline{w}^*$ in an absolute way, i.e. $\| \underline{w}^* - \underline{w} \| \to 0$.   This will only
have to happen when S' spans the n dimensional subspace of
R^{n+1} consisting of all those vectors whose last coordinate
is -1.

This concludes the study of learning and memory models
and it is hoped that their relation to one another has been
clarified.   In particular it has been shown that the crucial
distinction to be made is between the two methods of causing a
change to take place.   The method of causing a weight change
in a perceptron is very different to that which alters a switch
or weight in an associative net.   We now turn to the neurological
evidence for these types of change at a synaptic level in an
attempt to determine the importance of these ideas as biological
models.

CHAPTER 7   Synaptic plasticity

It is generally held that persistent changes in synapses - synaptic plasticity - form the physiological basis for learning and memory.   Of other possibilities, the proposals of Rashevsky[1] and Householder and Landahl[2] that the brain can sustain permanently circulating patterns of activity, (analogous to the mercury delay lines of some computers) do not seem compatible with the ability of conditioned behaviour to survive the suppression of all electrical activity in the brain by freezing or anaesthesia.   Nor is it likely that such activity would remain undistorted by epilepsy or electro-convulsive therapy.

The idea that permanent changes may take place at a neuronal level, for example changes in the threshold of a nerve cell is a possible alternative to synaptic plasticity although it certainly does not exclude it.   As an alternative it is less attractive since there are many more synapses than nerve cells in the central nervous system;   the ratio is of the order $10^4 - 10^5$ in the cerebral cortex (Cragg[3]) and the amount of information that could be stored synaptically would be correspondingly higher.   Moreover there is now a certain amount of neurophysiological evidence, which we shall shortly examine, which indicates that changes can take place in the conductivity of a pathway which are specific to a synapse, or a group of synapses in that pathway, rather than a whole cell.

First of all we shall describe some of the physiological demonstrations of persistent changes in the nervous system and the evidence that associates learning with changes in certain synaptic structures.   Later we shall discuss the possible neurological mechanisms which could effect such changes.

Hebb's[4] suggestion that synaptic changes could persist for the life of an animal has not yet been amenable to any substantial test.   But there are experiments which seem to indicate that some change takes place which can persist for at least the time for which one can record from a cell.   Such changes have been demonstrated in the cortex, but may well occur elsewhere: Lashley[5] has shown that decorticate animals can be conditioned in limited ways, and certainly some animals whose development has not progressed to the stage of having any neocortex can perform simple learning tasks.

Morell[6] found that epilepsy induced in an area on one side of the cortex would cause a similar epileptic focus in the corresponding contralateral area.   Even after the callosal fibres connecting these areas had been severed, the increased contralateral activity persisted indefinitely.   Related experiments were performed by Bindman, Lippold and Redfearn[7], and Gartside and Lippold[8].   In the first of these it was shown that a small current passed through the cortex of an anaesthetized rat would increase the spontaneous firing rates of cells and

this rate increase would persist at least for a few hours after the stimulating current had been turned off.   In the second experiment, a similar result was obtained by locally freezing the cortex.   Since the effect of freezing is to produce a potential gradient in a direction opposite to that in the first experiment, it was thought that the persistent increase in the firing rate may have been attributable to its initial increase rather than any other effect associated with the depolarisation.

Bliss, Burns and Uttley[9] were able to investigate this effect in very much more detail.   In their preparation, which was this time isolated but unanaesthetized cortex, they recorded from a cell which could be driven by independent stimuli to neighbouring parts of the cortex or white matter.   They were able to define a 'conductivity' for one of these stimuli, the test stimulus which was the ratio of the strength of response of the cell (number of spikes immediately following the stimulus) to the strength of the stimulus.   If we call the other stimulus the priming stimulus, their results may be by saying that the conductivity

increased when the priming stimulus alone had been used for some time;

decreased when the test stimulus alone had been used for some time;

decreased when these stimuli were 'paired' or used simultaneously.

In order to effect any change in the conductivity it was necessary to drive the cell (by the test, priming or paired stimuli) for at least six minutes, but, having done this, the conductivity changes often persisted for as long as it was possible to record from the cell.

From these experiments one can only guess that synaptic change is involved; the pathways could have involved a series of nerve cells and the change could have been caused by a change in, say, a cell threshold. In a similar set of experiments Kandel and Tauc[10] were able to demonstrate that in their preparation a change in conductivity was almost certainly attributable to a synaptic change.

Kandel and Tauc[10] worked with a nerve cell ganglion from the abdomen of a sea snail. While this is not a region of the nervous system or an animal which would be expected to show any retention, a change in the conductivity of a pathway was demonstrated which was produced by methods similar to those used by Bliss, Burns and Uttley[9]. The cells in this ganglion are large and it is possible to record the intracellulor potential changes. Such cells could be driven independently by two pathways. It was found that:

1.   The sensitivity of the cell to a weak (sub-threshold) stimulus in the test pathway could be greatly increased by driving the cell by a strong priming stimulus in the other pathway.

2. That this sensitivity change could last for some
    minutes after the priming stimulus had ceased.

3. For some cells this stimulus was conditional on
    the stimuli being paired, and in some cases specific
    to one of two possible test pathways.

In a second paper Kandel and Tauc[11] performed related
experiments and were able to hypothesise that the test pathway
was monosynaptic in that the fibre tthey were (test) stimulating
synapsed onto the cell from which they were recording.   They
also proposed that this increase in sensitivity could be attributed
to presynaptic facilitation (see figure 1) in which the priming
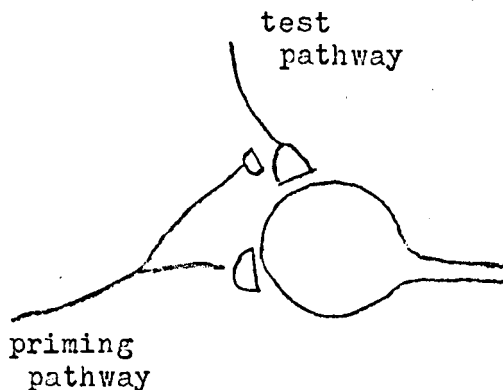

test
pathway

priming
pathway

Figure 1

pathway terminates in a synapse which serves only to modify the
behaviour of a synapse in the test pathway, but does not itself
influence the cell.   Such synapses could have the effect of a
gain knob on a variable gain amplifier and could be used to

control a synapse in a way which would be completely independent of the activity on either side of the synapse. However, there is no evidence to say that this is the underlying mechanism for control of synaptic change in the cortex and in the next section we shall examine the possible controls that could exist for synaptic change.

7.1. Mechanisms for control of synaptic change

It is not known what happens when a synaptic knob increases its effect on a membrane. It may simply grow (Young[12], Cragg[13]) and more effectively depolarise or hyperpolarise a membrane on account of its increased area of contact. Other experiments suggest (Brown and Pascoe[14]) that the ability to produce more transmitter substance is modified. Another proposal is that entirely new knobs may sprout (Eccles[15]). In spite of this it is reasonable to ask what controls such a change; and the answer may depend only in its details on the precise mechanism of change.

By supposing (like Pitts and McCulloch[16]) that the reaction of a nerve cell to its inputs can be described by a Boolean function of its inputs, Brindley[17] has listed some of these possible mechanisms and devised a notation for describing them. Brindley's notation assigns to each cell and its inputs an output function S of the form:

$$S = w \; v \; \alpha < u \mid v > \quad \ldots\ldots\ldots\ldots\ldots \quad 7.1.$$

w, u and v are Boolean functions of the inputs, the notation
means that if inputs satisfying v=1 are given sufficiently often,
then

1) S changes from w to w $\vee$ u if α =f (for a
facilitated synapse)

2) S changes from w $\vee$ u to w if α = h (for a
habituated synapse)

In addition it is stipulated that w $\wedge$ u = 0 and it is always
possible to choose w and u to satisfy these conditions.   As
examples of Brindley's notation can be used to describe the
following systems (the triangular synapses are modifiable).
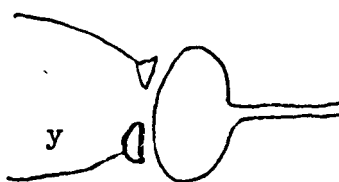
1) Presynaptic facilitation of an excitatory synapse
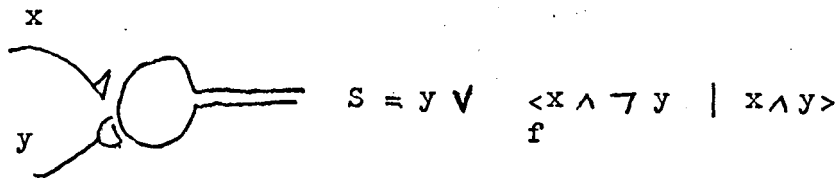
x $S = {}_f\!<x \mid x>$

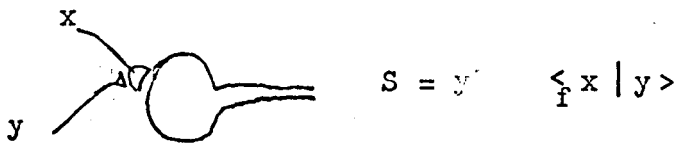
2) Postsynaptic facilitation of an excitatory synapse

x

y $S = y \vee {}_f\!<x \wedge \neg y \mid y>$

3) Excitatory synapse facilitated by pairing

$$S = y \vee \underset{f}{\langle x \wedge \neg y \mid x \wedge y \rangle}$$

4) Excitatory facilitated by a presynaptic knob

$$S = y \quad \underset{f}{\langle x \mid y \rangle}$$

Brindley points out that (3) is an example of the type of conditioning postulated by Hebb[4] and that various other of these functional arrangements have been found or postulated.

We may use Brindley's notation to formalise the idea that some synapses can change only as a function of the activity in the synaptic knob and the post-synaptic membrane. We call such a synapse <u>locally</u> <u>modifiable</u>.

If x is an input which terminates in a locally modifiable synapse, then if the output of the cell is described by

$$S = w \vee \underset{f}{\langle u \mid v \rangle}$$

v (the modifying stimulus) will be of the form $v = \phi (w,x)$ where ϕ is a Boolean function. If the output of the cell is described by

$$S = w \vee \underset{n}{\langle u \mid v \rangle}$$

then for a locally modifiable synapse: $v = \phi(w \lor u, x)$,

It is easy to verify that of the four types of synapse described above, the first three are locally modifiable and the last is not.

Brindley lists many more examples of modifiable synapses and classifies them according to the rules: Class A if $u \land v = u$; Class B if $u \land v = 0$; and others are placed in Class C.   Now although Brindley's classification does not imply this, it is interesting that of the examples he lists, all those in Class A and Class B are locally modifiable and that synapse in Class C (case 4 above) is not.   Naturally, if Brindley's Boolean description of synaptic plasticity is not totally accurate, which it probably isn't, our definition of a locally modifiable synapse fails and the definition of a locally modifiable synapse would have to be couched in the specific terms of another model.   Nevertheless we have been able to state in precise terms the fact that a locally modifiable synapse modifies only as a function of the activity on either side of the synaptic cleft.

Finally it is worth noting that there is copious anatomical evidence for the existence of presynaptic (for example Gray[13]) knobs and that these as we have seen, would provide a mechanism for non-local modifiability.   It has not yet been established whether all possible types of local modification can take place in the nervous system.   In particular there seems to be no

absolutely conclusive evidence of the synapses proposed by

Hebb.   As far as can be ascertained.from the individual record-

ings of Bliss, Burns and Uttley, their results could be explained

by local modifiability, even if the cortical pathways they tested

were monosynaptic.

CHAPTER 8 <u>Neurological restrictions on models</u>

Before examining the neurological relevance of the models advanced earlier, we put forward a hypothesis about synapses which will restrict the applicability of these models and allow us to put forward further testable neurological hypotheses. This is:

<u>Synapses in the neocortex can only be modified locally.</u>

If such is the case, then any model which is thought to describe any of the higher functions of the nervous system must conform with such a restriction. Also we may gain some insight into the exact nature of these 'higher' functions if we can advance plausible models meeting this restriction.

The anatomical evidence for this hypothesis, if not conclusive, indicates that if there are presynaptic knobs in the cortex, they are not frequently seen. Gray[1], in his famous electron microscope study of cortical synapses states that they are either axo-somatic or axo-dendritic. Cragg has remarked in personal communication that he has never observed such structures in the cortex. However, presynaptic effects may not be the only mechanism for non local modification and the alternatives are worth considering.

Non local modification may be the result of postsynaptic effects. For example a synapse might be capable of permanently

altering a small region of the postsynaptic membrane so that
the conductivity of a neighbouring synapse or synapses would
be impaired.   A rather striking piece of anatomy which illustrates
this possibility is the structure of the granule cells of the
cerebellum shown in Figure 1 which is taken from a drawing in
Eccles, Ito and Szentogothai[2].

granule cell   parallel fibre output   Golgi cell
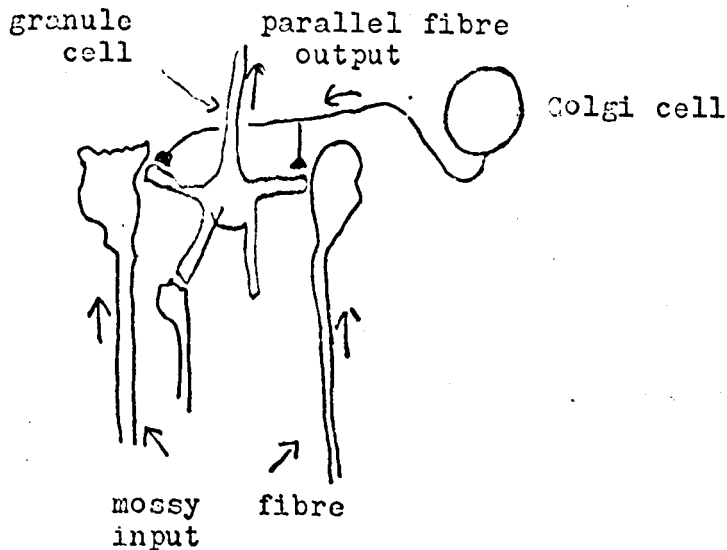
mossy input   fibre

Figure 1

The mossy fibre input excites the granule cell and the Golgi
cell inhibits it.   The appearance of the inhibitory Golgi
synapses indicates that they are in one to one correspondence
with the excitatory mossy fibre synapses.   Could it be that the
Golgi cell synapses exert a local effect on the granule cell
membrane?

If not, one wonders why the inhibitory effect of the Golgi cell is not achieved by axo-somatic contact. However there is no physiological evidence from the cerebellum to corroborate this postulate though Diamond[3] has shown that in the Mauthener cell of a goldfish inhibition can be specific to inputs distal (with respect to the axon) to the inhibitory synapse. Rall[4] has also demonstrated on theoretical grounds that it is in general wrong to regard the polarisation of a cell membrane as the sum of the polarisations and depolarisations caused by each synapse. Cells like the granule cells do not seem to be present in the cortex, the spines on the dendritic trunks of pyramidal cells might perform a similar function to this hypothetical property of the granule cells, but we have no anatomical or physiological evidence to confirm this and hence no certain objection to the hypothesis about neocortical synapses.

Griffiths[5] has suggested that non local modification could take place by the diffusion of some chemical throughout the neo-cortex and this is a serious possibility if it is not required that the modifying influences should be very specific. Another possibility that will be considered later, is that there may be some temporal coding of events at a synapse which would lead to modification. This would imply that the function of a synapse is rather more complicated than is normally supposed and would, in particular make Brindley's classification inadequate.

The adaptive nets of Chapter 2 and the perceptrons of
Chapter 6 differ fundamentally in respect of what type of
synapses would be required for their operation. We now take
the step of supposing that the weights, connections or switches
mentioned in these chapters are to be identified with synapses.
It was pointed out that the adjustments required to modify an
associative net when a new pair of patterns are to be stored
could be effected by Hebb synapses and we have seen that such
synapses are locally modifiable. At least our hypothesis about
modifiable synapses does not exclude the possibility that
associative nets are realised in one of their forms in the cortex.

For perceptron learning and its associated forms discussed
in Chapter 6, it is much more difficult to understand how the
weight adjustment could be accomplished with locally modifiable
synapses. The wiring diagram of a perceptron is illustrated
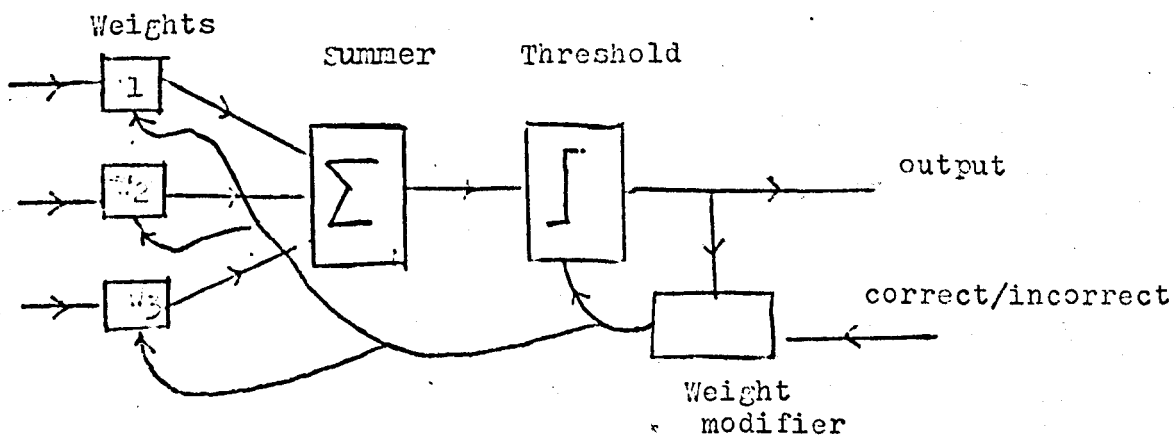in Figure 2.

Figure 2

Now although it is reasonable to suppose that the process of
taking a weighted sum of the inputs could be realised by one
nerve cell, it is not at all clear that the feedback to the
thresholds and synapses could be accomplished by the same cell.
The two natural possibilities are:

1. That this feedback is accomplished by pre-
   synaptic knobs which habituate or facilitate
   each synapse (together with some control on
   the threshold of the cell).

2. That the weights are associated with cells
   rather than synapses and weight adjustment
   is accomplished by altering a parameter of
   the whole cell. This requires there to be
   one cell for each weight. Taylor[6] has
   proposed a scheme for the cortex which works
   in this way.

The second possibility is somewhat unattractive on the
grounds that it is uneconomical in comparison with the first
and would indicate a set of "weight" cells with one effective
input and one effective modifier.

The former possibility would be excluded from the cortex
by our hypothesis since it involves the non local modification
of synapses.

But there is nothing to exclude this type of learning from other parts of the nervous system - first of all pre-synaptic knobs are seen in the spinal cord and in profusion in many thalamic nuclei (Gray[7], Purpura[8]). Secondly there is a certain amount of physiology which lends support to this: Kandel and Spencer in a review article of neurophysiology in the study of learning distinguish two types of conditioning.

In type 1 conditioning a behavioural stimulus (UCS) gives rise to some well defined response which is not elicited by some other stimulus (C.S.); often pairing these two stimuli a number of times the C.S. elicits the response without the U.C.S. This of course, is classical Pavlovian conditioning and the experiments of Kandel and Tauc provide a neural analogue of this situation.

Type II conditioning - often called operant conditioning is quite different and requires that some reinforcement take place if a response is given to a particular stimulus. For example, in a Skinner box an animal might be rewarded with a pellet of food if it presses a lever in response to the sound of a buzzer. Initially the animal presses the lever by accident and subsequently learns that the reward is produced by doing so in response to the buzzer. The reinforcement in this case is positive - in other cases the reinforcement can be negative, that is the animal learns to avoid a painful stimulus. The basic distinction is that in type II conditioning, one of the

stimuli is conditional on a response.

The possibility that either the C.S. or the U.C.S. in type I conditioning might be an electrical stimulus to the central nervous system was investigated by Loucks[10] who found that a C.S. consisting of a direct stimulus to the visual cortex of a dog could be effectively paired with a U.C.S. of a shock to the foreleg. Giurgea[11] used as U.C.S. and C.S. electrical stimuli to the motor on visual cortices respectively; the response being some movement elicited by stimulating the motor cortex. Finally Bures and Buresova,[12] were able to use a behavioural C.S. and the local polarization by an extracellular electrode as a U.C.S. which induced an increased or decreased firing rate in some cell from which their electrode would also record. The C.S. in this case an auditory stimulus was effective for certain cells in the thalamus and inferior colliculus though the conditioned response usually disappeared rapidly after pairing had ceased. To summarize: it seems that type I conditioning to electrical stimuli can occur anywhere in the nervous system, even in very primitive structures of simple animals. But in general the effects are most permanent when the stimulus is given to parts of the higher nervous system - in particular to the neocortex.

Type II conditioning has also been the subject of similar investigations. Delgado[13] showed that electrical stimulation of

parts of the brain, mainly in the region of the hypothalamus

could substitute as positive or negative reinforcement.

Stimulation in this region could cause the animal to manifest

rage or fear and the latter could be used to train the animal

to avoid certain types of food. Olds[14] advanced this work and

showed that rats would persistently stimulate themselves through

electrodes implanted near the hypothalamus in the medial forebrain

bundle. Other electrodes in the same region would produce

avoidance or aversive behaviour. Olds tested other parts of

the brain in this way and found that the most striking results

were produced from regions (such as the median forebrain bundle)

which are closely associated with the hypothalamus. No evidence

of positive or negative reinforcement was produced from electrodes

in the neocortex or in the sensory areas of the thalamus.

Olds and Olds[15] subsequently investigated whether or not they

could, by the same stimulus to the medial forebrain bundle, re-

inforce the firing rate of a single cell. Here they were able

to get positive results in the same regions as before including,

significantly, the mammilo thalamic tract - one of the pathways

by which connects the hypothalamus to the thalamus. Only with

great difficulty could they train cells in the neocortex and

then only rather weakly. Later, Olds and Milner[16] were able to

produce the same learning in cells by using a behavioural re-

inforcement such as food; again they found it very much easier

to train cells in areas connected directly with the hypothalamus.

It is an attractive conjecture that physiological type I conditioning requires only locally modifiable synapses but that type II conditioning cannot be achieved this way and perhaps requires presynaptic terminals. This would explain the absence of presynaptic knobs in the cortex and account for the apparent difficulty in producing type II conditioning in neocortical cell responses. Some caution is needed in interpreting the physiological results in this way. The cells or axons which could be conditioned were almost certainly involved in a more complicated pathway that was also being conditioned. It is not in general possible to say that the cell which is recorded is necessarily the cell at which the changes take place. It could be that it is one of the cell's predecessors which is responsible for the altered behaviour of that cell.

However, these results suggest some further experiments in order to confirm, both anatomically and physiologically the hypothesis about locally modifiable synapses.

1) Can a cortical cell be type I conditioned to a
behavioural stimulus? For example in the right
association area of the cortex is it possible to
condition a cell which normally responded to a
visual stimulus, also to respond to an auditory
stimulus? This experiment would be exceptionally

convincing if done in conjunction with a type II
conditioning experiment in which the animal was
rewarded if the cell did not respond or punished
if it responded.

2) Can areas of the hypothalamus be demonstrated to
    modify other pathways presynaptically?   It might
    be shown, for example, that the mammillo-thalamic
    tract or even the mammillo-tegmental tract operated in
    this way.

3) Is it possible to get further anatomical evidence
    for the previous conjecture?   Perhaps one of these
    tracts could be traced, by degeneration studies or
    otherwise, to presynaptic terminals.

4) The paleocortex cannot be classified with the neocortex
    in this way.   It is structurally different (Andersen)[17]
    and exhibits type II conditioning (Olds and Olds)[15].
    It is organised in a relatively simple manner and
    could perhaps be the subject of similar experimental
    tests.

On the other hand it has been conjectured that the hippo-
campus is part of the mechanism responsible for 'laying down'
memory traces, in this case we cannot at the moment attempt any
simple generalisations about the structure or function of its

synapses, but again it would be interesting to know if there are presynaptic or axo-axonal contacts in this region.

Relevant to this discussion of the modification of synapses is a suggestion by Young[18] on the function of the amacrine cells in the brain of an octopus.   In a model of type II conditioning which bears some relation to the learning machines of Chapter 6 he suggests the need for presynaptic inhibition and suggests that amacrine cells (small nerve cells with apparently no axons) might effect this by interposing themselves between the pre and post synaptic membranes.   Young's model has features which overcome many of the difficulties which would be encountered in taking perceptron learning as an exact description of a neurological process.   Whether or not it is realised entirely in the neocortex of mammals is, like the perceptron, subject to what has already been said about the local modification of synapses.

CHAPTER 9   The relevance of models

In the previous chapter the one crucial difference between learning and remembering devices. was examined in some detail. We now take the models individually and attempt to subject them to neurological criticisms.   As usual, this will not be entirely conclusive and demands answers to further neurological questions.

9.1.   Holography

In its general form holography requires a propagative medium and a stable periodic source.   If the brain or part of it constitutes such a medium and contains such a source, what is the speed of propagation and the frequency of the source? There seems to be little evidence that such a source exists; the E.E.G. rhythms might be taken as such evidence but these are neither stable nor very persistent.   Even if there was such a source - the simplest oscillator, consisting of two neurons connected in a loop would be unlikely to have a frequency of more than 100 c.p.s.

As for propagation it might happen along nerve fibres and in this case the slowest possible conduction rate would seem to be of the order of 1 metre per sec.   The wavelength is then not less than 1 cm.

Now the interference fringes produced by two such waves

would always be spaced at a distance greater than this wave-length. According to Sholl[1] a square centimetre of cortex contains, very roughly, 2×10^6 neurons and, in consequence of the order of 10^{10} synapses. To store a signal so far above the limits of resolution of the medium is at least uneconomical. Moreover, in a linear system, any superposition of interference fringes would not contain higher frequencies so this very high redundancy would apply as well to the storage of many holograms. It is hard to believe that a non-linear system which approximated this would exhibit a much more economical one of the available store (if this is synaptic).

A proposal advanced by Beurle[2] and van Heerden[3] suggests that the propagation may be much slower and consist of spreading waves of excitation throughout a mass of cells. If the conduction rates were much slower, the wavelength could be reduced to a more acceptable quantity. Again, if this mass of cells were in the cortex, the propagation would presumably take place laterally.

The phenomenon of spreading depression in which a wave of activity is observed to pass slowly across the cortex might be taken as evidence of such propagation, though it cannot be taken too literally since it results in prolonged inactivity in most of the cortex. Observations by Sperry[4] and Lashley[5] however, would not lend support to this idea. They were unable

to detect any change in the behaviour of animals in which they had attempted to interfere with lateral conduction in the cortex by cutting through it in many places, or by inserting lengths of electrically conducting material.   It is also hard to imagine how any recall could be sufficiently fast if the rate of conduction was only a few millimetres a second.   Finally, van Heerden's proposal involves propagation through a linear medium which does not reconcile with the known non-linearity of neural responses at least in the lower parts of the nervous system.

9.2. Associative nets

Two types of associative nets have been advanced, and each of these have points in their favour as neurological models. Since it is fairly certain that a synapse is either excitatory or inhibitory and cannot change from one type to the other, each input line of the linear net would have to branch and give rise to both an inhibitory and excitatory input to each output line. Also, since nerve cells seem to possess entirely inhibitory or entirely excitatory synapses, we have to postulate a set of interneurons to achieve this.   A very much more serious limitation on the linear associative net is that the weights can become very large.   The strength of a synapse is probably limited so that to get large weights may involve the growth of new synapses.   Eccles has suggested this as a mechanism for synaptic change, but this is only a conjecture.   These diffi- culties are avoided in the binary associative net which could

be realised with one type of excitatory synapse. The need
for an accurate threshold, and the need for the nervous system
to produce patterns with specific numbers active and inactive
lines present new obstacles to which we do not yet have any
complete solution.

Both types of associative net require that synaptic
change only takes place on the presentation of a pair of
patterns to be associated, the binary net requires just one
type of irreversible change.

Suppose that such synapses were of Hebb's type. How
could a synaptic change be prevented from happening, as
a result of noise, or in the linear net, simply in the use of
the net to recall a pattern? One solution is to suppose that
the pairs of patterns must be paired many times in order to
produce synaptic change and it is possible that the function of
the hippocampus and its related structures is to maintain such
a pairing, for bilateral destruction of these regions produces
Korsakoff's syndrome,[7] the inability to establish new memories
'or to form new associations.

The need for repetitive pairing in order to effect synaptic
change is certainly indicated by the experiments of Bliss, Burns
and Uttley.[8] A factor which may also be important is the relative
timing of the C.S. and U.C.S. to a synapse; it may be that the

most effective stimulus for producing synaptic change is not
the simultaneous depolarisation of both the presynaptic and
post-synaptic membrane but these two events occurring with a
slight time difference.   Suppose for example that the U.C.S.
were to precede the C.S. by a few milliseconds.   Some reverse
chemical diffusion could take place from the post-synaptic
membrane which would reach the presynaptic membrane at the
time of arrival of the U.C.S.

Note also that if the cells were being fired by the C.S.
alone, such a time delay would not occur so that the synapse
would 'know' when and when not to modify.   Such a timing
hypothesis could again be tested by methods similar to those
of Bliss, Burns and Uttley.[8]   It is of course equally possible
that the time delay could be the other way round, that is to
say, that the C.S. should precede the U.C.S.   The latter is the
most effective order for behavioural conditioning.

The synaptic contacts on the dendritic spines of pyramidal
cells (Gray)[9] are peculiarly suited to this function.   First
the spines contain a specialised process, the spine apparatus
which could form part of a system which communicates any post-
synaptic activity to some other part of the synapse.   Second,
such spines lie on the main dendritic trunks and would be
necessarily affected by any (active   or passive) depolarisation
of the apical dendrites.   Moreover the apical dendrites
(Lorente de No)[10] lie in the outer parts of the cortex which is

the region in which fibres afferent to the cortex (either from the thalamus, or from other parts of the cortex) terminate.

9.3. Group invariant associate nets

There are several possible ways in which such a net might be embodied in the brain and little to say which is more likely. All we can do at the moment is to make some rather general observations. First of all, the complex Hubel and Wiesel[11] cells which respond to a pattern of a certain orientation no matter what its position provide evidence that some sort of group invariant response obtains even in the primary visual cortex. Similarly there are sliding frequency detectors in the auditory cortex which detect rising of falling notes in almost any audible frequency range (Evans and Whitfield).[12] Recently, cells have been found in the visual cortex which respond maximally to patterns presented binocularly, of an arbitrary position but fixed disparity (Pettigrew, Nikara and Bishop).[13] All these are manifestations of simple forms of group invariance; there are probably more complicated forms which might be shown up by simultaneous recording of many cells.

Given that there are group invariant responses in the brain, are the necessary connections formed automatically during the development of the brain or is some sensory exper-ience required to establish them? The answer is likely to be

both of these.   There is an innate rough set of connections,
but the fine details are learnt.   Possibly this rather com-
plicated wiring can only be achieved by growth of nerve cells
over some distance, and in consequence can only happen fairly
early in life.   Such is the case with binocular vision, which
is, in cats, critically dependent on their being sighted in
the first few weeks of life (Hubel and Wiesel, [14]).   If
the acquisition of some group invariance is, as has been
suggested, also essential to the development of language, it
is reasonable to give a similar exploration of our inability
to learn to speak after the age of six or seven.   The latter
seems to be shown by children who have suffered damage to the
dominant hemisphere at an early age.

Finally, as they stand, the group invariant nets and
perceptrons seem to require too high a degree of precision to
be taken as precise neurological models.   The principle of
'tying' synapses together (i.e. altering many synapses at the
same time) nevertheless provides an interesting mechanism for
generalisation.   One of the most serious shortcomings of such
models is the difficulty of turning a continuous (perceptual)
group into a discrete group action on a set of input fibres.
It is possible that the concept of tolerance space (Zeeman)[15]
could be used to get round these obstacles.

9.4. Perceptrons

It has been shown that the fundamental distinction between perceptrons and associative nets lies in the method of adjusting the weights in these devices. Because of this distinction it was felt that the modifiable elements of a perceptron could not represent cortical synapses. We now ask if the method of weight adjustment or learning algorithm for perceptrons represents any other part of the nervous system.

First of all, perceptron learning is in no way a model of operant conditioning. To make this clear, consider for example, a thalamic neuron which we know can be operant conditioned to respond to a stimulus. While there is no response, no reinforcing stimulus is given and there is no evidence that any charges have taken place: a perceptron always alters when an incorrect output is given. Thus if perceptron learning describes thalamic conditioning we should expect that every conditionable cell in the thalamus is being conditioned even when no reinforcing stimulus is given.

After the first reinforcing stimulus is given (that is when the cell responds for the first time) it does not necessaril follow that the response has yet been conditioned: the cell can, and is likely to, not respond to the stimulus at its next occurrence, and many more trials are necessary in order to

condition the response.   A perceptron trained to one input

pattern will, when it has responded correctly once to that

input, will continue to do so.   Also, training a perceptron

to a particular input is likely to interfere with its previous

training on other inputs.   There does not seem to be any good

reason that this happens in the same way with operant conditioning.

Behaviourally, an animal being operant conditioned makes explor-

atory responses, apparently at random, until it makes the correct

response, which is reinforced.   A perceptron, which is limited

to two possible responses in a similar one-stimulus situation,

gives a series of incorrect responses and will thereafter give

a series of correct responses.

It was shown in the discussion of training algorithms for

perceptrons that a perceptron could be thought of as 'hill

climbing' and that each weight change effected a move up this

hill.   Certainly Uttley's[15] conditional probability machine and

Young's[16] diagram for operant conditioning in the octopus would

seem to present more accurate paradigms of operant conditioning.

If then, perceptrons do not provide a model for operant condition-

ing, they may well describe other types of conditioning especially

in improving the performance of an already established task.

Such conditioning could well involve this type of hill climbing

but a neurological representation of it has yet to be discovered.

Also, it has been shown that in certain cases the weights involved

can become very large - in such cases it may be difficult to find a reasonable mechanism for holding these large numbers. On the other hand there is no reason for supposing that predicate such as the parity predicate are at all natural in a behavioural sense.

CHAPTER 10   Conclusions and speculations

No theory of memory or learning, however well founded it is in physiology, is satisfactory unless it advances our knowledge of behaviour.   It is not hard to devise systems which exhibit some form of memory, nor is it usually difficult to build up some neural analogue of such a system.   Our assessment of any such theory must finally rely on our ideas about behaviour.   Some proposals will be made here concerning associative nets, and since it has been suggested that associative nets could exist in the cortex, such proposals will also be interpreted in terms of the function and evolution of the neocortex.   The central proposal is that it is the function of the neocortex to make predictive models of the environment.

To clarify this last statement, assume the following over simplified model of behaviour.   An animal receives a stimulus S at a certain time - it produces a response R and as a result receives a new stimulus S'.   S' will be partly or wholly predictable from S and R.   Consider some examples of this:

1)   In operant conditioning, if some stimulus is given (a bell) and a response is then made (pressing a lever) a new stimulus (food) ensues;   if some other response is made a different (or no) stimulus is received.

2) In the visual perception of some static scene,
   a movement of the eye causes the image on the
   retina to translate through a certain distance.

3) Any muscular contraction or relaxation causes
   altered signals from the various receptors
   associated with that muscle.

In each case here S' is a function of S and R, but in
general S and R do not determine a unique S'. The environment
might not be static as in (2), for example. To take a com-
plicated paradigm, in a game such as chess, suppose that S
is given by the state of the board before a player makes a
move (R). S' - the state before he makes his next move will
depend on his opponent's move so that S and R do not determine
a unique S'. However, to start with let us make the naive
assumption that S' is uniquely determined by S and R.

The claim is that an associative net is used to predict
S' given S and R: in neurological terms this cannot mean
that the stimulus S' is physically reproduced by the nervous
system nor does it mean that the sensory input fibres are
caused to fire in the same way that they would if S' is
actually received. It is suggested that the patterns of
firing of the pyramidal cells (which effect the output of the
neocortex) are similar whether S' is received or not. This

is compatible with the supposition that the neocortex can
be classically conditioned, the unconditional stimulus being
S with R and the conditional stimulus being S'.   It would be
interesting to know whether this effect obtains even in the
sensory cortex.   In the experiments suggested in Chapter
7: can for example, the primary visual cortex be conditioned?
It is then possible that the 'moving edge' detectors of Hubel
and Wiesel[1] are established by this mechanism, and perhaps
such cells continue to fire even if the stimulus is halted or
turned off in the middle of its traverse through the receptive
field.

In many situations there will be some very strong
connections between S and S'.   In one of the examples above,
if R consists of some movement of the eye, S' is approximately
a translate of S on the retina, it would seem unlikely that
each S' in this case has to be conditioned.   What is more
likely is that there is some mechanism for translating S, and
this may be achieved in the same way that group invariant nets
were constructed.   It is unlikely that the 'correlograph' is
an appropriate model here because a great deal of coding takes
place in the visual cortex and a translation of the retinal
image will not cause a similar modification in the pattern of
firing cells.   But it seems likely from Hubel and Wiesel's

work that units of all types and orientations exist with receptive fields in all parts of the fovea - if this is so then the set of such units is closed under the translation 'group' and it is still possible to build associative nets invariant under this operation.

We now digress and turn briefly to the possible use of group invariant associative nets in the use of language. It is not proposed here to expound any modern linguistic theory except to state that it is commonly held that the meaning of an English utterance can be in part represented by some sort of graphical structure. The nodes of such a structure represent words or grammatical constituents. Suppose a set of demonstrative sentences is given and these are passed into one or more of these structures. Any question can also be parsed into another structure and the answer to that question necessitates matching the question structure with part of the original demonstrative structure.

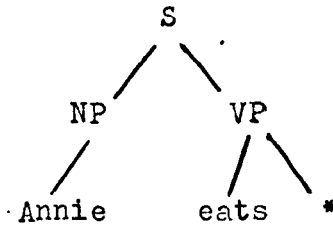Take, for example the demonstrative sentence:

A) Annie eats meat.

Which might be represented by:

```
            S
    NP           VP
    Annie      eats  meat
```
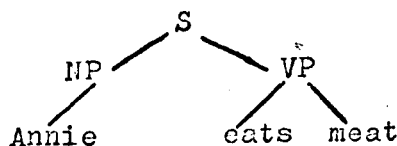
The question:

B) What does Annie eat?

    gives us an incomplete structure:

```
              S
            /   \
         NP      VP
        /       /  \
    ·Annie    eats   *
```

To answer the question B, given A, is now straightforward
the parsing of B must be matched with the parsing of A.   The
missing node in this case is the answer to the question.   In
another case the question:

C) Does Annie eat meat?

    Gets the answer yes if the parsing of A has been

    given in a demonstrative sentence.

In one case, therefore we are seeking to match one graph
with a subgraph of another and then continue it.   In the
other we are simply seeking a direct match of one graph with
another graph or subgraph.   It has been suggested [2] that a
set of demonstrative sentences could be represented by a
large connected graph with common terminals identified: so
that if "Annie" occurs in two sentences we link the parsings

of these two sentences by that word so that we can represent these sentences by one connected graph.

The problem of identifying a graph as a subgraph of some other graph is superficially like that problem which is resolved by holography. The graph recogniser in Chapter 4 shows that a group invariant associative net can perform this operation. In fact the graph recogniser worked on graphs with undistinguished nodes, that is it worked simply on the relation which defined the graph. In the linguistic problem posed above we would not want to allow a noun phrase node to be identified with a verb phrase node. Such a restriction would mean that an associative net built for this purpose would work on a subgroup of the full permutation group of the nodes and perhaps on a subset of all possible relations between nodes. These two restrictions would mean that the size of such a net would be reduced and, since fewer switches would be 'tied', that the information that could be stored in it would be very much greater.

It is not of course proposed that group invariant associative nets will properly solve any linguistic model, nor is it more than a very naive model for how the brain copes with a particular problem. It is intended to show that the graph

isomorphism problem which crops up in several areas of artificial intelligence[3] can be tackled by parallel systems rather than the relatively slow programmes that have been written to cope with it.[4]

To reconcile this proposal with what has been said about models of the environment is, at the moment, a very difficult task. But if it seems too far-fetched to believe that the mechanisms for producing language are fundamentally similar to those for making predictive models of the environment, it is at least worth noting that it is an important, and possibly the only, function of language to communicate expectations about the environment. It is certain that whatever mechanisms are involved in these abilities, they are closely linked.

What use is the ability to model the environment and why did it develop? Why is it a necessary addition to whatever subcortical learning machinery that exists? It has been seen that some devices which can be operant conditioned can suffer from drawbacks such as taking many trials to respond correctly to many different stimuli and 'forgetting' some old responses when it is being trained to new stimuli. Such a device, if it had in addition a past record of previous stimuli, would be able to avoid these drawbacks. It could

train on these stored stimuli, and when new stimuli were presented, could continue to check these against its store of old stimuli.   If, in fact the store contains every stimuli and the appropriate response, the operant conditioning machine becomes redundant, however many, if not most stimulus - response situations of the type described at the beginning of this chapter need not have any negative or positive reinforcement accompanying them.   *They are initially neutral and may only later become associated with some reinforcement.*   Here again we have a loose behavioural reason for supposing that the cortex, if it forms such models, should be neutral with respect to behavioural.or hypothalamic 'rewards'.

In connection with this, it has frequently been observed that only higher animals can perform well in delayed conditioning tasks, and it is reasonable to interpret 'higher' as meaning the presence of neocortex.   Furthermore, the destruction of neocortex, especially the frontal lobes,[5,6] can lead to the impairment of this ability.

In phylogenetic terms the neocortex is an outgrowth of the paleocortex which was originally a part of the brain associated with smell.[7]   This is perhaps the most interesting of senses in that it can be used for information about both distant and close objects.   Gregory[8] has pointed out that food

disappears from sight when it is taken in and that touch is
only effective when an object is close.   Smell and taste
together form a sense which can detect objects which secrete
chemically (and most food is of this form) whether they are
distant or have been taken in.   To indulge in teleology, a
group of cells receiving information of this nature would be
an ideal starting point for the outgrowth of a new structure
which could form predictive models.   In a general way this
also circumvents a problem posed by Gregory: how did the eye
develop when the brain did not have the 'computing power' to
deal with the information it would receive?   And why did the
brain develop such computing power when, if there was no well
developed eye, it had no data on which to compute?   The fact
that the neocortex, which includes the visual cortex, is an
extension of the rhinencephalon,   that part of the nervous
system associated with smell, shows how the necessary inter-
mediate development took place.

Such requirements of continuity in evolution also puts
up another obstacle for holographic theories.   How can a few
cells store a useful fourier transform or set of interference
fringes?   On the other hand, what sort of mutation could have
suddenly taken place in order to effect a working holograph in
the brain?

Having proposed this function for the neocortex that
it contains an internal 'world' of predictive models, what
can we say about Homunculus who explores this world?   He is
clearly very complicated, but he is not quite as complicated
as the nervous system he inhabits, for he himself does not
have the ability to make these models.   Homunculus no more
represents the individual that contains him, than a computer
without its core store represents the whole machine.   We have
suggested that it is possible to operant condition him and he
may be slightly frog-like (the latter animal having no neo-
cortex).   The point is that he is nevertheless complicated, and
a full understanding of the physiology of the neocortex may be
attained very much sooner than that of lower and phylogenetically
older structures.

REFERENCES

CHAPTERS 1 - 6

1.   GABOR, D. Nature, Vol. 161, p. 777. 1948

2.   GABOR, D. Proc. Roy. Soc. A. Vol. 197, p. 454. 1949

3.   LEITH, E.N. and UPATNIEKS, J. J. Opt. Soc. America,
         Vol. 52, p. 1123. 1962

4.   VAN HEERDEN, P.J. Applied Optics, Vol. 2, p. 393. 1963

5.   LIGHTHILL, M.J. An Introduction to Fourier Analysis and
         Generalised Functions. 1958. C.U.P.

6.   WILLSHAW, D.J., BUNEMAN, O.P. and LONGUET-HIGGINS, H.C.
         Nature, Vol. 222, p. 960. 1969

7.   MINSKY, M. and PAPERT, S. Perceptrons. 1969. M.I.T. Press.

8.   BUNEMAN, O.P. Machine Intelligence 5 (eds. B. Meltzer
         and D. Michie) p. 383, 1970. Edinburgh University Press.

9.   ROSENBLATT, F. Perceptrons. 1962. Spartan Books.

10. NILLSON, N. Learning Machines. 1965. McGraw Hill.

11. PRIBRAM, K.H. Macromolecules and Behaviour (J. Gaito ed.)
         1966. Appleton - Century - Crofts.

REFERENCES Cont'd.

CHAPTER 7

1. RASHEVSKY, N. Mathematical Biophysics. 1938. Chicago
    University Press.

2. HOUSEHOLDER, A.S. and LANDAHL, H.D. Mathematical Bio-
    physics of the Central Nervous System. 1945.
    Principia Press.

3. CRAGG, B.G. J. Anat. Vol. 101, p. 639. 1967.

4. HEBB, D,O. The Organisation of Behaviour. 1949. Wiley.

5. LASHLEY, K.S. Brain Mechanisms and Intelligence: A
    Quantitative Study of Injuries to the Brain. 1929.
    Chicago University Press.

6. MORRELL, F. Brain Mechanisms and Learning. (ed. J.F.
    Delafresnays) p. 375, 1961. Blackwell.

7. BINDMAN, L.J., LIPPOLD, O.C.J. and REDFEARN, J.W.T.
    Nature, Vol. 196, p. 534, 1962.

8. GARTSIDE, I.M. and LIPPOLD, O.C.J. J. Physiol. Vol. 139,
    p. 475, 1967.

9. BLISS, T.V.P., BURNS, B.D. and UTTLEY, A.M. J. Physiol.
    Vol. 195, p. 339, 1968.

REFERENCES Cont'd

CHAPTER 7 Cont'd

10.   KANDEL, E.R. and TAUC, L.   J. Physiol. Vol. 131, p. 1, 1965.

11.   KANDEL, E.R. and TAUC, L.   J. Physiol. Vol. 131, p. 28, 1965.

12.   YOUNG, J.Z.   A Model for the Brain.   1964.   Oxford
       University Press.

13.   CRAGG, B.G.   Proc. Roy. Soc.   Vol. 171, p. 319, 1969.

14.   BROWN, G.L. and PASCOE, J.E.   J. Physiol, Vol. 123,
       p. 565, 1954.

15.   ECCLES, J.C.   Physiology of Synapses.   1964.   Springer-
       Verlag.

16.   PITTS, W. and McCULLOCH, W.S.   Bull. Math. Biophysics,
       Vol. 9, p. 127, 1947.

17.   BRINDLEY, G.S.   Proc. Roy. Soc. B.   Vol. 168, p. 361, 1967.

18.   GRAY, E.G.   Nature, Vol. 193, p. 82, 1962.

REFERENCES Cont'd.

CHAPTER 8

1. GRAY, E.G.   J. Anat.   Vol. 93, p. 420, 1959.

2. ECCLES, J.C., ITO, M. and SZENTOGOTHAI, J.   The
        Cerebellum as a Neuronal Machine. 1967. Springer-
        Verlag.

3. DIAMOND, J.   J. Physiol.   Vol. 194, p. 669, 1968.

4. RALL, W.   Neural Networks (ed. E. Caioniello) p. 1, 1968.

5. GRIFFITHS, J.S.   A view of the Brain. 1967. O.U.P.

6. TAYLOR, W.K.   Proc. Roy. Soc. B.   Vol. 159, p. 466, 1964.

7. GRAY, E.G.   Nature, Vol. 193, p. 82, 1962.

8. PAPPAS, G.D., COHEN, E.B. and PURPURA, D.P. (eds.
        D.P. Purpura and M.D. Yahr) The Thalamus. p. 47,
        1966. Columbia University Press.

9. KANDEL, E.R. and SPENCER, W.A.   Physiol. Rev.   Vol. 43,
        p. 65, 1968.

10. LOUCKS, R.B.   J. Comp. Psychol.   Vol. 16, p. 439, 1933.

11. GIURGEA, C.   Studii Cercetari Fiziol. Neurol.   Vol. 4,
        p. 41, 1959.

12. BURES, J. and BURESOVA, O.   Proc. 23rd Inter Congr. Physiol.

REFERENCES Cont'd.

CHAPTER 8 Cont'd.

13.   DELGADO, J.M.R., ROBERTS, W.W. and MILLER, W.E.   Am. J.

Physiol.   Vol. 179, p. 537, 1954.

14.   OLDS, J.   Proc. 23rd Int. Congr. Physiol. Sci., p. 372, 1965.

15.   OLDS, J. and OLDS, M.E.   Brain Mechanisms and Learning.

(ed. J.F. Delafresnaye) p. 153, 1961.   Blackwell.

16.   OLDS, J. and MILNER, P.   J. Comp. Physiol.   Vol. 47,

p. 419, 1954.

17.   ANDERSEN, P.O.   Brain and Concious Experience.   (ed. J.C.

Eccles)   p. 59, 1966.   Springer-Verlag.

18.   YOUNG, J.Z.   A Model for the Brain.   1964.   O.U.P.

REFERENCES Cont'd.

CHAPTER 9

1.   SHOLL, D.A.   The Organisation of the Cerebral Cortex.
          1956.   Methuen.

2.   BEURLE, R.L.   Phil. Trans. Roy. Soc. B.   Vol. 240, p. 155,
          1956.

3.   VAN HEERDEN, P. J. Applied Optics.   Vol. 2, p. 393, 1963.

4.   SPERRY, R.W.   J. Neurophysiol.   Vol. 12, p. 325, 1947.

5.   LASHLEY, K.S.   J. Comp. Neurol.   Vol. 80, p. 257, 1944.

6.   ECCLES, J.C.   Physiology of Synapses.   1964.   Springer-
          Verlag.

7.   PENFIELD, W. and MILNER, B.   A.M.A. Arch. Neurol. Psychiat.
          Vol. 79, p. 475, 1958.

8.   BLISS, T.V.P., BURNS, B.D. and UTTLEY, A.M.   J. Physiol.
          Vol. 195, p. 339, 1968.

9.   GRAY, E.G.   Nature, Vol. 183, p. 1592, 1959.

10.   LORENTE, DE NO.   Chap. 15 in Physiology of the Nervous
          System by J.F. Fulton, 1943, O.U.P.

11.   HUBEL, D.H. and WIESEL, T.N.   J. Physiol.   Vol. 171.
          p. 476, 1964.

REFERENCES Cont'd.

CHAPTER 9 Cont'd.

12.   EVANS, E.F. and WHITFIELD, I.C.   J. Physiol. Vol. 171,

          p. 476, 1964.

13.   PETTIGREW, J.D., NIKARA, T. and BISHOP, P.D.   Exp. Brain

          Res.   Vol. 6, p. 353, 1963.

14.   HUBEL, D.H. and WIESEL, T.N.   J. Physiol.   Vol. 206,

          p. 419, 1970.

15.   ZEEMAN, E.C.   Mathematics and Computer Science in Biology

          and Medicine.   p. 277, 1965.   H.M.S.O.

16.   UTTLEY, A.M.,   Proc. 1st Int. Congr. on Cybernetics.   1965.

          Gouthier Villars.

17.   YOUNG, J.Z.   A Model for the Brain.   1964.   O.U.P.

REFERENCES Cont'd.

CHAPTER 10

1.   HUBEL, D.H. and WIESEL, T.N.   J. Physiol., 195,
        p. 339, 1968.

2.   ISARD, S.D. and LONGUET-HIGGINS, H.C.   To appear in
        Machine Intelligence 6.   Edinburgh University Press.

3.   RASTALL, J.   Internal Report, Research Memorandum,
        No. 62.   Department of Machine Intelligence and
        Perception, University of Edinburgh, 1969.

4.   UNGER, S.H.   Communications A.C.M., 7, p. 1, 1964.

5.   ROSSVOLD, H.E. and SWARBART, M.K.   The Frontal
        Granular Cortex and Behaviour, p. 1.   McGraw Hill,
        1964.

6.   MILNER, B.   The Frontal Granular Cortex and Behaviour,
        p. 313, McGraw Hill, 1964.

7.   PAPEZ, J.W.   Comparative Neurology, 1929, Hafner.

8.   GREGORY, R.L.   Proc. Roy. Soc. B., 171, p. 279, 1969.