1962

Analysis of a Four-Layer Series-Coupled Perceptron. II

H. D. Block
Bruce Knight
F. Rosenblatt

Follow this and additional works at: http://digitalcommons.rockefeller.edu/knight_laboratory

Part of the Life Sciences Commons

Recommended Citation
Block, H. D., Knight, B. W., Jr., Rosenblatt, F. (1962) Analysis of a four-layer series-coupled perceptron. II. Rev. Mod. Phys. 34: 135

This Article is brought to you for free and open access by the Laboratories and Research at Digital Commons @ RU. It has been accepted for inclusion in Knight Laboratory by an authorized administrator of Digital Commons @ RU. For more information, please contact mcsweej@mail.rockefeller.edu.
simple perceptron of Fig. 4 is no longer adequate. We shall find certain temporal effects in the paper which follows, but for others it is necessary to introduce time delays into the system.\(^1\) A speech recognizing perceptron which utilizes such delays is currently being built at Cornell University.

Other activities now in progress\(^3\) include quantitative studies of cross-coupled and multi-layer systems (by means of analysis and digital simulation), studies of selective attention mechanisms, the effects of geometric constraints on network organization, new types of reinforcement rules, and attempts at relating this research to biological data. Work is also in progress on development of electrolytic and other low-cost integrating devices and additional electronic components necessary for the construction of large-scale physical models.

It is clear that we are still far from the point of understanding how the brain functions. It is equally clear, we believe, that a promising road is open for further investigation.

---

**Analysis of a Four-Layer Series-Coupled Perceptron. II**

H. D. Block, B. W. Knight, Jr., and F. Rosenblatt

*Cornell University, Ithaca, New York*

1. INTRODUCTION

The preceding paper\(^1\) presented motivation and background for the general subject of perceptrons and gave some analysis and results for a simple three-layer perceptron. While it has been shown there that it is possible to associate any arbitrary set of responses to an arbitrary set of stimuli in a simple three-layer perceptron, such a perceptron characteristically requires a large representative sample of each kind of pattern (e.g., letters "A" and "B"), covering all parts of the retina, before it will recognize an arbitrarily positioned stimulus which is similar to one which it has seen before. In other words, a three-layer perceptron has no concept of "similarity" based on any criterion other than the intersections of sets of retinal elements. In a previous paper,\(^2\) Rosenblatt has shown that a "cross-coupled perceptron," in which \(A\) units are connected to one another by modifiable connections, should tend to develop an improved similarity criterion for generalizing responses from one stimulus to another when exposed to a suitably organized environment. In this paper a simpler network, consisting of four layers of units but without cross coupling, is analyzed in a more rigorous fashion, and is shown to possess the same property.

The perceptron of the present paper is "self-organizing" in the sense that during the training period the experimenter does not tell the machine the category of each stimulus. As the analysis below will show, the only contact between the experimenter and the machine is the presentation of the stimuli.

2. THE MODEL

The model to be analyzed here is a four-layer perceptron of the schematic type \(S\rightarrow A^1\rightarrow A^II\rightarrow R\), as indicated in Fig. 1.

---

\(^{1}\) Research sponsored by the Office of Naval Research.


There are \( n \) possible stimuli to be presented to the sensory field: \( S_1, S_2, \ldots, S_n \). The values of the \( S \) to \( A^1 \) connections do not change with time. The \( A^{11} \) units are in one-to-one correspondence with the \( A^1 \) units and have threshold \( \theta \). An active \( A^1 \) unit \( a^1 \) delivers a signal of \( \theta \) to its corresponding \( A^{11} \) unit \( a^{11} \) and also a signal \( v_{\mu r} \) to \( a^{11} \) for \( r = 1, 2, \ldots, N_\mu \) [see Eq. (1) below]. An inactive unit puts out no signal. We shall use the indices \( i, j, k \) throughout to designate various stimuli, while the indices \( \mu \) and \( r \) will be used to designate associator units. The values of \( v_{\mu r} \) are initially zero and change with time as follows. Stimuli are presented at times \( 0, \Delta t, 2\Delta t, 3\Delta t, \ldots \). If \( a^1 \) is active at time \( t \) and \( a^{11} \) is active at time \( t + \Delta t \), then \( v_{\mu r} \) receives an increment \((\eta \cdot \Delta \Delta)\); otherwise it does not receive this increment. At the same time each \( v_{\mu r} \) is decremented by \((\delta \cdot \Delta \Delta)\). [See Eq. (5), below.] These two effects represent a facilitation of used pathways and a decay, respectively. The \( A^{11} \) units are connected to response units with connections whose weights may be varied according to one of the standard rules of reinforcement. There is no time delay of transmission of signals through the system.

3. ANALYSIS

From the viewpoint of the \( R \) units we have a simple three-layer perceptron as described in the preceding paper with \( A^{11} \) as the associator set. Since the behavior of such a system is well understood in terms of the sets of associators in \( A^{11} \) activated by the various stimuli, our main concern here is to find the nature of these sets.

The set of associators in \( A^1 \) responding to \( S_t \) is denoted by \( A^1(S_t) \); the set responding to both \( S_t \) and \( S_j \) is \( A^1(S_t) \cap A^1(S_j) \). The number of associators in \( A^1 \) responding to both \( S_t \) and \( S_j \) is denoted by \( n_{ij} \) and is equal to the number of elements in \( A^1(S_t) \cap A^1(S_j) \): 

\[
n_{ij} = \sum_{\mu} e_{\mu i} e_{\mu j},
\]

where

\[
e_{\mu i} = \begin{cases} 1 & \text{if } a^1_\mu \in A^1(S_t) \\ 0 & \text{if } a^1_\mu \not\in A^1(S_t) \end{cases}
\]

None of these quantities change in time.

Let \( a^{11}_\mu (t) \) denote the total input signal to association unit \( a^{11}_\mu \) at time \( t \), if stimulus \( S_t \) were to be applied to the sensory field at time \( t \). Then

\[
a^{11} \mu (t) = \theta e_{\mu i} + \sum_{\mu} v_{\mu i}(t) e_{\mu j}.
\]

Let

\[
\beta_{ij} = \theta e_{ri} + \gamma_{ir}(t),
\]

and

\[
\gamma_{ij}(t) = \sum_{\mu} v_{\mu j}(t) e_{\mu i}.
\]

Then

\[
a^{11} \mu (t) = \beta_{ij} + \gamma_{ij}(t).
\]

Note that \( \beta_{ij} \) is \( \theta \) or 0 according as \( a^1 \) is in \( A^1(S_t) \) or not; it does not change with time. On the other hand \( \gamma_{ij}(t) \) represents the effect of the variable \( A^1 \) to \( A^{11} \) connections whose values are \( v_{\mu i}(t) \).

Suppose that at time \( t_0 \) stimulus \( S_t \) is presented and at time \( t_0 + \Delta t \) stimulus \( S_j \) is presented. Then the consequent change in \( v_{\mu i}(t) \) will be

\[
v_{\mu i}(t_0 + 2\Delta t) - v_{\mu i}(t_0 + \Delta t) = (\eta \cdot \Delta \Delta)(\gamma_{ij}(t_0 + \Delta t) - (\delta \cdot \Delta \Delta) v_{\mu i}(t_0 + \Delta t),
\]

where

\[
\phi(x) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta. \end{cases}
\]

From (3) and (5) we get

\[
\gamma_{ij}(t_0 + 2\Delta t) - \gamma_{ij}(t_0 + \Delta t) = \sum_{\mu} [v_{\mu i}(t_0 + 2\Delta t) - v_{\mu i}(t_0 + \Delta t)] e_{\mu i},
\]

\[
= (\eta \cdot \Delta \Delta) \phi(\alpha^{11}_{ij}(t_0 + \Delta t))
\]

\[
\times \sum_{\mu} e_{\mu i} e_{\mu j} - (\delta \cdot \Delta \Delta) \gamma_{ij}(t_0 + \Delta t),
\]

Hence

\[
\gamma_{ij}(t_0 + 2\Delta t) - \gamma_{ij}(t_0 + \Delta t) = (\eta \cdot \Delta \Delta) \phi(\alpha^{11}_{ij}(t_0 + \Delta t)) n_{ij} e_{\mu i} - (\delta \cdot \Delta \Delta) \gamma_{ij}(t_0 + \Delta t),
\]

where, for brevity, we have dropped the subscript \( r \), and will remember that the \( \gamma \) and \( a \) refer to any particular associator \( a^{11}_\mu \). Now suppose the sequence of stimuli \( S_{j_1}, S_{j_2}, \ldots, S_{j_M} \) is presented at the successive times \( t_1, t_2, \ldots, t_m + M \cdot \Delta t \). In Eq. (6) we take \( t = t_m + \Delta t \) and obtain

\[
\gamma_{ij}[t + (m + 2)\Delta t] - \gamma_{ij}[t + (m + 1)\Delta t] = (\eta \cdot \Delta \Delta) \phi(\alpha^{11}_{ij}(t + (m + 1)\Delta t)) n_{ij} e_{\mu i} - (\delta \cdot \Delta \Delta) \gamma_{ij}[t + (m + 1)\Delta t].
\]

Summing on \( m \) from 0 to \( M - 1 \), we get the change in \( \gamma_{ij}(t) \) due to the entire sequence of stimuli:

\[
\gamma_{ij}[t + (M + 1)\Delta t] - \gamma_{ij}[t + \Delta t] = \sum_{m=0}^{M-1} [(\eta \cdot \Delta \Delta) \phi(\alpha^{11}_{ij}(t + (m + 1)\Delta t)) n_{ij} e_{\mu i} - (\delta \cdot \Delta \Delta) \gamma_{ij}(t + (m + 1)\Delta t)].
\]

We divide by \( M \Delta t \) and let \( \Delta t \) approach zero to obtain

\[
\frac{d\gamma_{ij}}{dt} = \sum_{m=0}^{M-1} \frac{1}{M} \phi(\alpha^{11}_{ij}(t + (m + 1)\Delta t)) n_{ij} e_{\mu i} - \delta \gamma_{ij}(t).
\]

Let \( F_{jk} \) be the number of times the pair \( S_j S_k \) occurs in the given sequence \( S_{j_1}, S_{j_2}, \ldots, S_{j_M} \); and let \( f_{jk} = F_{jk} / M \) be the average frequency of the pair \( S_j S_k \). Then from (9) we get

\[
\frac{d\gamma_{ij}}{dt} = \sum_{k} \eta_{jk} \phi(\alpha^{11}_{ij}(t)) n_{ij} e_{\mu i} - \delta \gamma_{ij}(t).
\]

Defining the matrix

\[
K_{ij} = \sum_{k=1}^{n} n_{ik} f_{kj}
\]
we have from (10)

\[
\frac{d\gamma^{(i)}}{dt} = \eta \sum_j K_{ij} \left[ \beta_j + \gamma^{(i)}(t) \right] - \delta \gamma^{(i)}(t). \tag{11}
\]

This is a system of nonlinear differential equations for \( \gamma^{(i)}(t), \cdots, \gamma^{(n)}(t) \), with initial conditions \( \gamma^{(i)}(0) = 0 \).

If the \( f_{ij} \) vary with \( t \), then \( K_{ij} \) are time dependent but in any case they are non-negative and bounded; \( \phi \) is non-negative, monotonically increasing in \( \gamma \), bounded and continuous on the right. We shall treat here only the case in which the \( K_{ij} \) are constants.

Before discussing the solution of (11) we consider the equilibrium equation

\[
\gamma^{(i)} = \frac{\eta}{\delta} \sum_j K_{ij} \left[ \beta_j + \gamma^{(i)} \right]. \tag{12}
\]

The system of equations (12) may have more than one solution. However we shall show that there is a unique minimal solution (by this we mean a solution, none of whose components \( \gamma^{(i)} \) exceed the corresponding components of another solution); and this minimal solution is obtained in a finite number (at most \( n \)) of iterations of (12), starting with all \( \gamma^{(i)} = 0 \) on the right side and then finding the new values of \( \gamma^{(i)} \) from (12), putting these back into the right side and so on. That is, we take \( \gamma^{(0)} = 0 \) and

\[
\gamma^{(i+1)} = \frac{\eta}{\delta} \sum_j K_{ij} \left[ \beta_j + \gamma^{(i)} \right]. \tag{13}
\]

We prove first that the process terminates in at most \( n \) iterations. This can be seen from the following considerations. Since the right-hand side of (13) is non-negative and \( \gamma^{(i)} = 0 \), it follows that \( \gamma^{(i)} \geq \gamma^{(0)} \).

Now since the right-hand side of (13) is a non-decreasing function of the \( \gamma \)'s, it follows that

\[
\gamma^{(i)} \geq \gamma^{(0)} \geq \gamma^{(1)}, \cdots, \gamma^{(n+1)} \geq \gamma^{(n)}.
\]

Therefore also

\[
\phi[\beta_j + \gamma^{(n+1)}] \geq \phi[\beta_j + \gamma^{(n)}],
\]

that is, successive \( \phi \)'s cannot decrease. If, at a particular step no \( \phi \) increases then we are at a solution. The \( \phi \)'s have only the values zero or one, so even if only a single \( \phi \) changes at each step, the process terminates in at most \( n \) steps.

The solution of (12) thus obtained we denote by \( \gamma^{(0)} \). We now prove that this solution is minimal. Let \( \tilde{\gamma}^{(i)} \) be any solution of the equilibrium equation (12). Clearly \( \tilde{\gamma}^{(i)} \geq 0 \). Then for the iteration process (13), we have \( \gamma^{(i)} \leq \tilde{\gamma}^{(i)} \), for all \( i \). Since the right-hand side of (13) is a monotone function of \( \gamma \), we have

\[
\gamma^{(i)} = \frac{\eta}{\delta} \sum_j K_{ij} \left[ \beta_j + \gamma^{(i)} \right]
\]

\[
\leq \frac{\eta}{\delta} \sum_j K_{ij} \left[ \beta_j + \tilde{\gamma}^{(i)} \right] = \tilde{\gamma}^{(i)}.
\]

Similarly, \( \gamma^{(m)} \leq \tilde{\gamma}^{(m)} \), hence \( \gamma^{(0)} \leq \tilde{\gamma}^{(0)} \). Hence \( \gamma^{(0)} \) is minimal.

To avoid consideration of a special pathological case, we make a mild assumption. Consider the sum

\[
(\eta/\delta) \sum_{j \in S} K_{ij}
\]

taken over a subset \( S \) of the possible values of \( j \), \( 1,2,\cdots,n \). We assume that no such sum is equal to \( \theta \). This is not a serious assumption, since by an arbitrarily small change in \( \eta/\delta \) we can satisfy this requirement.

Now suppose that the \( \gamma^{(i)}(t) \) satisfy the system of differential equations (11) and the initial conditions \( \gamma^{(i)}(0) = 0 \). Then we assert that the \( \gamma^{(i)}(t) \) are non-decreasing and \( \lim_{t \to \infty} \gamma^{(i)}(t) = \gamma^{(0)} \). That is, the solution obtained by the iterative process (13) is indeed the solution of the differential equation (11), with initial conditions zero in each case.

First we shall show that \( d\gamma^{(i)}/dt \geq 0 \). Moreover,

\[
\text{if } \gamma^{(i)}(t) > 0 \text{ then } d\gamma^{(i)}(t)/dt > 0. \tag{A}
\]

As a preliminary step, consider the nature of the solution of the equation \( dx/dt = M - \delta x \), where \( M \) and \( \delta \) are positive constants and \( x(0) = a \), where \( 0 \leq a < M/\delta \). The solution \( x = (M/\delta) - e^{-\delta t}(M/\delta) - a \), has the appearance of Fig. 2. The solution approaches \( M/\delta \) monotonically from below, and \( dx/dt > 0 \) for all \( t > 0 \). If at time \( t = t_1 \) we replace \( M \) by \( M_1 > M \) the solution appears as Fig. 3.

As \( t \) goes from 0 to \( t_1 \) the solution approaches \( M/\delta \) monotonically from below; as \( t \) increases beyond \( t_1 \) the solution approaches \( M_1/\delta \) monotonically from below.

The solution is continuous; so it is derivative, except at \( t_1 \), where the left- and right-hand derivatives are not equal; but both are positive.
If instead of $M > a \beta_0 \geq 0$, we take $M = a = 0$, the solution $x(t) = 0$ for $0 \leq t \leq t_1$.

We now proceed to the proof of (A). Let

$$M^{(i)}(t) = \eta \sum_j K_{ij} \phi[\beta_j + y^{(j)}(t)].$$

Then (11) can be written

$$\frac{dM^{(i)}}{dt} = M^{(i)}(t) - \delta y^{(i)}(t),$$

where here and in the following paragraph, $i$ is a generic index of the set $(1, 2, \cdots, n)$, while $j$ and $k$ will refer to specific indices to be defined below.

Each function $M^{(i)}(t)$ can take on at most $2^n$ possible values. Let $k$ be a specific value of $i$ and suppose first that $M^{(k)}(0) > 0$. The only times at which $M^{(k)}(t)$ can change its value are when one of the $y^{(j)}$ (indeed one whose corresponding $\beta_j = 0$) reaches the value $\theta$. Suppose the first time at which this happens is $t_1 > 0$.

Suppose then that $y^{(k)}(t_1) > 0$. Since in the interval $0 < t < t_1$ all $\frac{dM^{(i)}}{dt} \geq 0$, we have $M^{(k)}(t_1) > M^{(k)}(0)$. Thus the solution $y^{(k)}(t)$ appears as in Fig. 3; in particular for all $k < \epsilon$ such that $M^{(k)}(0) > 0$, we have $\gamma^{(k)}(t_1) < M^{(k)}(0) \delta \leq M^{(k)}(t_1) / \delta$; and for the others $\gamma^{(k)}(t_1) \leq M^{(k)}(t_1) / \delta$. Furthermore, since both the left and right derivatives of $\gamma^{(k)}$ at $t_1$ are positive we have, for $t > t_1$ and sufficiently close to $t_1$, $\gamma^{(k)}(t) = \theta$, so that it will not be until time $t_2$, with $t_1 < t_2$, that there will again be a $\gamma^{(k)}(t)$ having the value $\theta$. In the interval $t_1 < t < t_2$ we have the same pertinent conditions as we had in the interval $0 < t < t_1$; namely $\frac{dM^{(i)}}{dt} = M^{(i)}(t_1) - \delta y^{(i)}(t)$, with initial values $y^{(i)}(t_1) \leq M^{(i)}(t_1) / \delta$ and in particular $\gamma^{(k)}(t_1) < M^{(k)}(t_1) / \delta$. Thus in the interval $(t_1 < t < t_2)$ we again have $\frac{dM^{(i)}}{dt} \geq 0$, and $\frac{dM^{(k)}}{dt} > 0$.

We repeat the same argument for the successive intervals $(t_2, t_3), (t_3, t_4)$ and so on. Since the $y^{(i)}(t)$ are monotone there are at most $n$ such intervals.

If $M^{(k)}(0) = 0$ then $y^{(k)}(t) = 0$ for $0 \leq t \leq t_1$. If $M^{(k)}(t_1) > 0$, then we use the previous argument starting at $t_1$; otherwise $y^{(k)}(t)$ remains zero at least until $t_2$ and so on. In any case we have proved (A).

Next we show that

$$\lim_{t \to \infty} y^{(i)}(t) = \gamma^{(i)*}; \quad i = 1, 2, \cdots, n. \quad (B)$$

Since, from the proof of (A) it is clear that each $y^{(k)}(t)$ is monotone and bounded, $\lim_{t \to \infty} y^{(k)}(t)$ exists; call it $y^{(k)*}$; it is a sum of the form

$$(\eta/\delta) \sum_{j \in \mathbb{S}} K_{ij}$$

(which was assumed at the outset to be unequal to $\theta$) and thus $\phi[\beta_j + y^{(j)}(t)]$ is continuous when $\gamma^{(j)} = y^{(j)*}$. Letting $t \to \infty$ in Eq. (11) we see that $\gamma^{(j)*}$ is a solution of the equilibrium equation (12).

Hence $\gamma^{(k)*} \geq \gamma^{(k)}$, since $\gamma^{(k)}$ is minimal. We next show that for all $t \geq 0$, $\gamma^{(k)}(t) \leq \gamma^{(k)*}$.

Note that initially $\gamma^{(k)}(0) \leq \gamma^{(k)*}$. Suppose that $t_1$ is the first time at which some $\gamma^{(k)}(t) = \gamma^{(k)*}$. From (11) and the fact that $\phi$ is nondecreasing we see that at $t_1$

$$\frac{dy^{(k)}}{dt} = \eta \sum_j K_{ij} \phi[\beta_j + \gamma^{(j)}(t)] - \delta y^{(k)}(t) \leq \eta \sum_j K_{ij} \phi[\beta_j + \gamma^{(j)*}] - \delta y^{(k)}(t) = \delta y^{(k)*} - \delta y^{(k)}(t) = 0,$$

i.e., $\frac{dy^{(k)}}{dt} \leq 0$ at $t \leq t_1$. If $\gamma^{(k)*} > 0$, we see from (A) that $\frac{dy^{(k)}}{dt} > 0$ at $t_1$, a contradiction. Suppose that $\gamma^{(k)*} = 0$, so that also $t_1 = 0$. Then, as long as no $\gamma^{(k)}(t)$ reaches a nonzero $\gamma^{(k)*}$, we have, since $\phi$ is a nondecreasing function of its argument,

$$M^{(k)}(t) = \eta \sum_j K_{ij} \phi[\beta_j + y^{(j)}(t)] \leq \eta \sum_j K_{ij} \phi[\beta_j + \gamma^{(j)*}] = \delta y^{(k)*} = 0.$$

Hence over this period $\gamma^{(k)}(t) = 0$. But no nonzero $\gamma^{(k)*}$ can ever be attained by $\gamma^{(k)}(t)$, since, by the above argument, we would have $\frac{dy^{(k)}}{dt} \leq 0$ at the first time it occurred, in contradiction to (A).

Thus we have shown that: if $\gamma^{(k)} > 0$, then $\gamma^{(k)}(t) < \gamma^{(k)*}$; and if $\gamma^{(k)} = 0$, then $\gamma^{(k)}(t) = \gamma^{(k)*}$. In general $\gamma^{(k)}(t) \leq \gamma^{(k)*}$.

Hence $\gamma^{(k)} = \lim_{t \to \infty} y^{(k)}(t) \leq \gamma^{(k)*}$ and (B) follows.

From this point on we shall be concerned with the steady-state values $\gamma^{(k)*}$, and, for brevity, we shall drop the $*$. In the terminal condition the associator $a_{11}$ is activated by $S_1$ if $\beta_{11} + y^{(1)*} \geq \theta$. The set of $A_{11}$ associators which are activated by stimulus $S_1$ will be denoted by $A_{11}(S_1)$. In the initial state, the set $A_{11}(S_1)$ is denoted by $A_{11}(S_1)$, and in the terminal state by $A_{11}(S_1)$. The number of $A_{11}$ associators which are activated by both $S_1$ and $S_2$ is called $n_{11}$ and is equal to the number of units in $A_{11}(S_1) \cap A_{11}(S_2)$.

Once the $n_{11}$ are known, the behavior of the perceptron can be predicted, as outlined in the preceding paper. To determine the $n_{11}$ we might proceed as follows. First the set of $A_{11}$ associators is broken into the smallest elements of the Venn diagram representing the sets responding to different combinations of stimuli. Each of these cells is characterized by a certain $\beta$ vector. For each such $\beta$ vector we solve Eq. (12) for the terminal values of $\gamma^{(i)}$. Here we regard $n_{1k}$ and $f_{ij}$ as given. (In the present paper $n_{1k}$ represents the actual number of $A_{11}$ units responding to both $S_1$ and $S_2$ in the particular perceptron being studied. In general, if the $S$ to $A_{11}$ connections are chosen at random, subject to statistical constraints, then the $n_{1k}$ represent random variables whose expected values are $N_{Q_{1k}}$, where $Q_{1k}$ is the probability that an associator $a_{1k}$ is activated by both $S_1$ and $S_2$. These random variables have been studied rather thoroughly for several main classes of networks.)

Initially, $n_{11} = n_{11}$. Knowing $\beta_{11}$ and $\gamma^{(1)}$ we can compute the region of the $A_{11}$ Venn diagram to which

---

each of the $A^{II}$ units moves. Then we have the complete
terminal distribution in the Venn diagram of $A^{II}$
and hence in particular the $n_{ij}^{II}$. While this program
is simple in principle the actual computations can be
quite formidable. It can be seen in advance that the
motion will be for associators to go into regions of
higher intersections but points which are initially
outside all the $A^{II}(S_i)$ will stay outside all the $A^{II}(S_i)$.

The analysis is, approximate, in course, that we have
replaced the difference equation (8) by the
differential equation (9). For small increments of reinforcement $\eta \Delta t$ and small increments of decay
$(\Delta t)^2$ at each step the behavior of the discretized
system will be approximated by that of the continuous
one.

In this way the performance of these perceptrons
can be analyzed for a variety of particular cases.\note{\textsuperscript{4}}
We shall instead consider here two "training programs"
in which, by a suitable choice of parameters, the
motion of the $\beta$ vectors can be controlled and the
performance predicted.

4. APPLICATIONS

Let $P_j$ denote the fraction of occurrences of $S_j$ in
the given sequence $S_j, S_{j1}, \ldots, S_{jm}$ and let $P_{jk}$ denote
number of times $S_k$ immediately follows $S_j$ divided by
the number of times $S_j$ occurs. Then in a long sequence
the equilibrium equation (12) takes the form

$$
\gamma^{(i)} = \frac{1}{\delta \Delta t} \left[ \sum_{j=1}^{K} \sum_{k=1}^{K} P_j P_{jk} \phi(\beta^{(k)} + \gamma^{(k)}) \right],
$$

where $P_j$ corresponds to the probability of $S_j$, and $P_{jk}$
corresponds to the transition probability

$$
S_j \rightarrow S_k = \text{Prob}(S_j S_k | S_j).
$$

Training Program I

The stimuli are divided into two classes:
$\{S_1, S_2, \ldots, S_K\}$ is class $X$, while $\{S_{K+1}, \ldots, S_n\}$ is
class $Y$. There is assumed to be no appreciable difference
in the retinal overlaps; we assume $n_{ij} = (q+s\delta_{ij})$,
where $s>0$, $q \geq 0$. Thus the diagonal elements of the
$n_{ij}$ matrix are all $(q+s)$ and all other elements are $q$.
(Note that by raising thresholds of the $A^I$ units, the
ratio $q/s$ can be made as small as desired.) We shall
assume that the probability of transition to a member of
the same class is $p$, nearly unity, and to a member of
the opposite class is $(1-p)$, nearly zero. Inside a
class all members are equally likely. Thus

$$
P_j = \begin{cases} 
1/2K & \text{for } S_j \text{ in } X \\
1/2L & \text{for } S_j \text{ in } Y
\end{cases}
$$

where $L=n-K$;

$$
P_{jk} = \begin{cases} 
p/K & \text{for } S_j \text{ in } X, S_k \text{ in } X, \\
(1-p)/L & \text{for } S_j \text{ in } X, S_k \text{ in } Y, \\
p/L & \text{for } S_j \text{ in } Y, S_k \text{ in } X, \\
(1-p)/K & \text{for } S_j \text{ in } Y, S_k \text{ in } X.
\end{cases}
$$

Then we get from (14)

$$
\gamma^{(i)} = \frac{1}{\delta \Delta t} \left[ \sum_{j=1}^{K} \sum_{k=1}^{K} P_j P_{jk} \phi(\beta^{(k)} + \gamma^{(k)}) \right] 
+ \frac{1}{2K} \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{q+s\delta_{jk}}{2L} \sum_{l=1}^{n} \phi(\beta^{(l)} + \gamma^{(l)})
+ \frac{1}{2L} \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{q+s\delta_{jk}}{2K} \sum_{l=1}^{n} \phi(\beta^{(l)} + \gamma^{(l)})
\right].
$$

Let us now assume that $S_{q}$ is one of the stimuli of
class $X$. Then

$$
\gamma^{(i)} = \frac{1}{\delta \Delta t} \left[ \frac{(q+s)P_j}{2K^2} \sum_{k=1}^{K} \phi(\beta^{(k)} + \gamma^{(k)}) 
+ \frac{(s(1-p)+q)P_j}{2KL} \sum_{k=1}^{K} \phi(\beta^{(k)} + \gamma^{(k)}) \right].
$$

We now observe the following
(a) If $\eta(sP+qK)/2K^2 \geq \delta$, then

$$
A^{II}(S_q) \supseteq \bigcup_{S_j \in X} A^{II}(S_j).
$$

In words, if the stated inequality holds, then,
in the terminal condition, each of the stimuli of
class $X$ activates the union of all sets which were
initially activated by any of the stimuli of class $X$.
[See Fig. 4(a).]

The proof follows from the fact that any associator
which originally responded to any of the stimuli in

\[\text{FIG. 4. Superimposed initial and terminal Venn diagrams of } A^{II}\text{layer, for two cases.}\]
class $X$ has $\beta^{(k)} = \theta$ for some $k \leq K$. Hence there is a nonzero term in $\sum_{k=1}^{K} (16)$. The postulated inequality then guarantees that the associator will be active in the terminal state.

(b) If $\eta s(1-p) + qK) \leq 2K \delta < \theta$, then

$$A_{\theta}^{(1)}(S_2) \subseteq \bigcup_{s_j \in X} A_{\theta}^{(1)}(S_j).$$

In words, if the stated inequality holds, then, in the terminal condition, no stimulus in class $X$ activates any associator outside of the union of sets initially activated by stimuli of class $X$. [See Fig. 4(b).]

The proof follows from the fact that, if we were to solve (16) by iteration, then any associator which is activated by none of the $X$ stimuli has, on the first pass, no contribution from the term

$$\sum_{k=1}^{K} \text{in Eq. (16)}.$$ 

Thus on the first iteration we get from (16)

$$\gamma^{(s)}(a) = \left(\frac{s}{L}\right) \sum_{k=1}^{K} \frac{\phi(\beta^{(k)} + \gamma^{(k)})}{2K \delta} \leq \left(\frac{s}{L}\right) \frac{\eta s(1-p) + qK) \leq 2K \delta < \theta}{2K} < \theta$$

in virtue of the assumed inequality (b). Similarly, on the next iteration, the term

$$\sum_{k=1}^{K} \text{ in Eq. (16)}$$

is again zero for such an associator and, as in (17), $\gamma^{(s)}(a) < \theta$. Since there are only a finite number (indeed less than $L$) iterations we have $\gamma^{(s)}(a) < \theta$ for such an associator and this associator does not become active.

(c) If the inequalities of (a) and (b) hold then

$$A_{\theta}^{(1)}(S_2) = \bigcup_{s_j \in X} A_{\theta}^{(1)}(S_j).$$

Necessary and sufficient conditions for both (a) and (b) to hold are easily found. They are

(i) $s > qK(K-1)$,
(ii) $p > [Kq + qK(K-1)]/s(K+1)$,
(iii) $K^2/s(p + qK) \leq \eta/206 < K/(s(1-p) + qK)$.

Condition (i) ensures that a $p$, $0 < p < 1$ can be chosen to satisfy (ii). Condition (ii) ensures that $\eta/206$ can be chosen to satisfy (iii). The conditions can be put in the alternative form

(i') $p > K/(K+1)$
(ii') $s > qK(K-1)/[p(K+1) - K]$

and (iii).

If the parameters satisfy the stated inequalities then

$$A_{\theta}^{(1)}(S_2) = \bigcup_{s_j \in X} A_{\theta}^{(1)}(S_j).$$

This means that in the terminal state each stimulus of class $X$ activates the same set of $A^{(1)}$ units. Similarly there is a second set of $A^{(1)}$ units activated by each member of class $Y$. If the $A^{(1)}$ to $R$ connections are random then, in general, one pattern of activated $R$ units will respond to all stimuli of class $X$ and another pattern of activated $R$ units will respond to all stimuli of class $Y$. Thus the machine has dichotomized the classes, its output being in terms of this intrinsic code on the $A^{(1)}$ units. Alternatively, with a single $R$ unit and zero initial values on the $A^{(1)}$ to $R$ connections, a single corrective reinforcement applied to one stimulus of each class will serve to establish the dichotomy, yielding the correct response for all the stimuli.

The problem in which the stimuli are separated into more than two categories can be analyzed in a similar manner.4

Training Program II

Consider the stimuli $S_1, S_2, \ldots, S_K$ and their transforms $S_{K+1} = T(S_1), S_{K+2} = T(S_2), \ldots, S_{2K} = T(S_K)$ under some one-to-one transformation $T$ of the retinal points. For example, $S_1, \ldots, S_K$ may be in the left half of the field and $T$ a transformation which moves them to the right half. $S_{2K}(x = 2K+1)$ is not shown during the training but is a test stimulus to be applied after the perceptron is trained. $S_y = T(S_y), y = 2K+2 = \mu$. Let us assume $S_x$ intersects $S_1, \ldots, S_K, L < K$ to a larger extent than does the others and hence $S_y$ intersects mostly with $S_{K+1}, \ldots, S_{2K+L}$. Specifically (cf. Fig. 5)

$$n_{xy}^1 = \begin{cases} (q + \delta_0) & j > L, \\ (q + r) & j \leq K, \end{cases}$$

$$n_{xj}^1 = \begin{cases} q & K+1 \leq j \leq K+L, \\ (q + \delta_0) & j > K+L. \end{cases}$$

We also assume that no associator is activated by more than $\mu$ of the stimuli $S_1, S_2, \ldots, S_K$, where $\mu < K/L$.

A stimulus $S_i$ from $(S_1, \ldots, S_K)$ is picked at random and the next stimulus is the transform $T(S_i)$. Then another is picked at random from $(S_1, \ldots, S_K)$ and this is followed by its transform, and so on.

Then

$$P_j = \begin{cases} 1/2K, & j \leq 2K, \\ 0, & j > 2K, \end{cases}$$

$$P_{jk} = \begin{cases} 1, & j \leq K, \ k = j + 1, \\ 1/K, & j > K, \ k \leq K, \\ 0, & \text{otherwise}. \end{cases}$$
From Eq. (14) we obtain

\[
\gamma'(x) = \frac{\eta}{2K\delta} \sum_{j=1}^{K} n_j \phi(\beta^{(k+i)} + \gamma^{(k+i)}) \\
+ \frac{1}{K} \sum_{j=K+1}^{2K} \sum_{k=1}^{K} n_{j,k} \phi(\beta^{(k+i)} + \gamma^{(k+i)})
\]

\[\text{(18)}\]

\[
\gamma'(x) = \frac{\eta}{2K\delta} \left[ q + r \sum_{j=1}^{L} \phi(\beta^{(k+i)} + \gamma^{(k+i)}) \\
+ q \sum_{j=L+1}^{K} \phi(\beta^{(k+i)} + \gamma^{(k+i)}) + \frac{q}{\phi(\beta^{(i)} + \gamma^{(i)})} \right].
\]

Hence if

(i) \( \eta(q+r)/2K\delta \geq \theta \)

then

\[ A_{x}^{11}(S_x) \supseteq A_{0}^{11}(S_x) + \bigcup_{j=1}^{L} A_{x}^{11}[T(S_j)]. \]

In words, if the stated inequality holds then, in the terminal state, \( S_x \) activates all those elements originally excited either by itself or by any of the transforms \( T(S_1), \ldots, T(S_L) \).

(ii) If \( \eta(K+\mu-L)/2K\delta < \theta \), then

\[ A_{x}^{11}(S_x) \supseteq A_{0}^{11}(S_x) + \bigcup_{j=1}^{L} A_{x}^{11}[T(S_j)]. \]

If both inequalities hold, then

\[ A_{x}^{11}(S_x) = A_{0}^{11}(S_x) + \bigcup_{j=1}^{L} A_{x}^{11}[T(S_j)]. \]

Consider next \( \gamma^{(p)} \). Equation (18) applies (with \( x \) replaced by \( y \)) and

\[
\gamma^{(p)} = \frac{\eta}{2K\delta} \sum_{j=1}^{K} \phi(\beta^{(k+i)} + \gamma^{(k+i)}) \\
+ \frac{1}{K} \sum_{j=K+1}^{2K} \sum_{k=1}^{K} n_{j,k} \phi(\beta^{(k+i)} + \gamma^{(k+i)}) \\
+ \frac{q}{\phi(\beta^{(i)} + \gamma^{(i)})}
\]

Thus if

(iii) \( \eta(q+K)/2K\delta + \frac{Lr+Kq}{K} < \theta \),

then \( A_{x}^{11}(S_y) = A_{0}^{11}(S_y) \). If all three inequalities hold then the stimulus \( S_x \) generalizes to \( T(S_y) \), but the transform \( T(S_y) = S_y \) does not generalize to the stimulus \( S_x \). Thus, when the machine has reached its terminal state the stimulus sequence \( S_x \) followed by \( S_y \) is characterized by a decreasing amount of activity, while the sequence \( S_y \) followed by \( S_x \) would yield an increasing pattern of activity. By connections having a time delay to the response units the machine can thus distinguish between a motion to the left (decreasing activity) and a motion to the right (increasing activity).

Of course the above asymmetry between \( [S, T(S)] \) and \( [T(S), S'] \) is due to the training method. If a symmetrical training is applied, the generalization goes both ways.\(^4\)

Necessary and sufficient conditions that all three inequalities hold are easily found, [with \( r \geq 0 \), condition (iii) implies (ii) and are:

\[
\begin{align*}
\text{(a)} & \quad r \geq \frac{K(K+\mu-1)}{(K-L\mu)}, \\
\text{(b)} & \quad \frac{2K}{q+r} \leq \frac{2K}{q+r} = \frac{Kq(K+\mu)+Lr\mu}{Kq(K+\mu)+Lr\mu}.
\end{align*}
\]

In particular let \( L = 1 \). Then \( A_{x}^{11}(S_x) = A_{0}^{11}(S_x) + A_{x}^{11}[T(S_1)], \) and \( A_{x}^{11}[T(S_2)] = A_{x}^{11}[T(S_2)]. \) Thus due to the intersection between \( A_{x}^{11}(S_x) \) and \( A_{x}^{11}[T(S_i)] \) the test stimulus \( S_x \) generalizes to its transform, even though neither one has occurred during the training sequence. This is the effect which was originally predicted for cross-coupled perceptrons,\(^2\) and has since been demonstrated in digital simulation experiments. Specifically, suppose that we have another test stimulus \( S_x \) analogous to \( S_x \) but its chief intersection is with \( S_y \) say also \( q+r \). Then, if (a) and (b) are satisfied (with \( L = 1 \) and \( \mu < K \)),

\[ A_{x}^{11}(S_x) = A_{0}^{11}(S_x) + A_{x}^{11}[T(S_2)]. \]

and

\[ A_{x}^{11}[T(S_y)] = A_{0}^{11}[T(S_y)]. \]

Suppose the perceptron has zero initial values on the \( A^{11} \) to \( R \) connections. Let it be shown \( S_x \) and let all active \( A^{11} \) to \( R \) connections be given a positive increment. Then let it be shown \( S_x \) and let all active \( A^{11} \) to \( R \) connections be given a negative increment. Now if the perceptron is shown \( T(S_x) \), a stimulus it has never seen before) the input to the response unit is the number of associators in

\[ A_{x}^{11}[T(S_x)] \cap \{A_{x}^{11}[T(S_1)] \cup A_{x}^{11}(S_x)\} \]

minus the number of associators in

\[ A_{x}^{11}[T(S_2)] \cap \{A_{x}^{11}[T(S_2)] \cup A_{x}^{11}(S_2)\}, \]

which, in general, is positive. Similarly if the machine is shown \( T(S_y) \) the response is negative. [Note that
the same result is obtained if the final training is done in the opposite order; that is, show \( T(S_o) \) and increment positively; then show \( T(S_o) \) and increment negatively; now ask for the responses to \( S_o \) and to \( S_r \).

Thus in the terminal state, after training with the transformation applied to unrelated stimuli, the machine, when taught the response to \( S_o \) and to \( S_r \), automatically gives the same response to \( T(S_o) \) as to \( S_o \) and to \( T(S_o) \) as to \( S_r \). It has learned to identify stimuli as equivalent under the transformation \( T \) (and similarly \( T^{-1} \)).

It has been noted\(^2\) that this ability to identify objects equivalent under a transformation is greatly enhanced by using random blobs, rather than completely random pepper and salt patterns during the initial training period. The reason for this is clear from the fact that the intersection of the pepper and salt patterns with the test letters are all approximately equal, so that \( r/q \) is small and condition (a) is violated.

5. CONCLUSIONS

We have shown that for the model described, autonomous learning is possible. In particular, from the analysis, the following types of performance are possible.

(a) The perceptron is shown a random sequence of letters of the alphabet, each letter occurring in various forms, fonts, and positions. The sequence is composed in such a way that a given letter “A,” is more likely to be followed by another form or position of the same letter, A, than by a different letter. Ultimately, the perceptron will have seen a number of “runs” of each letter of the alphabet, each such run consisting of a sample of possible positions and variations. At the end the machine should assign a distinctive response to any letter presented; one response for “A”’s another for “B”’s etc. Of course, the particular assignment of responses cannot be specified in advance, since at no time does the experimenter give the machine any instructions based on his knowledge of what the letters are; he merely shows it one letter at a time, distorting and transforming it. It is not the topological similarity of the “A”’s with each other, nor the point-set overlap that is crucial here, but rather the fact that the “A”’s occur contiguously in time. Thus any set of objects that occur contiguously in time can be classified separately from any other sets whose members have the same property.

(b) The machine is shown blobs which jump from the left to the right half of the retina. It is then shown an “A” and a “B” on the right and taught to give the response \( R_1 \) to “A” and \( R_2 \) to “B”. Then, shown an “A” on the left, it gives response \( R_1 \), and, for “B” on the left, \( R_2 \). It can also be designed so that on being shown any object moving to the right it gives one response; for all objects moving to the left the opposite.

6. CROSS COUPLED SYSTEMS

The more general cross-coupled systems of Fig. 2 of the previous paper\(^1\) present several additional complications.

(a) Closed-loop reverberations are possible. Activity can go on indefinitely, independent of the stimuli presented. The question of whether these reverberations die out, stabilize, or spread to activate all units is crucial to the design.

(b) The sets \( A^1(S) \) are no longer constant in time. This complicates the analysis. Moreover they depend on the sequence of stimuli preceding \( S \) rather than on only \( S \) itself. However with suitable modifications, an analysis analogous to that given here has been carried through and equations analogous to (12) obtained.\(^3\) The application of these results to particular problems is being studied.

Cross coupled systems with feedback from the \( R \) to \( A \) units are now being studied both analytically and by simulation, with the belief that such systems or related models might be capable of dealing with the problems of figure-ground discrimination, relations of figures in complex visual fields, selective attention and recall, and temporal pattern recognition.\(^3\)