JOURNAL OF MATHEMATICAL PSYCHOLOGY: 6, 209–239 (1969)

Embedding Fields: A Theory of Learning with Physiological Implications[1]

STEPHEN GROSSBERG

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

A learning theory in continuous time is derived herein from simple psychological postulates. The theory has an anatomical and neurophysiological interpretation in terms of nerve cell bodies, axons, synaptic knobs, membrane potentials, spiking frequencies, transmitter production and release, etc. In particular, a new hypothesis concerning transmitter production is presented. Backward learning, a connection between reaction times and learning speeds, learning without neural reverberation, and the variation through time of contexts in response to shifting environmental demands are discussed. Some qualitative connections with the work of Guthrie, Hull, Pavlov, and the Gestaltists are noted. Linear vs nonlinear, as well as Markovian vs non-Markovian properties of the theory's mathematical formalism are mentioned.

1. INTRODUCTION

Some recent papers (Grossberg, 1967, 1968a–e, 1969a–h) have introduced a new theory of learning in a rigorous setting. In its simplest form, this theory provides a mathematical description of the following kind of experiment. An experimenter E, confronted by a machine (or learning subject) M, presents M with a list of "letters" or "events" to be learned. Suppose, for example, that E wishes to teach M the list of letters AB, or to predict the event B given the event A. E does this by presenting A and then B to M several times at prescribed instants of time. To find out if M has learned the list as a result of these list presentations, the letter A alone is then presented to M. If M responds with the letter B, and M does this whenever A alone is said, then we have good evidence that M has indeed learned the list AB.

Our learning theory thus concerns itself with a description of the stimuli and responses of an individual subject through time. It is a *deterministic* theory, and not a statistical one.

Surely the construction of machines which "learn" in a sufficiently naive sense is not a difficult task. On the other hand, the machines which we have discussed can be derived from plausible psychological axioms, and once derived exhibit some interesting

[1] This work was supported in part by NONR Contract 1841/38.

properties of learning. For example, our simplest machine (Grossberg, 1967, 1968a, b) has, among others, the following properties:

(1) *Practice Makes Perfect.* The more often AB is practiced, the better is the machine's prediction of B given A at a prescribed later time, and the prediction becomes as good as we please after a sufficient amount of practice. One can modify the machine in a trivial way to guarantee that the learning of a short list such as AB seems to occur in an "all-or-none" fashion. Practice is by respondant conditioning.

(2) *An Isolated Machine Suffers No Memory Loss.* Once learning trials end, our simplest machine remembers what it has been taught without any memory loss. This is not true of all our machines. Some of them spontaneously forget at an approximately exponential rate if they do not practice continuously. These various machines all obey the same laws, however. They differ only in the way in which their several components are interconnected. We are led to a study of the "geometry of learning," namely, a study of how to interconnect the components of our machines to guarantee that they learn and remember special tasks in the best possible way.

(3) *An Isolated Machine Remembers Without Practicing Overtly.* After learning trials cease, our simplest machine also stops producing guesses for the experimenter. Even when the machine produces no overt behavior after learning, its memory of the preceding experiment remains unimpaired.

(4) *The Memory of an Isolated Machine Sometimes Improves Spontaneously Without Practice.* After the simplest machine receives a moderate amount of practice, and shortly after practice ceases, we find that its memory is better on a recall trial than it was at the instant practice stopped. The magnitude of this improvement depends on the degree to which practice is massed or distributed when the learning trials cease. This effect strikingly resembles the experimental phenomenon of "reminiscence," otherwise known as the Ward-Hovland phenomenon (Osgood, 1953).

(5) *All Errors Can Be Corrected.* If a list such as AB is learned to an arbitrary degree of accuracy, we can nonetheless teach the machine the new list AC.

(6) *Response Interference Sometimes Occurs.* The rate with a list AC can be taught to replace a previously learned list AB depends on the degree to which AB had been learned, as well as on the number of other response alternatives. However, this is not true of error correction in long lists. One can change a long list in its middle after the first few learning trials without substantially delaying the rate with which the new items are learned. The effect of other response alternatives also depends on list length, on list position, on the rate of list presentation, and on the degree of learning at any time.

These properties do not exhaust the list of mathematical effects which arise in our machines, and one can find formal analogs of such familiar empirical phenomena as

backward learning, bowing, chaining, and chunking (Jensen, 1962; Miller, 1956; Osgood, 1953). It is also possible to interpret the mathematical variables of the machines in a way that permits us to compare them with known neural facts. Geometrical objects exist in them that readily call to mind nerve cell bodies, axons, endbulbs, and synapses (Crosby, 1962). Processes occur within these objects that remind one of the generation of cellular potentials in cell bodies, of the fluctuation of spiking frequencies in axons, of transmitter production and release at the endbulbs, and of various trophic and plastic effects (De Robertis, 1964; Eccles, 1957, 1964).

Our machines, therefore, provide a single mathematical picture within which at least formal analogs of both psychological and neural phenomena of some interest can be discussed. All of these phenomena, or at least their formal analogs, are a consequence of a rather simple mathematical mechanism. Since these machines do learn and can, at least roughly, be interpreted in a neural way, they embody a definite proposal concerning the manner in which real neural structures might learn.

Because of these various facts, it seems desirable to try to analyze the psychological principles which give rise to these machines. This paper aims at such an analysis and, in particular, at a description that is as intuitive and nontechnical as possible to emphasize the simplicity of the basic ideas. We begin by discussing in a rather philosophical way some psychological facts known to all of us from daily life, and then gradually translate these facts into definite mathematical terms until we have explored enough facts to construct a well-defined mathematical system. We cannot, of course, hope by such a one-sweep procedure to "construct a brain," teeming with representations of countless macromolecules and ions interwined in exotic combinations of variable duration and strength. Nor should we want to, since such a representation would blind the unprepared beholder with complexities. Three later papers (Grossberg, 1968e, 1969b, c) will continue this task by successive approximation.

2. THE EXISTENCE OF BEHAVIORAL ATOMS

Language Seems to Be Spatio-Temporally Discrete

Consider the vocabulary of a standard English-speaking adult. This vocabulary contains 26 letters and no more than several thousand words of various sorts, of which only several hundred are most frequently used in daily discourse. Consider the way in which we hear and say the simplest verbal units of daily discourse, such as single letters like A. An obvious feature of this usage is that we never try to decompose A into two or more finer subparts, as for example we can with a word consisting of more than one syllable. Yet even complicated words may be decomposed into no more than finitely many simple parts, and clearly there are only a finite number of simple pieces in any one person's vocabulary.

If we wish to understand our usage of such simple verbal units as A, we must take seriously our impression that A is a *single* unit that is never decomposed in actual speech. We do this by assuming that A is represented in M by a *single state*. That is, we assign to A a single point p_A in M. We also assign a point p_B to B, p_C to C, and so on. In more mathematical terminology, given any n simple behavioral units r_i, $i = 1,2,..., n$, we define n points p_i in M, $i = 1, 2,..., n$, to stand for these units, as in Fig. 1.

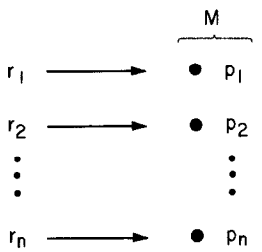

FIG. 1

The reader interested primarily in our mathematical postulates can proceed to the next section, but we will linger momentarily to discuss our impressions of simple behavioral units, since these reveal a rather deep property that any theory of learning might profitably have, and which the present theory has.

When a standard English speaking adult hears a word spoken or speaks a word himself, the word seems to occur at a single instant in time. That is, we can say either that the word has, or has not, been said at a given time in a perfectly definite way. Moreover, no more than a finite number of words are spoken in a lifetime. Thus, both "spatially" (the number of verbal units) and "temporally" (the number of time instants at which verbal units occur), language seems to have many properties of a finite, or discrete, phenomenon.

THE REPRESENTATION OF SENSORY CONTINUA BY DISCRETE SYMBOLS

One of the most vital uses of language is to report our sensory experiences, such as variations in tactile pressure, light intensity, loudness, taste, etc. Many of these sensory impressions seem to vary in a continuous way both in space and in time. A basic characteristic of much sensory experience is that it seems to be *spatio-temporally continuous.*

Yet we successfully use language, which seems to be quite spatio-temporally discrete, to express—or to represent—sensory experience, which seems to be spatio-temporally continuous. The representation by language of sensations requires that the two kinds of phenomena interact, and so, mathematically speaking, we must envisage the interaction of spatio-temporally discrete and continuous processes of

such a kind that the relatively discrete process provides an adequate representation of the relatively continuous process. Moreover, although each sensory modality seems to provide us with essentially different varieties of experience, the very same language tools are adequate for describing at least the rudiments of all of these various modalities. Thus, the discrete representation of continuous processes must be a *universal representation* of some kind. For this reason, we expect conclusions about the dynamics of language behavior to generalize to many other psychological phenomena.

Learning as a Bridge from Continuity to Discreteness

The centrality of the connection between relatively discrete and continuous phenomena in behavior is better understood by considering several simple examples. Consider the phenomenon of walking for specificity. When a child begins to learn how to walk, he must concentrate much effort on the endeavor, and must attend continually to his efforts. An observer is struck by the many motions of the child that are inessential to the walking process, and by the total absorption of the child in the process. In an adult, walking takes on a different appearance. A first step is automatically followed by a second, the second by a third, etc. Once the decision to walk is made, the walk essentially takes itself, and one can pay attention to other matters so long as a minimal amount of obstacle avoidance is accomplished. After walking to one's destination, one "decides" to stop walking and the walk comes to an end. Whereas a child must continuously attend to the walking process until he has mastered it, the adult attends essentially only to starting and stopping the walk, and the mechanics of walking are entirely automatic. Starting and stopping are "on"–"off" responses, which are discrete. Thus, walking requires *continuous* attention before its mechanism is mastered, but only *discrete* attention thereafter. The very process of learning how to walk involves a passage from a relatively continuous representation of voluntary efforts at walking to a relatively discrete representation of these efforts.

A comparable example can be found in language learning. When a young child first begins to learn a letter such as A, an observer is aware of the relatively slow and seemingly continuous juxtaposition of complicated lip, tongue, and associated motions governing pronunciation of the letter. Once A is learned, A can be emitted rapidly and in a seemingly simple integrated motion occurring at a given instant of time. Saying the letter A becomes after learning a simple and discrete act. This situation is analogous to the example of walking, where again an initial state that is continuous both in space and time converges (or contracts) to an asymptotic state, approximately discrete both in space and time. Examples can be drawn from many varieties of learning experience. The fundamental conclusion is that learning often involves a passage from continuous representations of the control of a given act to a more discrete representation of this control.

The Pyramid of Discrete Acts

The intuitive significance of such a passage is easy to see. Once the saying of a verbal unit seems to the performer to be a simple act rather than a tremendously complicated juxtaposition of delicately poised muscular motions, he can proceed to integrate several of these units into more complicated composite units constructed from sequences of seemingly simple acts. After these composite units also seem to be simple, the composite units themselves can be organized into still more complicated composites, and so on. Without the reduction of continuous (and complicated) acts to discrete (and simple) acts, the integration of more complicated behavior based on these acts would seem hopelessly complicated. We would be doomed to paying attention day and night to walking and other menial endeavors. The passage from initially continuous representations of behavioral controls to asymptotically discrete representations is thus no casual event. It makes possible the emergence of new organized behavior patterns, and is a prerequisite for effective learning.

The Continuous and Discrete Pictures Coexist

Since different behavioral sequences in different stages of learning can often coexist, all intermediates between continuity and discreteness can in principle coexist at any time.

The pervasiveness of the coexistence of discrete and continuous representations can be seen from the following example. When a single letter, such as A, is said to a standard English speaking adult, his impression is that A is presented at a single instant of time and that A seems to be a simple behavioral unit. Nonetheless, if scalp electrodes are placed on his head when A is presented, there will ensue a temporally prolonged and spatially widespread alteration in his brain waves (Walter, 1953). Thus the impression that A is spatio-temporally discrete must be reconciled with the fact that A's presentation *simultaneously* causes spatio-temporally continuous alterations in neural potentials. This conclusion is not surprising if only because of the representation of the sound of A as it travels through the air as a complicated series of waves.

Properties of discreteness and continuity coexist at every stage of learning. The continuous background is never wholly eliminated. We must study how certain processes superimposed on this background become increasingly discrete relative to an initially prescribed standard of continuity, and will have at our disposal at least two different levels of dynamical graining such that the degree of continuity of one level takes on meaning only relative to the degree of continuity of the other.

To postulate that A is represented by a single point p_A in M amounts to the hypothesis that A, as a simple behavioral unit, has already been learned by M. We therefore enter the learning process in the middle, and seek to know how known simple behavioral units are integrated into more complicated units, such as the alphabet

ABC...Z. Once we see how new units are formed from old, whether we call our original points p_A, p_B,..., etc., or by another name will seem irrelevant.

3. THE TIME SCALE OF THE MACROSCOPIC WORLD SEEMS CONTINUOUS

The impression from daily experience that time flows continuously is taken for granted in all physical theories. Since we wish to maintain as close a contact to daily experience as possible, we too will suppose that both E and M have a continuous time scale t.

A theory constructed in continuous time has the substantial formal advantage of being able to consider arbitrary input spacing without *ad hoc* changes in parameter values. For example, suppose that E tries to teach M the alphabet *ABC...Z* by presenting the letters with an intratrial interval of w. As w approaches 0 or ∞, the list becomes impossible to learn, whereas the list can more readily be learned at some intermediate value of w. The explanation of even this fact can be cumbersome when discussed in terms of a model in discrete time, but it is trivial in the continuous time theory to be discribed.

4. THE EXISTENCE OF CONTINUOUSLY DIFFERENTIABLE STATE FUNCTIONS

The word "see" and the letter "*C*" sound alike in daily discourse. If I say "see" to someone, he might well reply, "See what?" But if I say "*ABC*" to him, it is far more likely that he will reply by saying "*D*."

To make this latter assertion with confidence, we must specify the rate w at which A, then B, and then C are said. If w is a few seconds, then D is certainly a likely reply to *ABC*. If w is 24 hours, then "See what?" is a more likely reply. And as w varies smoothly from seconds to hours, the effect of the "context" *AB* gradually wears off in the determination of a reply to C. This is only one example of many where the effects of prior events linger and then gradually fade away.

We must be able to represent in M that an event such as A has occurred at a recent time. The point p_A alone does not suffice to do this, since there is no time variation in p_A. There must exist some function, or functions, *of time t* that do this for M. Since we have, in Sec. 2, emphasized that A seems simple in daily experience, we should try to restrict ourselves to just *one* function of time at p_A. We denote this function by $x_A(t)$. Thus to every simple behavioral unit r_i, we postulate the existence in M of a point p_i, and a function $x_i(t)$ representing a process taking place at p_i, $i = 1, 2,..., n$. We now discuss several properties of $x_i(t)$.

$x_i(t)$ is continuously differentiable. $x_i(t)$ was introduced to represent within M the occurrence and gradual fading away through time of the event r_i presented to M at a

given time. This is a question about the rate of change of $x_i(t)$ through time, or about $\dot{x}_i(t)(= dx_i(t)/dt)$. Since the effect of an event wears off gradually, we assume that $\dot{x}_i(t)$ is continuous.

$x_i(t)$ is Nonnegative (M is Observable). The data available to a psychological experimenter E is of two kinds: either a stimulus or response does not occur at a given time, or it does. We are predisposed to express the occurrence of "nothing" by a statement that some quantity is zero. Thus, if A is never presented to M, we set $x_A(t) \equiv 0$. Suppose A is presented to M for the first time at $t = t^{(A)}$. Then surely $x_A(t) = 0$, for $t \leqslant t^{(A)}$. But $x_A(t)$ cannot remain zero for all $t > t^{(A)}$, since $x_A(t)$ was, after all, introduced to represent the occurrence of A. When something occurs, we are predisposed to assign a positive weight to the quantity representing the event, and therefore we suppose that $x_A(t)$ becomes positive when $t > t^{(A)}$.

As t increases, the effect of A's occurrence at time $t = t^{(A)}$ gradually wears off. Thus $x_A(t)$ must gradually return to the level signifying that A has not recently occurred, namely zero. The graph of $x_A(t)$, given exactly one occurrence of A at time $t = t^{(A)}$, thus takes on approximately the form described in Fig. 2.

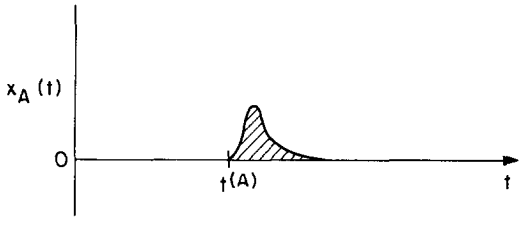

FIG. 2

In particular $x_A(t)$ is nonnegative for all t. (By a change in our sign conventions, we could have just as well assumed that $x_A(t)$ is nonpositive for all t.)

To express Fig. 2 mathematically, we need a way to translate the occurrence of A at time $t = t^{(A)}$ into mathematical terms. There is a standard mathematical way of doing this. That is, let an input $I_A(t)$ perturb $x_A(t)$ at time $t = t^{(A)}$. $x_A(t)$ grows most quickly when $I_A(t)$ is large, and decays towards zero when $I_A(t)$ is zero. The simplest mathematical way of saying this is

$$\dot{x}_A(t) = -\alpha x_A(t) + I_A(t), \tag{1}$$

where α is a positive constant, and the initial data of x_A, say $x_A(0)$, is nonnegative.

We can readily determine some of the basic properties of $I_A(t)$ from (1) and our previous remarks. Since both $x_A(t)$ and $\dot{x}_A(t)$ are continuous, (1) implies that $I_A(t)$ is also continuous. $x_A(t)$ is nonnegative to represent the effect on M of the occurrence or nonocurrence of A. Since $I_A(t)$ is E's way of presenting A to M, (1) shows that $I_A(t)$ should be nonnegative. In the present example, $I_A(t)$ stands for the presentation of

A to M at time $t = t^{(A)}$. Thus, $I_A(t)$ becomes positive once t exceeds $t^{(A)}$. Since this presentation takes only a finite amount of time to occur, $I_A(t)$ becomes zero once again after a finite amount of time. We summarize these conclusions about $I_A(t)$ in Fig. 3.

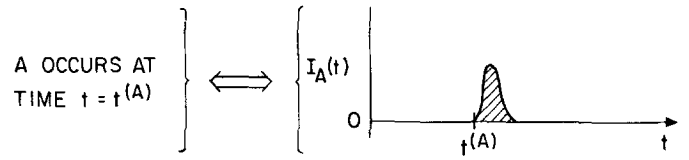

FIG. 3

Equation 1 describes a machine in which $x_A(t)$ can become as large as we please if $I_A(t)$ is taken sufficiently large. In a machine within which $x_A(t)$ has a fixed maximum M_A, (1) is replaced by

$$\dot{x}_A(t) = -\alpha x_A(t) + (M_A - x_A(t))\, I_A(t), \tag{1'}$$

where $0 \leqslant x_A(0) \leqslant M_A$. It is obvious that $x_A(t) \leqslant M_A$ for all $t \geqslant 0$ no matter how large $I_A(t)$ becomes. That is, $x_A(t)$ saturates at M_A. Throughout the following discussion, we will always consider (1) for specificity, but all our conclusions apply to (1') as well with obvious modifications.

Let us consider experiments in which E presents A to M at more than one time instant. Suppose that A occurs at the times $t_1^{(A)}, t_2^{(A)}, \ldots, t_{N_A}^{(A)}$, where $t_i^{(A)} < t_{i+1}^{(A)}$, $i = 1, 2, \ldots, N_A - 1$. Our previous discussion of $I_A(t)$ can be extended to this situation if we suppose that $I_A(t)$ becomes large momentarily at all the times $t = t_i^{(A)}$, $i = 1, 2, \ldots, N_A$, as we show in Fig. 4.

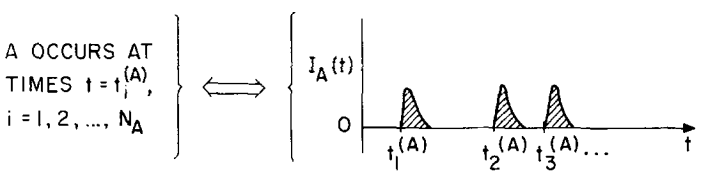

FIG. 4

Figure 4 can be expressed mathematically in the following way. Let $J_A(t)$ be a fixed nonnegative and continuous function which is positive in an interval of the form $(0, \lambda_A)$, $\lambda_A > 0$. Then Fig. 4 can be expressed as

$$I_A(t) = \sum_{k=1}^{N_A} J_A(t - t_k^{(A)}); \tag{2}$$

that is, as a succession of input pulses $J_A(t - t_k^{(A)})$ at the times $t = t_k^{(A)}$. The waveform described by a particular choice of $J_A(t)$ is the "signature" of the given event A in M.

The above remarks are true for all simple behavioral units r_i, $i = 1, 2, \ldots, n$, and not merely A. We can therefore generalize (1) and (2) by writing

$$\dot{x}_i(t) = -\alpha x_i(t) + I_i(t), \; i = 1, 2, \ldots, n, \tag{3}$$

where

$$I_i(t) = \sum_{k=1}^{N_i} J_i(t - t_k^{(i)}),$$

and each $J_i(t)$ is a nonnegative continuous function that is positive in an interval $(0, \lambda_i)$, $\lambda_i > 0$. Equation 3 translates the occurrence of *any* sequence of symbols chosen from $r_1, r_2, \ldots, r_n$, and occurring at *any* times $t_k^{(i)}$, into a definite choice of inputs delivered to M.

Having defined the input $I_i(t)$ to p_i, we remark in passing that the output $O_i(t)$ from p_i will ultimately be given by

$$O_i(t) = \max\{x_i(t)\, \bar{H}(t) - \Gamma_i, O\},$$

where Γ_i is a positive "response threshold" and

$$\bar{H}(t) = 1 + \frac{\sum_{k=1}^{n} X_k(t) \ln_2 X_k(t)}{\ln_2 n},$$

with

$$X_k(t) = x_k(t) \left[\sum_{m=1}^{n} x_m(t) \right]^{-1}.$$

The mathematical properties of this definition are discussed in Grossberg (1968b). In brief these properties are as follows. $\bar{H}(t)$ is closely related to the familiar entropy function of probability theory, which is defined for any probability distribution $p_1, p_2, \ldots, p_n$ by

$$H(p_1, \ldots, p_n) = -\sum_{k=1}^{n} p_k \ln_2 p_k,$$

since

$$\bar{H}(t) = 1 - \frac{H(X_1(t), X_2(t), \ldots, X_n(t))}{\ln_2 n}.$$

It is well known (Khinchin, 1957) that (i) H achieves its maximum of $\ln_2 n$ if and only if all $p_i = 1/n$; (ii) H achieves its minimum of 0 if and only if, for some fixed i, $p_i = 1$ and all $p_j = 0$, $j \neq i$; and (iii) H is a continuous function. Therefore, (i') $\bar{H}(t)$ approximates its minimum of 0 if and only if all stimulus traces $x_i(t)$ are approximately

equal; and (ii') $\overline{H}(t)$ approximates its maximum of 1 if and only if one stimulus trace $x_i(t)$ is much larger than all other stimulus traces.

By (i'), it is clear that all outputs $O_i(t)$ equal 0 if all stimulus traces are approximately equal. That is, if there exists no preference within the machine for any symbol at time t, then no guess is made at time t. In this sense, equal stimulus traces, no matter how large, inhibit each other away before an output can be generated by them. By contrast if, as in (ii'), only one stimulus trace $x_i(t)$ is large at time t, then

$$O_i(t) \cong \max(x_i(t) - \Gamma_i, 0)$$

whereas

$$O_j(t) = 0, \qquad j \neq i.$$

Thus, no output will arise from the weak stimulus traces $x_j(t), j \neq i$. An output will arise from the strong stimulus trace $x_i(t)$ just so long as $x_i(t) > \Gamma_i$; that is, if p_i has been excited recently by a sufficiently large input that the output threshold Γ_i is achieved. The onset of a positive output at time t from v_i is translated as the occurrence of the guess r_i by the machine at time t. The input pulses $J_i(t)$ which create these outputs are fixed once and for all in a given machine before an experiment begins. Many of our qualitative conclusions hold for any choice of continuous $J_i(t)$ with a single maximum and a duration less than τ, as Grossberg (1968e) shows.

The function $\overline{H}(t)$ expresses a kind of mutual inhibition of associations in the production of outputs, whereas the constants Γ_i describe output thresholds. Grossberg (1969b) shows how to improve these inhibitory and threshold effects using a simple formal argument, and thereby derive equations which agree, at least formally, with empirically measured physiological mechanisms of lateral inhibition (Ratliff, 1965) and spiking thresholds (Eccles, 1957). The empirical Hartline-Ratliff equation for lateral inhibition is also derived as a special case. The main heuristic point of these deductions is that the physiological mechanisms can then, at least formally, be discussed as provisions needed to make perfect learning and efficient guessing possible.

The discussion above shows that $O_i(t)$ reduces essentially to $x_i(t)$ minus a constant threshold shift Γ_i if only a couple of stimulus traces are large at time t. Since our thought experiments in this paper involve only a couple of $x_i(t)$ functions at a time, the assumption that the output from p_i is $x_i(t)$ is quite satisfactory.

5. THE PRODUCTION OF OUTPUTS BY INPUTS AFTER LEARNING HAS OCCURRED

Our remarks to now have discussed only how the presentation of an r_i to M is represented within M by suitable fluctuations of $x_i(t)$ at p_i; that is how M "recognizes" or "perceives" these events. We have said nothing about how M learns. We now begin to fill this gap.

Consider M after it has learned the list AB. Suppose AB has been presented many times by E to M in the past. Since M now knows AB, if E presents A alone to M, then M must reply a short time later by saying B. We now ask how this can happen.

The presentation of A to M at time $t = t^{(A)}$ has been conceptualized as the occurrence of an *input* $J_A(t - t^{(A)})$ delivered by E to p_A at time $t^{(A)}$. Thus, M's reply to E a short time later should be an *output* delivered by p_B to E. This output arises from p_B at (say) time $t = t^{(A)} + \tau_{AB}$, where τ_{AB} is some positive reaction time. The τ_{AB} is positive simply because responses to stimuli take some time to arise. Which function at p_B is the output? Only one function, namely $x_B(t)$, is associated with p_B. We suppose for simplicity that $x_B(t)$ is the desired output. In summary, after AB has been learned, an input to p_A at time $t = t^{(A)}$ gives rise to an output from p_B at time $t = t^{(A)} + \tau_{AB}$.

An input to p_A at time $t = t^{(A)}$ creates a momentary increase in $x_A(t)$. Thus presenting A to M at time $t = t^{(A)}$ and receiving B in reply at time $t = t^{(A)} + \tau_{AB}$ has the effect on M which we have diagrammed in Fig. 5.

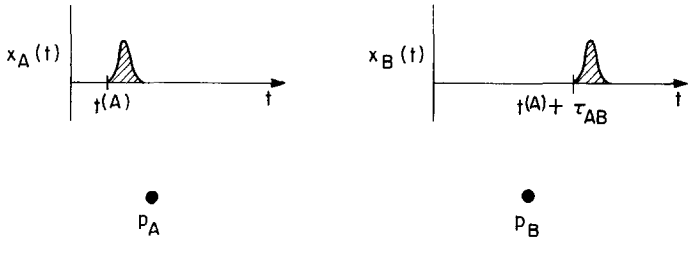

FIG. 5

E causes only the increase in $x_A(t)$. The mechanism of M itself must cause the increase in $x_B(t)$ τ_{AB} time units later. Figure 5 shows, however, that the only possible cause of this increase in $x_B(t)$ is the prior increase in $x_A(t)$. A *signal* from p_A is thus carried to p_B with a delay of τ_{AB} time units, and this must be true whenever $x_A(t)$ is large after AB has been learned. Since the signal reaching p_B from p_A is large at time t if and only if $x_A(t - \tau_{AB})$ is large, we suppose for simplicity that the signal is proportional to $x_A(t - \tau_{AB})$, and choose positive proportionality constants β and p_{AB} such that the signal equals $\beta p_{AB} x_A(t - \tau_{AB})$.

To write this conclusion in mathematical terms, we need only observe that the signal from p_A to p_B is an input to p_B, just as $I_B(t)$ is an input to p_B. We therefore replace the equation

$$\dot{x}_B(t) = -\alpha x_B(t) + I_B(t),$$

by the slightly more complicated equation

$$\dot{x}_B(t) = -\alpha x_B(t) + I_B(t) + \beta p_{AB} x_A(t - \tau_{AB}),$$

which also takes into account the signal from p_A to p_B.

The previous argument must hold when the list AB is replaced by any list $r_i r_j$ which M can learn. Thus, after M learns $r_i r_j$,

$$\dot{x}_j(t) = -\alpha x_j(t) + I_j(t) + \beta p_{ij} x_i(t - \tau_{ij}), \tag{4}$$

where, just as in the special case $r_i = A$ and $r_j = B$, p_{ij} and τ_{ij} are positive constants.

How is the signal from p_A to p_B carried to p_B? We will envisage some pathway over which the signal travels without decrement at a finite velocity so as not to reach p_B until τ_{AB} time units after it is emitted by p_A. We denote this pathway by e_{AB}. Since the list AB is not the same list as the list BA, $e_{AB} \neq e_{BA}$. That is, e_{AB} is a *directed* pathway from p_A to p_B. We denote it by an arrow facing from p_A to p_B. Thus, for every list $r_i r_j$ which M can learn, an arrow e_{ij} will face from p_i to p_j in order to carry the signal $\beta p_{ij} x_i(t - \tau_{ij})$ after $r_i r_j$ has been learned. Figure 6 diagrams this situation.

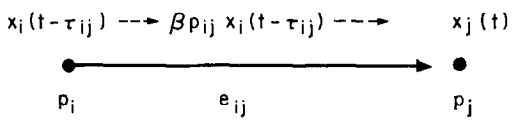

FIG. 6

If it is *impossible* for M to learn $r_i r_j$, then no signal can reach p_j from p_i, and we set $p_{ij} = 0$.

6. THE MECHANISM OF LEARNING

Equation 4 holds for any sequence $r_i r_j$ which has already been learned by M, say AB. Before learning AB, on the other hand, there must exist other possible lists AC, AD, etc., which M could learn instead of AB, for if B were the only possible reply to A, then by definition, AB would have already been learned. This means that p_A must be able to send signals to all points p_B, p_C, p_D,..., which stand for possible successors of A, or else no possible connection between p_A and these alternatives could ever be established. In particular, the points p_j, $j = B, C, D,...$, could never possibly satisfy (4).

We are thrown, therefore, into the following dilemma: *After* learning occurs, we want p_A to send a signal such as (4) *only* to the correct point p_B so that a presentation of A to M creates the reply B. *Before* learning occurs, p_A must be able to send signals to all the points p_j which correspond to symbols r_j that might be learned. The process of learning thus eliminates the signals from p_A to all incorrect points p_C, p_D,..., at the same time that it preserves and strengthens the signal from p_A to p_B.

This can happen in essentially only one way in the picture we have thus far constructed. The only effect on M of saying AB several times, say at a rate w, is to make both $x_A(t - w)$ and $x_B(t)$ large during and shortly after the times $I_A(t - w)$ and $I_B(t)$ are large, respectively. Saying AB more often ensures that $x_A(t - w)$ and $x_B(t)$ are both large more often. If AC were said instead, $x_A(t - w)$ would sometimes be large, but $x_B(t)$ would always remain small. If only B were said, $x_B(t)$ would sometimes be large, but $x_A(t - w)$ would always remain small. If nothing were said to M, then both $x_A(t - w)$ and $x_B(t)$ would always remain small. Thus, the learning of AB occurs if and only if the *product*

$$x_A(t - w)\, x_B(t), \tag{5}$$

is often large, and all other products $x_A(t - w)\, x_j(t)$, $j = C, D, \ldots$, remain small, where $w > 0$ is some "reasonable" learning rate.

In order for M to be capable of learning, a mechanism exists in M which computes these products, or else M would have no way of distinguishing one ordering of inputs from another. Therefore, we postulate the existence of a process $z_{AB}(t)$ somewhere in M which grows only if $x_A(t - w)\, x_B(t)$ is large. $z_{AB}(t)$ can only take place at some position in M where both the values $x_A(t - w)$ and $x_B(t)$ are simultaneously present, but there is only one place in Fig. 6 at which past x_A values (such as $x_A(t - w)$) and present $x_B(t)$ values are simultaneously present. This place is at the *arrowhead* N_{AB} of e_{AB}, since only here is the signal $\beta p_{AB} x_A(t - \tau_{AB})$ from p_A contiguous with the $x_B(t)$ value of p_B. We therefore replace the product (5) by the product

$$\beta p_{AB} x_A(t - \tau_{AB})\, x_B(t), \tag{6}$$

and say that $z_{AB}(t)$ grows if and only if (6) is large. The simplest way to express this mathematically is to say that $z_{AB}(t)$ grows at a rate equal to (6), minus perhaps a spontaneous decay (or "forgetting") term $uz_{AB}(t)$. That is, we let

$$\dot{z}_{AB}(t) = -uz_{AB}(t) + \beta p_{AB} x_A(t - \tau_{AB})\, x_B(t).$$

In the same way, we can define a $z_{ij}(t)$ function at the arrowhead N_{ij} of each e_{ij} by

$$\dot{z}_{ij}(t) = -uz_{ij}(t) + \beta p_{ij} x_i(t - \tau_{ij})\, x_j(t), \tag{7}$$

where $\beta > 0$, $u > 0$, $p_{ij} \geqslant 0$, $\tau_{ij} > 0$, and $z_{ij}(0) \geqslant 0$. Figure 6 now becomes Fig. 7.

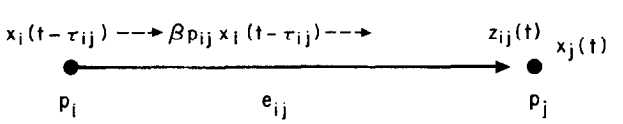

FIG. 7

If z_{ij} has a fixed finite maximum M_{ij}, then we can replace (7) by

$$\dot{z}_{ij}(t) = -uz_{ij}(t) + \beta p_{ij}(M_{ij} - z_{ij}(t))\, x_i(t - \tau_{ij})\, x_j(t), \tag{7'}$$

just as we replaced (1) by (1'). In summary, the functions $z_{ij}(t)$ exist at the arrowheads because this is the only place where past signals from p_i and present signals from p_j coexist, and the past signal from p_i is needed so that an input to p_i will give rise after a short pause to a correct output from p_j once $r_i r_j$ has been learned.

We have defined functions such as $z_{AB}(t)$ not only to record whether or not AB has been frequently presented to M, but also to guarantee that after AB has been learned, an output to p_A generates an output *only* from p_B τ_{AB} time units later. To achieve this mathematically, note the following heuristic requirements.

If A is said but AB has not been learned, then B will not be said in reply τ_{AB} time units later. If A is not said, then B will not be said τ_{AB} time units later even if AB has been learned. And if A is not said and AB has not been learned, then surely B will not be said in reply. Saying A amounts to momentarily increasing $x_A(t)$. Saying B in reply amounts to momentarily increasing $x_B(t + \tau_{AB})$. And having learned AB amounts to keeping $z_{AB}(w)$ large at least for w chosen within the times that $x_A(t)$ and $x_B(t + \tau_{AB})$ are large. Since $x_B(t + \tau_{AB})$ will become large in this situation only if the signal received by p_B from p_A is large, our heuristic requirements show that $z_{AB}(w)$ must influence the size of the signal $\beta p_{AB} x_A(t)$ while it is being transferred through the arrowhead N_{AB} from e_{AB} to p_B. This occurs at time $w = t + \tau_{AB}$. Indeed, our heuristic requirements imply that $x_B(t + \tau_{AB})$ becomes large only if both $\beta p_{AB} x_A(t)$ and $z_{AB}(t + \tau_{AB})$ are large, or only if the *product*

$$\beta p_{AB} x_A(t)\, z_{AB}(t + \tau_{AB}),$$

is large. In terms of arbitrary indices i and j, this means that the input to p_j from p_i at time t is

$$\beta p_{ij} x_i(t - \tau_{ij})\, z_{ij}(t).$$

Equation 4 is therefore replaced by

$$\dot{x}_j(t) = -\alpha x_j(t) + I_j(t) + \beta p_{ij} x_i(t - \tau_{ij})\, z_{ij}(t), \tag{8}$$

and Fig. 7 is replaced by Fig. 8.

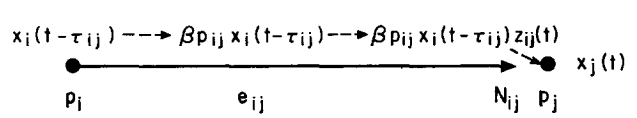

FIG. 8

7 THE INDEPENDENCE OF LISTS

Consider now a machine M in which the lists AB and CB can both be learned; that is, $p_{AB} > 0$ and $p_{CB} > 0$. We want to be able to learn AB independently of C if C is never said during the learning process, and to be able to learn CB independently of A if A is never said. (We temporarily ignore higher order conditioning effects. Some of these will be an automatic consequence of our considerations.) That is, we want the two inputs $\beta p_{AB} x_A(t - \tau_{AB})\, z_{AB}(t)$ and $\beta p_{CB} x_C(t - \tau_{CB})\, z_{CB}(t)$ to combine *independently* at p_B. Mathematically speaking, combining two quantities in an independent way means: add them. Thus, the total input received by p_B from p_A and p_C at time t is

$$\beta[x_A(t - \tau_{AB})\, p_{AB} z_{AB}(t) + x_C(t - \tau_{CB})\, p_{CB} z_{CB}(t)].$$

Or more generally, the total input received by r_j from all r_k, $k = 1, 2, \ldots, n$, at time t is

$$\beta \sum_{k=1}^{n} x_k(t - \tau_{kj})\, p_{kj} z_{kj}(t),$$

and (8) is replaced by

$$\dot{x}_j(t) = -\alpha x_j(t) + \beta \sum_{k=1}^{n} x_k(t - \tau_{kj})\, p_{kj} z_{kj}(t) + I_j(t) \tag{9}$$

for every $j = 1, 2, \ldots, n$.

Equations 7 and 9 together form a mathematically well defined proposal for a learning machine M. The next section shows how to modify such a machine slightly to make it learn much better. This modification is suggested both by a heuristic oversight in our derivation and by a corresponding formal difficulty. The modification is made without observability. It then suggests a deeper set of equations with further physiological implications in Grossberg (1969b).

8. THE NORMALIZATION OF $p_{ij} z_{ij}$

Consider the problem of learning AB vs AC once again. The letters B and C are heuristically thought of as a "set of response alternatives" to A, and the strengthening of B as a reply to an isolated presentation of A carries with it the weakening of C as a reply to A. Otherwise expressed, the choice of B as a response to A is made only relative to the strength of other response alternatives, or response alternatives compete with one another.

We will show that by overlooking this rudimentary fact, we have constructed a system with some unpleasant formal properties. Then we will include the competition between response alternatives in a simple way, and simultaneously, automatically overcome the formal difficulties. Grossberg (1969h) studies a related case.

Consider the problem of learning AB vs AC once again, and suppose for simplicity that no other lists can be learned, so that only p_{AB} and p_{AC} are positive. We assume for

simplicity that all $\tau_{ij} = \tau > 0$, so that M has a well defined "reaction time" τ. Then (9) becomes

$$\dot{x}_B(t) = -\alpha x_B(t) + \beta x_A(t - \tau)\, p_{AB} z_{AB}(t) + I_B(t),$$

and

$$\dot{x}_C(t) = -\alpha x_C(t) + \beta x_A(t - \tau)\, p_{AC} z_{AC}(t) + I_C(t). \tag{10}$$

We assume that A, B, and C have occurred at least once in the remote past of (10). By (7), we can therefore suppose that $z_{AB}(t) > 0$ and $z_{AC}(t) > 0$ for all the times t which we will consider. In this setting, we reinvestigate the task of teaching AB to M, and observe the following technical difficulties:

a) z_{AC} *Remains Too Large.* When A occurs, the signal $\beta x_A(t - \tau)\, p_{AC}$ from p_A along e_{AC} grows. Since $z_{AC}(t) > 0$, a positive signal $\beta x_A(t - \tau)\, p_{AC} z_{AC}(t)$ reaches p_A and causes, by (10), a momentary increase in the value of $x_C(t)$. Consequently $\beta p_{AC} x_A(t - \tau)\, x_C(t)$ also grows momentarily, and so, by (7), $z_{AC}(t)$ is momentarily boosted in its value as well. Then the cycle repeats itself, with the net effect that saying A alone helps to keep $z_{AC}(t)$ from decaying at an exponential rate, even though C is never said. Of course, $z_{AB}(t)$ grows much faster than $z_{AC}(t)$ during this time. We can surely guarantee that $z_{AB}(t) \gg z_{AC}(t)$ as a result of saying AB sufficiently often, but we cannot guarantee that *only* the flow from p_A to p_B eventually survives the learning process. This is the main formal deficiency of the process (7) and (9). A related secondary difficulty is the following one:

(b) *Instability of the Transformation from inputs to outputs.* If AB has occurred very often in the recent past, then $z_{AB}(t)$ can grow very large. Even a very small input $I_A(t)$ to p_A can therefore create a very large output $x_B(t + \tau)$ from p_B, because the signal $\beta x_A(t)\, p_{AB} z_{AB}(t + \tau)$ from p_A to p_B will be large even though $x_A(t)$ is small. We desire, however, an equation such as (4) after learning has occurred, in which an input generates a correct output of comparable size.

These examples suggest that we replace the functions $p_{ij} z_{ij}(t)$ which control the size of the flow from p_i to p_j by new functions $y_{ij}(t)$ which avoid the formal difficulties of (a) and (b), and which express the intuitive idea that response alternatives compete. Then (9) is replaced by

$$\dot{x}_j(t) = -\alpha x_j(t) + \beta \sum_{k=1}^{n} x_k(t - \tau)\, y_{kj}(t) + I_j(t), \tag{11}$$

$j = 1, 2, \ldots, n$. We now list several properties which y_{ij} should have, and then exhibit a simple function that realizes all of these properties.

Consider $y_{AB}(t)$ for specificity. $y_{AB}(t)$ should be a function only of $p_{AB} z_{AB}(t)$, $p_{AC} z_{AC}(t), \ldots$, and $p_{AZ} z_{AZ}(t)$, since only these functions control the size of the flow from p_A to possible response points $p_B, p_C, \ldots, p_Z$. That is,

$$y_{AB}(t) = f_{AB}(p_{AB} z_{AB}(t), p_{AC} z_{AC}(t), \ldots, p_{AZ} z_{AZ}(t))$$

for some as yet unknown f_{AB}.

Consider now a learning experiment in which only AB occurs, and C, D,..., Z have occurred only in the remote past. Then we should be able to lump together the non-occurring letters C, D,..., Z, since they are never distinguished one from the other by any experimental operation. That is

$$y_{AB}(t) = g_{AB}(p_{AB}z_{AB}(t), p_{AC}z_{AC}(t) + \cdots + p_{AZ}z_{AZ}(t)), \tag{12}$$

for some as yet unknown function $g_{AB} = g_{AB}(u, v)$ of $u \geqslant 0$ and $v \geqslant 0$. We now itemize various desirable properties of g_{AB}. g_{AB} is nonnegative since the function $p_{AB}z_{AB}$ which it replaces is nonnegative. To avoid the problem of (b) we also assume that g_{AB} is bounded from above. Since an as yet unspecified positive constant β multiplies y_{AB} in (11), we can take this bound to be 1 without loss of generality. That is,

$$0 \leqslant g_{AB} \leqslant 1. \tag{13}$$

As M learns AB better and better, we want $p_{AB}z_{AB}$, and thus y_{AB}, to grow. That is,

$$g_{AB}(u, v) \text{ is monotone increasing in } u. \tag{14}$$

Similarly, if the incorrect alternatives $p_{AC}z_{AC} + \cdots + p_{AZ}z_{AZ}$ get to be learned better, then learning of AB is jeopardized and y_{AB} decreases. That is,

$$g_{AB}(u, v) \text{ is monotone decreasing in } v. \tag{15}$$

The difficulty in (a) shows that, at best, saying AB very often implies for t sufficiently large that

$$p_{AB}z_{AB}(t) \gg p_{AC}z_{AC}(t) + \cdots + p_{AZ}z_{AZ}(t).$$

We also want $y_{AB}(t)$ to be very close to its maximum 1 at such times. That is,

$$u \gg v \text{ implies } g_{AB}(u, v) \cong 1. \tag{16}$$

Similarly, if AB has been very poorly learned, then

$$p_{AB}z_{AB}(t) \ll p_{AC}z_{AC}(t) + \cdots + p_{AZ}z_{AZ}(t)$$

and also $y_{AB}(t)$ is very close to its minimum 0. That is,

$$u \ll v \text{ implies } g_{AB}(u, v) \cong 0. \tag{17}$$

And certainly,

$$g_{AB}(u, v) \text{ is continuous in } u \text{ and } v. \tag{18}$$

We now ask if a function satisfying all the conditions (12)–(18) exists. The answer is "yes" and perhaps the simplest such function is given by

$$y_{AB}(t) = \frac{p_{AB}z_{AB}(t)}{p_{AB}z_{AB}(t) + p_{AC}z_{AC}(t) + \cdots + p_{AZ}z_{AZ}(t)}.$$

That is, we need merely change $p_{AB}z_{AB}(t)$ into the ratio of $p_{AB}z_{AB}(t)$ compared with

all the functions $p_{Ai}z_{Ai}(t)$ that control a flow from p_A to a possible response point p_i, $i = 1, 2,..., n$. This definition of $y_{AB}(t)$ immediately generalizes to

$$y_{ij}(t) = p_{ij}z_{ij}(t)\left[\sum_{k=1}^{n} p_{ik}z_{ik}(t)\right]^{-1},$$

for all $i, j = 1, 2,..., n$. This definition of $y_{ij}(t)$ clearly embodies the idea that the choice of B given A is made only relative to other response alternatives. For example, if B and C are the only response alternatives to A, then $p_{AA} = p_{AD} = p_{AE} = \cdots = p_{AZ} = 0$, so that

$$y_{AB}(t) = \frac{p_{AB}z_{AB}(t)}{p_{AB}z_{AB}(t) + p_{AC}z_{AC}(t)}$$

and

$$y_{AC}(t) = \frac{p_{AC}z_{AC}(t)}{p_{AB}z_{AB}(t) + p_{AC}z_{AC}(t)}.$$

By nonnegativity of $p_{AB}z_{AB}(t)$ and $p_{AC}z_{AC}(t)$, an increase in B given A (i.e., in $y_{AB}(t)$) implies a decrease in C given A (i.e., in $y_{AC}(t)$), and achieves this competition between alternatives by "relativizing," or dividing, $p_{AB}z_{AB}(t)$ by the sum of $p_{AB}z_{AB}(t)$ and $p_{AC}z_{AC}(t)$. The conditions (12)–(18) can therefore be thought of as some formal prerequisites for competitive choices among response alternatives to occur in our machines. In Grossberg (1969b), this competition between choices is shown to be closely related to the physiological process of lateral inhibition in much the same way that the outputs $O_i(t)$ are. We have hereby derived the following system of nonlinear difference-differential equations to describe M.

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta \sum_{m=1}^{n} x_m(t - \tau)\, y_{mi}(t) + I_i(t), \tag{19}$$

$$y_{jk}(t) = p_{jk}z_{jk}(t)\left[\sum_{m=1}^{n} p_{jm}z_{jm}(t)\right]^{-1}, \tag{20}$$

and

$$\dot{z}_{jk}(t) = -uz_{jk}(t) + \beta p_{jk}x_j(t - \tau)\, x_k(t), \tag{21}$$

for all $i, j, k = 1, 2,..., n$. This completes our derivation of the mathematical laws governing the machines M. We now single out a particularly important collection of the machines that are currently undergoing a systematic mathematical analysis. If the y_{ij}'s are not used, then the numerical parameters in Eqs. (7) and (9) must be carefully chosen to avoid (a) and (b) (Grossberg 1969h).

9. LOCALLY UNBIASED MACHINES

If $p_{jk} = 0$, then (21) becomes $\dot{z}_{jk} = -uz_{jk}$, or $z_{jk}(t) = z_{jk}(0)\, e^{-ut}$, and $z_{jk}(t)$ decays to zero at an exponential rate. Since $p_{jk} = 0$ also implies that $y_{jk}(t) \equiv 0$, or that no

flow whatsoever passes from p_j to p_k, we can for convenience set z_{jk} identically equal to zero without changing M in any nontrivial way. We therefore replace (21) by

$$\dot{z}_{jk}(t) = \begin{cases} -uz_{jk}(t) + \beta p_{jk} x_j(t-\tau)\, x_k(t), & \text{if} \quad p_{jk} > 0 \\ 0, & \text{if} \quad p_{jk} = 0, \end{cases} \tag{21'}$$

and by the initial condition that $z_{jk}(0) > 0$ if and only if $p_{jk} > 0$.

Our mathematical studies of these machines (Grossberg, 1967, 1968a–d, 1969a, d–f) consider only cases where the positive values of p_{jk} have the form

$$p_{jk} = \frac{1}{\lambda_j} > 0.$$

That is, the positive weights leading from a fixed point to all other points are the same. We then call the geometry of M *locally unbiased*. In this case, (20) and (21) can be simplified by letting $Z_{jk}(t) = \lambda_j z_{jk}(t)$ for all j, $k = 1, 2, \ldots, n$, and noting that

$$y_{jk}(t) = p_{jk} Z_{jk}(t) \left[\sum_{m=1}^{n} p_{jm} Z_{jm}(t) \right]^{-1}, \tag{20'}$$

and

$$\dot{Z}_{jk}(t) = \begin{cases} -uZ_{jk}(t) + \beta x_j(t-\tau)\, x_k(t), & \text{if} \quad p_{jk} > 0 \\ 0, & \text{if} \quad p_{jk} = 0. \end{cases} \tag{21''}$$

The main advantage of using Z_{jk} instead of z_{jk} is that the coefficients p_{jm} now occur only in (20'). Since all common factors can be divided out of the positive values among $p_{j1}, p_{j2}, \ldots$, and p_{jn} which appear in (20'), we can assume without loss of generality that

$$\sum_{m=1}^{n} p_{jm} = 0 \quad \text{or} \quad 1, \qquad j = 1, 2, \ldots, n.$$

10. THE NEURON HYPOTHESIS

A considerable amount of anatomical and physiological investigation has gone into the demonstration of the existence of nerve cell bodies, axons, endbulbs, synapses, and the directed transmission of neural impulses from the nerve cell body towards the synapse (Crosby, 1962; Eccles, 1957, 1964). These investigations show that membrane potentials at the cell body give rise to spikes traveling down the axon in frequencies that vary systematically with variations in membrane potential. Once these spikes reach the endbulb they cause a release in transmitter that travels across the synaptic cleft and influences the postsynaptic potential.

Striking analogs of all these processes exist in our machines M. Each point p_i can

roughly be thought of as a collection of cell bodies, each edge e_{ij} can roughly be thought of as the collection of axons leading from cells in p_i to cells in p_j, each arrowhead N_{ij} as the endbulbs attached to these axons, and the gap between N_{ij} and p_j as the corresponding synapses. Given this obvious candidate for a neural interpretation of the geometry of M, the following interpretation of the dynamical variables of M is readily suggested. $x_i(t)$ roughly corresponds to the average membrane potential over the cells corresponding to p_i, $\beta x_i(t)$ is the spiking frequency in the axons corresponding to e_{ij}, and $y_{ij}(t)$ is the state of transmitter production in the endbulbs corresponding to N_{ij}. Once these identifications are made, then the flow of $\beta x_i(t)$ to N_{ij} followed by the input $\beta x_i(t)\, y_{ij}(t + \tau)$ to p_j reads: after the membrane potential generates a spike, it travels along the axon to the endbulb, where it activates the transmitter control process at the endbulb and releases a quantity that increases both with increases of spiking frequency and with the amount of available transmitter. This statement has a very familiar neurological ring to it. See Grossberg (1969b) for a more detailed physiological account.

In a clear sense, therefore, we have been led, from purely psychological postulates to some of the basic qualitative facts of the neuron hypothesis, in particular the existence of directed transmissions along a network-like structure, the existence of a process at the network arrowheads, and the interaction of the transmissions and arrowhead processes to produce inputs to the recipient "cell bodies." These conclusions are independent, moreover, of the detailed functional form of Eqs. 19–21. They follow quite readily from our remarks concerning the existence of reaction times, and the places at which processes could possibly exist to distinguish one ordering of inputs from another.

11. A POSSIBLE MECHANISM OF NEURAL LEARNING

We have also been led to a new idea of how learning occurs. Thus, the functions $z_{ij}(t)$ grow only if *both* the presynaptic influence from p_i via the signal $\beta p_{ij} x_i(t - \tau)$ *and* the postsynaptic value $x_j(t)$ are large. That is, a coupling of both pre- and postsynaptic influences is needed to increase the level of transmitter production and, thereupon, the strength of the connection from p_i to p_j.

In forthcoming papers, we explore the possible physiological means whereby such a "trophic" effect of postsynaptic influences on the endbulb can take place by replacing the postulate of observability by a more realistic one (Grossberg, 1969,b c).

12. REACTION TIMES AND LEARNING RATES

In Sec. 3, we observe that a variation from 0 to ∞ of the presentation rate w of a long list takes us from an impossible learning task to a more tractable task, and back again to an impossible task. We now show that our machines also have this property.

We begin with a locally unbiased machine M in a state of "maximal ignorance." Suppose, for example, that

$$p_{ij} = \begin{cases} \dfrac{1}{n-1}, & i \neq j \\ 0, & i = j, \end{cases}$$

that M's initial data satisfies $x_i(v) = \gamma(v)$, $v \epsilon [-\tau, 0]$, where γ is continuous and nonnegative, and that $z_{jk}(0) = \delta p_{jk}$ where $\delta > 0$, and $i, j, k = 1, 2, \ldots, n$. Let us present the list $r_1, r_2, \ldots, r_n$ to M at a rate w. Thus $I_1(t) = J(t)$, $I_2(t) = J(t - w)$, $I_3(t) = J(t - 2w), \ldots$, and $I_n(t) = J(t - (n-1)w)$, where $J(t)$ is some input pulse. Suppose $w = 0$. Then $I_1(t) = I_2(t) = \cdots = I_n(t)$, and by symmetry, $x_i(t) = x_j(t)$ and $z_{ij}(t) = z_{km}(t)$ for all $i \neq j$, $k \neq m$, and $t \geqslant 0$. Thus M remains in a state of maximal ignorance for all $t \geqslant 0$, and nothing is learned. Similarly, if w is very small relative to the duration of $J(t)$, then again by symmetry, we will expect M to remain close to a state of maximal ignorance.

Now suppose $w = \tau > 0$. Then for any $i = 1, 2, \ldots, n-1$, the signal created by I_i at p_i reaches p_{i+1} at the same time that I_{i+1} becomes large at p_{i+1}. This means that the product $x_i(t - \tau)\, x_{i+1}(t)$ will become large relative to all the products $x_i(t - \tau)\, x_j(t)$, $j \neq i + 1$. By (21), the function $z_{i,i+1}$ will be given a strong boost in its values as compared to the functions z_{ik}, $k \neq i + 1$. Thus $y_{i,i+1}$ will grow considerably, whereas all y_{ik}, $k \neq i + 1$, will decay. Substantial learning therefore occurs. The same argument manifestly holds for values of w which are of the order of τ.

If $w \gg \tau > 0$, then the signal created by I_i at p_i reaches p_{i+1} long before I_{i+1} becomes large at p_{i+1}. Since I_i becomes zero long before I_{i+1} occurs, the signal from p_i to p_j also becomes very small before I_{j+1} becomes large. Thus, all the products $x_i(t - \tau)\, x_j(t)$ are always either equal or very small, and so the function $z_{i,i+1}$ grows little more than the functions z_{ik}, $k \neq i + 1$. All the functions y_{ik} remain approximately equal, and little learning occurs.

This argument shows that the maximal learning rates in M are of the order of magnitude of its reaction time τ. Once we decided that M's reply to an input should be delayed in time, we tacitly prescribed the places in M where functions z_{ij} could possibly be computed, and thus in turn the relative timing of inputs which could lead to efficient learning. Grossberg (1969b) describes machines which can effectively predict items at a rate somewhat faster than the rate at which they learned them.

13. STIMULUS TRACES, ASSOCIATIONAL STRENGTHS, AND SPACE–TIME CONTEXTS

The functions y_{ij} determine the strength of M's reply r_j to an isolated presentation of r_i. We therefore call $y_{ij}(t)$ the *associational strength of $r_i r_j$ at time t*, by analogy with classical theorizing, such as that of Hull (Hilgard, 1956). The associational strengths $y_{ij}(t)$ collectively contain M's memory of past experiments.

The size of $x_i(t)$ determines how actively p_i has been perturbed by recent inputs, including presentations of r_i. $x_i(t)$ thus, in part, plays the role of a *stimulus trace* (Hilgard, 1956).

Equations 20 and 21 show that changes in associational strength are due to changes in the stimulus traces. Equation 19 shows that changes in the associational strengths alter the stimulus traces. Insofar as the stimulus traces are "perceptual" or "recognition" variables, and the associational strengths are "learning" variables, we see that processes of learning and perception in M form a single unified process rather than being two qualitatively different aspects of M's experience.

Because of this strong binding between perception and learning in M, we can see how M automatically forms an appropriate "perceptual set," or "learning set," or "space–time context" in response to particular experiences. For example, suppose that only a small subset of points p_i are perturbed by inputs I_i during a time interval that is long compared to the rate of decay of the x_i's. These x_i's will grow as a result of the inputs, and will send large signals only to the points x_j whose y_{ij} values are large, i.e., only to those states that are "contextually" related to the x_i's by past experiences. In this way, a "spatial context" is automatically created by the inputs. This context changes continually as time goes on and new inputs I_j perturb new points p_j while the previously perturbed x_i values decay. That is, a "space–time context" is automatically carved out by the inputs.

Suppose that a time interval exists that is long relative to the decay time of the x_i in which only a few points p_i are perturbed in rapid succession. Then only the y_{ij} values which connect these points to one another will grow, so that even as a space–time context is being formed by the mere distribution of inputs, the "learned" contextual associates of given points also change. Hence, the space–time context formed by the very same distribution of inputs might well change if these inputs are repeated again and again.

These facts illustrate one way by which a machine which stores a very large vocabulary can call upon small subsets of this vocabulary as experiences demand without activating all of its repertoire unnecessarily. The inputs automatically carve out those channels in M which correspond to the experiences, and the remembrances of related past experiences, that the inputs represent. These channels fluctuate through time as the demands of experience do.

14. MARKOVIAN AND NON-MARKOVIAN

The context which is formed at any time in M depends on the rate at which inputs are presented. Suppose, for example, that $ABCD$ is presented once at rate w. That is $I_A(t) = I_B(t + w) = I_C(t + 2w) = I_D(t + 3w)$. If w is not large compared to τ and to the rate of decay of the x_i's, then all of the point strengths $x_A(t)$, $x_B(t)$, $x_C(t)$, and $x_D(t)$

will be large right after D is presented. $x_D(t)$ will often be largest since D has just occurred, and $x_A(t)$ will often be smallest since A occurred some time before, but all these point strengths will have some influence on the determination of the nodal strengths $y_{ij}(t)$, the magnitude of their influence depending on their relative size at any time. Events which occur prior to D (i.e., ABC) will influence the behavior of M after D occurs, and so the "past" affects the "future." In the mathematical literature such an effect is said to be "non-Markovian" (Kemeny and Snell, 1960).

Now let w increase. Suppose w is chosen far larger than both τ and the decay time of the point strengths $x_i(t)$. Again let $ABCD$ be presented at rate w, and consider M shortly after D has occurred. Then each of $x_A(t)$, $x_B(t)$, and $x_C(t)$ will be very small when $x_D(t)$ is large, because their inputs occurred so long ago that they have decayed back to their resting position. Therefore, only D determines the future behavior of M when D is presented to M, i.e., the "future" depends only on the "present." Such a dependence is mathematically called "Markovian."

We see in this simple way that our systems can behave in both a Markovian or non-Markovian fashion depending on the particular choices of inputs to which they are exposed. This fact suggests that our systems can also behave in an "all-or-none" or "gradualist" fashion, depending on the particular experiment, since all-or-none learning is distinguished from gradual learning by different effects of past on future events. In a later paper we show that this is the case. See Grossberg (1969d), for example.

15. LINEAR AND NONLINEAR

Our systems combine linear and nonlinear effects in an unusual way. For example, consider (19)–(21) when $\sum_{m=1}^{n} p_{im} = 1$ for all $i = 1, 2, \ldots, n$; i.e. every point p_i sends an edge to *some* point p_j. Then it is seen by summing over $i = 1, 2, \ldots, n$ in (19) and dividing by n that the average output $x = (1/n)\sum_{k=1}^{n} x_k$ is related to the average input $I = (1/n)\sum_{k=1}^{n} I_k$ by a linear equation

$$\dot{x}(t) = -\alpha x(t) + \beta x(t - \tau) + I(t),$$

even though the interaction of the $x_i(t)$'s along the edges e_{jk} is nonlinear. Thus our systems are often "linear in the large" although they are "nonlinear in the small." This linear behavior in the large is independent of the y_{jk}'s, and thus of all learning effects.

Consider (19)–(21) when all associational strengths $y_{ij}(t)$ have approached limiting values θ_{ij} as a result of a learning experiment; that is, $y_{ij}(t) \cong \theta_{ij}$ for times $t \geqslant T$, where T is some large time after practice has been going on for awhile. Then (19) becomes

$$\dot{x}_i(t) \cong -\alpha x_i(t) + \beta \sum_{k=1}^{n} x_k(t - \tau)\, \theta_{ki} + I_i(t),$$

which is approximately a system of linear equations for the outputs $x_i(t)$ in terms of the inputs $I_i(t)$. If the inputs to M are sufficiently regular in time that learning occurs, then M's behavior automatically passes from a nonlinear phase to a linear phase. Since M's output will seem linear after a sufficient amount of practice has occurred in an experiment, it is tempting to try to model M's mechanism in a linear way. Nonetheless, the learning mechanism of M is nonlinear, so that such an extrapolation will work well only after M has already learned. Thus linearizing in the present situation destroys the very mechanism of learning that we wish to study. This is proved rigorously in Grossberg (1969e).

16. GESTALT, GUTHRIE, AND PAVLOV

Consider a machine M before it has learned anything. Suppose that M is capable of learning any list chosen from $r_1, r_2, \ldots, r_n$ in which no symbol r_i occurs more than once. Suppose also that M is unbiased for specificity. Then

$$p_{ij} = \begin{cases} \dfrac{1}{n-1}, & i \neq j \\ 0, & i = j. \end{cases}$$

Since M begins in a state of maximal ignorance, all $x_i(v)$ are equal, $i = 1, 2, \ldots, n$, for $v \in [-\tau, 0]$. All $z_{jk}(0)$ are also equal, $j \neq k$, and are *positive*. Now let any symbol be presented to M, say r_1. Then x_1 grows momentarily and large signals are transmitted to all the other points $p_j, j \neq 1$. If r_2 then occurred, p_2 sends large signals to all the other points $p_j, j \neq 2$. And so on. Before learning occurs, therefore, the entire "field" of points is influenced by an event at a single point, i.e., a kind of "Gestalt" effect "in space" occurs (Hilgard, 1956).

Similarly, if the list $r_1 r_2 \ldots r_n$ is presented to M at a rate w which is not large compared to τ and the decay rate of the point strengths, then several point strengths will determine together the alterations in nodal strength at that time, as pointed out in Sec. 14, i.e., a Gestalt effect "in time" occurs. In summary, if M begins in a state of ignorance, then M exhibits Gestalt effects in space–time whenever it is exposed to a long and rapidly occurring list of symbols.

Let us now consider M after it has learned the list $r_1 r_2 \ldots r_n$. Then, by definition,

$$y_{12}(t) \cong y_{23}(t) \cong y_{34}(t) \cong \cdots \cong y_{n-1,n}(t) \cong 1$$

for all times t during which M knows the list, and all other $y_{ij}(t)$ are approximately zero. Thus, a chain of associational strengths leads from p_i to p_2, from p_2 to p_3, and so on until p_{n-1} and p_n are reached. This chain has been embedded into the field of M's alternatives—hence the name "embedding fields" for our theory.

We can now readily see that M's behavior after learning is qualitatively different from its behavior before learning. After learning, an input to p_1 creates a large signal τ time units later only at p_2, and so only p_2 delivers a large output to E at this time. Similarly a large input to p_2 creates a large signal τ time units later only at p_3, and so only p_3 delivers a large output to E at this time. This is so for every p_i and p_{i+1} with $i = 1, 2, \ldots, n-1$, and the Gestalt effect in space–time has been substantially eliminated. It has been replaced by a simple succession of "stimuli" and "responses," and the response which occurs depends on the contiguity of the stimulus in the list $r_1 r_2 \ldots r_n$. This kind of behavior is often associated with the name of Guthrie (Hilgard, 1956).

We have, therefore, at our disposal a machine which would be a delight to the Gestaltists before learning occurs and a discomfort to them after learning occurs. The same machine would be a comfort to Guthrie after learning and a stranger in his house before.

Before learning occurs, M is a complicated network of transitions representing the many possible alternative choices at M's disposal. After learning occurs, M becomes a simple chain, or circuit, in which no choices remain. That is, M's behavior is reduced to a series of reflexes, or to a "Pavlovian circuit" (Hilgard, 1956).

We wish to suggest by these examples that our machines contain within them formal properties that are highly suggestive of various theoretical movements drawn from the history of psychology. As in the case of Gestalt vs Guthrie, these formal properties can sometimes appear at different times during the very same experiment, and a learning or perceptual mechanism which seems adequate to describe the effects of one kind of experiment often seems hopelessly unsuitable for the description of a closely related experiment. We wish to suggest that this difficulty arises when theorizing is done by tacitly or explicitly assuming mathematical properties, such as linearity and locality, which are simply not generally valid but which nonetheless work quite well for specific kinds of experiments. Our machines also exhibit some of these properties for one kind of experiment (i.e., initial data and inputs) and different properties for another kind of experiment. It will be of interest to test whether these changes in formal properties correspond, at least qualitatively, to sources of controversy in classical psychological theories.

17. BACKWARD LEARNING

Consider an unbiased machine which can learn both AB and BA, for example, the machine M depicted in Fig. 9. Thus, $p_{AB} = p_{AC} = p_{AD} = p_{BA} = p_{BC} = p_{BD} = \frac{1}{3}$, and all other $p_{ij} = 0$. Let M begin in a state of "maximal ignorance" and "at rest." That is,

$$x_A(v) = x_B(v) = x_C(v) = x_D(v) = 0, v \epsilon [-\tau, 0],$$

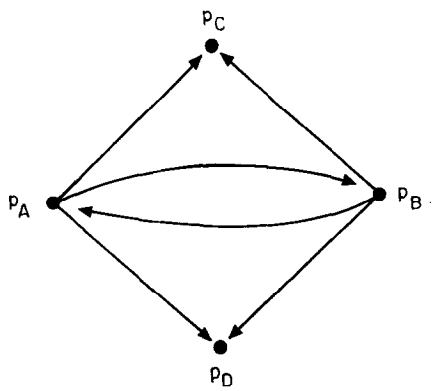

FIG. 9

and

$$z_{AB}(0) = z_{AC}(0) = z_{AD}(0) = z_{BA}(0) = z_{BC}(0) = z_{BD}(0).$$

We now show that teaching the machine AB at a speed $w = \tau$ automatically teaches the machine BA, but to a lesser extent; i.e., backward learning occurs. Thus, learning BA given only the occurrence of AB follows from the mere possibility of learning BA at all! The list BA is certainly a short list. Grossberg (1969d) shows that this conclusion must be substantially qualified in the case of long lists. We now show how BA is learned given only the occurrence of AB.

When A occurs at time $t = 0$, $x_A(t)$ grows and equal signals are sent along e_{AB}, e_{AC}, and e_{AD} towards p_B, p_C, and p_D, respectively. As these signals reach the arrowheads N_{AB}, N_{AC}, and N_{AD} at time $t = \tau$, B occurs. p_C and p_D thereafter receive only signals from N_{AC} and N_{AD}. p_B, on the other hand, receives a signal from N_{AB} as well as an input I_B. Thus, $z_{AB} > z_{AC} = z_{AD}$ for all times after B occurs. Consequently $y_{AB} > y_{AC} = y_{AD}$ as well, and at least partial learning of AB has occurred.

After B occurs, $x_B(t)$ sends out equal signals along e_{BA}, e_{BC}, and e_{BD} to p_A, p_C, and p_D, respectively. These signals begin to reach their destination at time $t = 2\tau$. p_A has, however, also received the input I_A 2τ time units earlier. Although the effect of I_A has partially worn off by time $t = 2\tau$, x_A is still larger than $x_C(t)$ and $x_D(t)$. After the signal from p_B arrives at p_A, p_C, and p_D, therefore, $y_{BA}(t) > y_{BC}(t) = y_{BD}(t)$, and thus at least partial learning of the backward list BA occurs. Whereas the overlap in time of the signal from p_A to p_B and the input $I_B(t)$ to p_B is perfect, the signal from p_B to p_A arrives only after the effects of $I_A(t)$ have partially worn off at p_A. Thus we expect $y_{AB}(t) > y_{BA}(t)$, or learning in the forward direction is better than learning in the backward direction.

As the speed w of saying AB is allowed to approach zero, the asymmetry between the states p_A and p_B due to differences in the timing of the inputs I_A and I_B gradually vanishes. Indeed, when $w = 0$, $y_{AB}(t) = y_{BA}(t)$ for all $t \geq 0$, by symmetry. Thus, the relative advantage of forward learning over backward learning in M arises simply because the relative timing of inputs from E to M and of signals within M favors the forward direction when the presentation speed w is sufficiently close to the optimal learning speed τ.

If $ABAB...$ is presented very often to M at a periodic speed $w = \tau$, then clearly the initial bias of saying A first will gradually wear off, and the difference $y_{AB}(t) - y_{BA}(t)$ will decrease as t increases. By contrast, if ABC is presented to M, then the forward association y_{BC} will competitively diminish the backward association y_{BA}, so that the rudiments of a forward "arrow in time," namely, $y_{AB} > y_{BA}$ and $y_{BC} > y_{CB}$, will be established within M.

18. LEARNING WITHOUT REVERBERATION

In his classic book, Hebb (1949) discusses the possibility that neural memories are preserved by a form of persistent reverberation within neural networks. We now remark that reverberation is quite unnecessary for memories to be preserved in our machines. Indeed, reverberation is one of the processes most destructive of M's memory that can occur.

Reverberation in M means that large x_i signals pass cyclically between the points p_i. Firstly, we show that M's memory is perfect if all the signals are maximally small; that is, if all $x_i(t)$ are identically zero.

Suppose $x_i(t) = 0$, $i = 1, 2,..., n$. Then by (21), if $p_{jk} > 0$, then $\dot{z}_{jk} = -uz_{jk}$, or $z_{jk}(t) = z_{jk}(0)\, e^{-ut}$. Thus by (20),

$$\begin{aligned} y_{jk}(t) &= \frac{p_{jk} z_{jk}(0)\, e^{-ut}}{\sum_{m=1}^{n} p_{jm} z_{jm}(0)\, e^{-ut}} \\ &= \frac{p_{jk} z_{jk}(0)}{\sum_{m=1}^{n} p_{jm} z_{jm}(0)} \\ &= y_{jk}(0). \end{aligned}$$

The associational strengths remain constant for all time, and M's memory is perfect. Hence, reverberation is surely unnecessary for M to remember very well.

Reverberation harms M's memories because whenever too many x_i's have large values, the values of many y_{ij}'s will also change, just as in the formation of spatio-temporal contexts. Changes of the y_{ij}'s mean changes in M's memory.

Moreover, we want the values $x_i(t)$ to become small whenever the inputs $I_i(t)$ are zero over long time intervals, because these values, or at least a subset of them, are the

outputs of M, and we would like M to be able to remember without persistently spelling out its memories in large outputs to the outside world. Small outputs are often desirable, and small outputs imply little reverberation and good memory within M.

19. BRAIN AND BEHAVIOR

We now have two ways by which we can, at least roughly, interpret our mathematical variables: one psychological and one neurological. The point strength $x_i(t)$ stands both for a stimulus trace and for an average membrane potential. The nodal strength $y_{jk}(t)$ stands both for an associational strength and for the average state of transmitter production in a collection of endbulbs. To the extent that our machines M are realistic models of simple behavioral systems, we can now translate a psychological fact into a neural property, and conversely. We have at our disposal at least a partial proposal for a language to suggest how "brains control behavior." This translation table between neural and psychological variables will be extended systematically in later papers of this series.

20. M IS NOT ENTIRELY OBSERVABLE

We constructed M with nonnegative states $x_i(t)$ to try to guarantee that the states be observable to E. Nonetheless the functions $z_{jk}(t)$ cannot be directly measured by E, and these functions contain the heart of M's learning mechanism. $z_{jk}(t)$ is a hidden or intervening variable, and our mathematical papers prove rigorously that various fundamental features of M's behavior are not directly measurable by E, in spite of all our efforts to maximize M's observability to E. In particular, the protocol of M's stimuli and responses does not provide a complete picture of M's learning mechanism.

21. INPUTS AND OUTPUTS VS STIMULI AND RESPONSES

Much psychological theorizing is based on the use of the concepts of stimulus and of response to a stimulus. In complicated experimental situations, one is then sometimes forced to discuss stimuli which share some response properties, and responses which share some stimulus properties, the degree of sharing depending on the situation. We believe that the stimulus–response terminology is often an inconvenient one because it does not correspond in a simple way to the way in which we learn. This is clear even in our simple machines M.

A stimulus r_i to M is an input to p_i. Then p_i sends out signals to other p_j. These signals reach the p_j as inputs also. Are these inputs stimuli to p_j? Since p_j can distin-

guish only the size of an input and not its source, our answer must be "yes," using the "principle of sufficient reason." Thus a stimulus becomes identical with any input that a point receives. In a similar fashion, a response becomes identical with any output that M emits. But this is certainly not the customary way in which S–R terminology is used.

It is well known that realistic behavioral systems benefit substantially from the feedback created by their own behavior; for example, we can organize our speech better when we can hear our words. In M, this means that an output from a state should create a subsequent input to that state via some form of feedback through the physical medium surrounding M. By virtue of our previous remarks, this means that every response also has stimulus properties. To avoid a terminology which does not clearly distinguish the process that decides how important the stimulus or response aspects of an event are, we propose instead that one simply classify the inputs and outputs which occur, and study systemically the mechanisms that connect them.

REFERENCES

CROSBY, E. C., HUMPHREY, T., AND LAUER, E. W. *Correlative anatomy of the nervous system.* New York: Macmillan, 1962.

DE ROBERTIS, E. D. P. *Histophysiology of synapses and neurosecretion.* New York: Macmillan, 1964.

ECCLES, J. C. *The physiology of nerve cells.* Baltimore: Johns Hopkins Press, 1957.

ECCLES, J. C. *The physiology of synapses.* New York: Academic Press, 1964.

GROSSBERG, S. Nonlinear difference-differential equations in prediction and learning theory. *Proceedings of the National Academy of Sciences of the United States of America*, 1967, **58**, 1329–1334.

GROSSBERG, S. Global ratio limit theorems for some nonlinear functional-differential equations, I, II. *Bulletin of The American Mathematical Society*, 1968, **74**, 95–105. (a)

GROSSBERG, S. A prediction theory for some nonlinear functional-differential equations. I. learning of lists. *Journal of Mathematical Analysis and Applications*, 1968, **21**, 643–694. (b)

GROSSBERG, S. A prediction theory for some nonlinear functional-differential equations. II. learning of patterns. *Journal of Mathematical Analysis and Applications*, 1968, **22**, 490–522. (c)

GROSSBERG, S. Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 1968, **59**, 368–372. (d)

GROSSBERG, S. Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Sciences of the United States of America*, 1968, **60**, 758–765. (e)

GROSSBERG, S. On the global limits and oscillations of a system of nonlinear differential equations describing a flow on a probabilistic network. *Journal of Differential Equations*, 1969, **5**, 291. (a)

GROSSBERG, S. On learning, information, lateral inhibition, and transmitters. *Mathematical Biosciences*, 1969, in press. (b)

GROSSBERG, S. On the production and release of chemical transmitters and related topics in cellular control. *Journal of Theoretical Biology*, 1969, **22**, 325. (c)

GROSSBERG, S. On the serial learning of lists. *Mathematical Biosciences*, 1969, in press. (d)

GROSSBERG, S. On the variational systems of some nonlinear difference-differential equations. *Journal of Differential Equations*, 1969, in press. (e)

GROSSBERG, S. Some networks that can learn remember, and reproduce any number of complicated space-time patterns, I. *Journal of Mathematics and Mechanics*, 1969, in press. (f)

GROSSBERG, S. On learning of spatiotemporal patterns by networks with ordered sensory and motor components, I — excitatory components of the cerebellum. *Journal of Mathematics and Physics*, 1969, in press. (g)

GROSSBERG, S. Some networks that can learn, remember, end reproduce any number of complicated space-time patterns, II. SIAM *Journal of Applied Mathematics*, 1969, submitted for publication. (h)

HEBB, D. O. *The organization of behavior*. New York: Wiley, 1949.

HILGARD, E. R. *Theories of learning*. New York: Appleton, Crofts, 1956.

JENSEN, A. R. An empirical theory of the serial position effect. *Journal of Psychology*, 1962, **53**, 127.

KEMENY, J. G., AND SNELL, J. L. *Finite Markov chains*. Princeton, New Jersey: Van Nostrand, 1960.

KHINCHIN, A. I. *Mathematical foundations of information theory*. New York: Dover, 1957.

MILLER, G. A. The magic number seven, plus or minus two. *Psychological Review*, 1956, **63**, 81.

OSGOOD, C. E. *Method and theory in experimental psychology*. New York: Oxford Univer. Press, 1953.

RATLIFF, F. *Mach bands: quantitative studies on neural networks in the retina*. San Francisco: Holden-Day, 1965.

WALTER, W. G. *The living brain*. New York: Norton, 1953.

RECEIVED: November 10, 1967