Reprinted from:
PROGRESS IN THEORETICAL BIOLOGY, VOL. 3

ACADEMIC PRESS, INC.
New York San Francisco London

# Classical and Instrumental Learning by Neural Networks

Stephen Grossberg*

*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*



* This work was supported in part by the Alfred P. Sloan Foundation and the Office of Naval Research (N00014-67-A-0204-0051).

## I. Introduction

### A. Embedding Fields: A Psychophysiological Theory

This article reviews results chosen from the theory of *embedding fields*. Embedding field theory discusses mechanisms of pattern discrimination and learning in a psychophysiological setting. It is derived from psychological postulates that correspond to familiar behavioral facts. The theory tries to isolate facts which embody fundamental principles of neural design, and which therefore imply and illuminate many less evident facts and predictions. The postulates reveal their implications by being translated into rigorous mathematical expressions. On various occasions, the precision of this mathematical language has uncovered unsuspected physical properties of the postulates, or corrected erroneous conclusions of prior heuristic thinking. In particular, the mathematics can be given a natural anatomical and physiological interpretation. The neural networks hereby derived can thus be rigorously analyzed both behaviorally and neurally.

### B. The Method of Minimal Anatomies

The theory introduces a particular method to approach the several levels of description that are relevant to understanding behavior. This is the method of *minimal anatomies*. At any given time, we will be confronted by laws for neural components, which have been derived from psychological postulates. The neural units will be interconnected in specific anatomies. They will be subjected to inputs that have a psychological interpretation, which create outputs that also have a psychological interpretation. At no given time could we hope that all of the more than $10^{12}$ nerves in a human brain would be described in this way. Even if a precise knowledge of the laws for each nerve were known, the task of writing down all the inter-

actions and analyzing them would be bewilderingly complex and time-consuming. Instead, a suitable method of successive approximations is needed. Given specific psychological postulates, we derive the *minimal* network of embedding field type that realizes these postulates. Then we analyze the psychological and neural capabilities of this network. An important part of the analysis is to understand what the network cannot do. This knowledge often suggests what new psychological postulate is needed to derive the next, more complex network. In this way, a hierarchy of networks is derived, corresponding to ever more sophisticated postulates. This hierarchy presumably leads us closer to realistic anatomies, and provides us with a catalog of mechanisms to use in various situations. Moreover, once the mechanisms of a given minimal anatomy are understood, variations of this anatomy having particular advantages or disadvantages can be readily imagined. The procedure is not unlike the study of one-body, then two-body, then three-body, and so on, problems in physics, leading ever closer to realistic interactions; or the study of symmetries in physics as a precursor to understanding mechanisms of symmetry-breaking; or the study of thermodynamics as a preliminary to statistical mechanical investigations.

At each stage of theory construction, natural formal analogs of nontrivial psychological and neural phenomena emerge. We shall denote these formal properties by their familiar experimental names. This procedure emphasizes at which point in theory construction, and ascribed to which mechanisms, these various phenomena first seem to appear. No deductive procedure can justify this process of name-calling, and incorrect naming of formal network properties does not compromise the formal correctness of the theory as a mathematical consequence of the psychological postulates. Nonetheless, if ever psychological and neural processes are to be unified into a coherent theoretical picture, such name-calling, with all its risks and fascinations, seems inevitable, both as a guide to more microscopic theory construction and as a tool for a deeper understanding of relevant data. The following pages will attempt to distinguish clearly between postulates, mathematical properties, factual data, and mere interpretations of network variables.

This policy of theory construction has more than practical convenience to recommend it. Even a routine behavioral act can utilize billions of nerves distributed along complexly interacting pathways that extend from sensory receptors to motor effectors. The organization of these pathways—the *global* properties of the network—powerfully influence the transformation of stimuli into responses. To the extent that these properties are ignored, one loses insight into the behavioral constraints which guide neural development and design. Even if one's neural data about individual



cells are precise, the meaning of these data can remain obscure until global information about the role of these cells in behavior is obtained.

## C. Overview

This article summarizes some main results that are distributed in several papers (Grossberg, 1968a,b, 1969a-f, 1970a,b, 1971a-c, 1972a-d; Grossberg and Pepe, 1971). An intuitive description of mathematical results will replace mathematical details wherever possible. Emphasis will be placed on theoretical ideas. Relevant data are discussed in the references.

The theory begins by analyzing simple facts about classical, or Pavlovian, conditioning (Grossberg, 1969a, 1971a). This form of learning is illustrated by the following experiment. A hungry dog is presented with food and thereupon salivates. A bell is rung, but the dog does not salivate. Then the bell is rung just before food presentation on several learning trials. Thereafter presentation of the bell alone yields salivation. Food is called the unconditioned stimulus (UCS), salivation is called the unconditioned response (UCR), and the bell is called the conditioned stimulus (CS). Thus Pavlovian conditioning is a problem in nonstationary prediction: The CS eventually predicts the UCR if it is paired sufficiently often with the UCS. Alternatively, this learning process can be described by considering an experimentalist, $\mathfrak{E}$, who interacts with a machine, $\mathfrak{M}$, to teach $\mathfrak{M}$ to predict B given A by practicing the list AB. The sensory presentation of A is analogous to a CS, the sensory presentation of B is analogous to UCS, and the motor response, B, is analogous to a UCR.

The first derivation of the theory asks how a particular CS $\rightarrow$ UCR transition can be embedded in memory by sequential pairing of CS and UCS, and how future presentation of the CS can elicit the UCR. This derivation is reviewed in Section II.

Given the derivation, which is based on only the most rudimentary facts about classical conditioning, a number of psychophysiological and mathematical surprises ensue. For example, the mathematical systems that arise already have a natural anatomical and neurophysiological interpretation which includes cell bodies, axons, synaptic knobs, cell potentials, spiking thresholds and frequencies, and transmitter substances. In psychological terms, one finds such items as short-term memory (STM) traces, long-term memory (LTM) traces, a stimulus sampling theory, Now Print or Amplifier mechanisms, imprinting mechanisms, serial learning phenomena, a way to learn arbitrary patterns, influences of overarousal on paying attention, and a teleology for attacking problems of sensory filtering and pattern discrimination. Pattern discrimination problems will not be discussed herein.

Given this foundation, the theory proceeds in several directions. First, it studies the minimal anatomy that can learn an arbitrary space–time pattern, such as a piano sonata or a dance (see Sections IV and V). Only one command cell is needed to encode the memory of such a pattern, although there exist larger networks that can also do the job with interneurons. The main liability of such a cell is that performance is ritualistic: Once performance of the pattern begins, it cannot be terminated in midcourse even if more urgent environmental demands are made on the network. Anatomies in which sensitivity to environmental feedback exists typically encode sequences of events using many cells. This fact motivates a study of learning in arbitrary anatomies (Section VI). The goal is to find those constraints that make learning in a general context possible. Thereafter, one can specialize the anatomy to perform particular tasks. Studies of serial (or list) learning, and some related problems concerning dependence of serial learning parameters on arousal level, arise as special cases in these investigations (Section VII).

Next the theory is developed in the direction of instrumental, or operant, conditioning; namely, it approaches the question of how learning is influenced by rewards, punishments, drives, etc. (Section VIII). These questions arise naturally from a closer investigation of classical conditioning. For example, the time lags between CS and UCS presentation on successive learning trials need not be the same, since the two events are usually independent of each other. Also, after learning has occurred, the CS elicits the UCR on recall trials in the absence of the UCS. These and a few other simple facts can be used as postulates to derive networks that include mechanisms of reinforcement, drive, and incentive motivation. In short, classical and instrumental learning mechanisms are not conceptually independent. Given these networks, one imposes postulates which are aimed at preventing the network from seeking previously rewarded goals that presently lead to punishment, and which permit learned avoidance of such goals. The above postulates eventually yield networks containing rudimentary formal analogs of midbrain reinforcement centers. These analogs include the interaction of two formal transmitter systems, whose properties can be compared with data concerning cholinergic and adrenergic effects at midbrain sites. Various other facts and predictions about punishment and avoidance formally emerge in these systems.

The theoretical equations can also be refined in several directions to provide a deeper insight into possible chemical substrates of network mechanisms (Section IX). This refinement procedure uncovers various transient interactions, and suggests an important concept: that a cell capable of learning is a chemical dipole, with the two ends of the dipole existing near the cell body and synaptic knobs.



## II. Classical Conditioning

The derivation below (Grossberg, 1969a, 1971a) is given in storybook form to emphasize its intuitive basis. It studies how an experimentalist, $\mathcal{E}$, can teach a machine, $\mathfrak{M}$, to predict B given A by practicing the list AB.

### A. Each Letter Seems Simple

In daily speech and listening, a letter is never decomposed into two parts. To maintain close contact with experience, we assume that a single state, $v_A$, in $\mathfrak{M}$ corresponds to A. In a similar fashion, let $v_B$ correspond to B, $v_C$ to C, etc. We designate each $p_i$ by a point, or vertex. (A vertex is not necessarily an individual cell, but can represent a cell population acting as a control unit.)

### B. Presentation Times

The times at which letters are presented to $\mathfrak{M}$ must be represented within $\mathfrak{M}$. For example, presenting A and then B with a time spacing of 24 hours should yield different behavior than does presentation with a time spacing of 2 seconds. Thus various functions of time should be associated with each vertex to designate how recently a given letter has been presented. To maintain contact with the "one-ness" of each letter, and to maximize the simplicity of our derivation, we let one function $x_A(t)$ be associated with $v_A$, one function $x_B(t)$ be associated with $v_B$, etc., as in Fig. 1.

### C. Continuous Vertex Functions

The functions $x_A(t), \ldots, x_Z(t)$ will be chosen continuous, and in fact differentiable. Several reasons for this exist. One reason is the following. Consider the question: What follows ABC? It is tempting to say D, but really the problem is ill-defined if the letters are presented one at a time with time spacing, $w$, between successive letters. If indeed $w$ is small, say $w \cong 2$ seconds, then D might well be the correct response, but if $w \cong 24$ hours, then to the sound C (= "see") one can also reply "See what?" That is, as $w$ varies from small to large values, the influence of A and B on

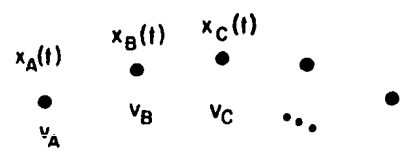


Fig. 1. Vertex functions register how recently given events occur


FIG. 2. Sequential presentation of an event induces sequential perturbation of its vertex function.

the prediction following C gradually wears off. Since $x_A(t)$ and $x_B(t)$ describe the relevance at time $t$ of A and B in $\mathfrak{M}$, we conclude that these functions also vary gradually in time.

## D. Perturbations Instead of Presentations

Suppose that A is never presented to $\mathfrak{M}$. Corresponding to the occurrence of "nothing" is the natural mathematical predisposition to set $x_A(t) = 0$ at all times $t$. (The equilibrium point, 0, can, it turns out, be rescaled ultimately relative to the signal thresholds.)

Suppose that A is presented to $\mathfrak{M}$ for the first time at time $t = t_A$. Then $x_A(t)$ must be perturbed from 0 for certain $t > t_A$, or else $\mathfrak{M}$ would have no way of knowing that A occurred. We associate the occurrence of "something" with a positive deflection in the graph of $x_A$. (The theory could also, in principle, be carried out with negative deflections.)

Shortly after A is presented, A no longer is heard by $\mathfrak{M}$. That is, $x_A(t)$ gradually returns to the value signifying no recent presentation of A, namely 0. In a similar fashion, if A is presented at times $t_A^{(1)} < t_A^{(2)} < \cdots < t_A^{(N_A)}$, then we find the graph of Fig. 2. The same construction holds true for all letters. In this way, we have translated the presentation of any letters A, B, C, ... in the alphabet at prescribed times into a definite sequence of perturbations of the vertex functions $x_A(t), x_B(t), x_C(t), \ldots$.

## E. Linearity

For notational convenience, we replace the alphabet A, B, C, ... by any sequence $r_i$, $i = 1, 2, \ldots, n$, of $n$ behavioral atoms; the vertices $v_A, v_B, v_C, \ldots$ by the vertices $v_i$, $i = 1, 2, \ldots, n$; and the vertex functions $x_A(t), x_B(t), x_C(t), \ldots$ by the vertex functions $x_i(t)$, $i = 1, 2, \ldots, n$. Now $r_i$ corresponds to $[v_i, x_i(t)]$, $i = 1, 2, \ldots, n$.

What is the simplest way to translate Fig. 2 into mathematical terms? Since we are constructing a system whose goal is to adapt with as little bias as possible to its environment, we are strongly advised to make the system as linear as possible. In Section VI, we shall discuss which of these linearities



is really essential. The simplest linear way to write Fig. 2 is in terms of the equations

$$\dot{x}_i(t) = -\alpha_i x_i(t) + C_i(t) \tag{1}$$

with $\alpha_i > 0$, $x_i(0) \geq 0$, and $i = 1, 2, \ldots, n$. The input $C_i(t)$ can, for example, have the form

$$C_i(t) = \sum_{k=1}^{N_i} J_i(t - t_i^{(k)}) \tag{2}$$

where $J_i(t)$ is some nonnegative and continuous function that is positive in an interval of the form $(0, \lambda_i)$. Thus $x_i(t)$ decays at the exponential rate $\alpha_i$ unless it is perturbed by an input pulse $J_i(t - t_i^{(k)})$.

## F. After Learning

In order that $\mathfrak{M}$ be able to predict B given A after practicing AB, interactions between the vertices $v_i$ must exist. Suppose, for example, that $\mathfrak{M}$ has already learned AB, and that A is presented to $\mathfrak{M}$ at time $t_A$. We expect $\mathfrak{M}$ to respond with B after a short time interval, say at time $t = t_A + \tau_{AB}$, where $\tau_{AB} > 0$. The term $\tau_{AB}$ is called the *reaction time* from A to B. Let us translate these expectations into graphs for the functions $x_A(t)$ and $x_B(t)$. We find Fig. 3. The input, $C_A(t)$, controlled by $\mathcal{S}$ gives rise to the perturbation of $x_A(t)$. The internal mechanism of $\mathfrak{M}$ must give rise to the perturbation of $x_B(t)$. In other words, after AB is learned, $x_B(t)$ gets large $\tau_{AB}$ units after $x_A(t)$ gets large.

There exists a linear and continuous way to say this; namely, $v_A$ sends a linear signal to $v_B$ with time lag $\tau_{AB}$. Then Eq. (3) with $i = B$ is replaced by

$$\dot{x}_B(t) = -\alpha_B x_B(t) + \beta_{AB} x_A(t - \tau_{AB}) + C_B(t) \tag{3}$$

with $\beta_{AB}$ some positive constant. More generally, if $r_i r_j$ has been learned, we conclude that

$$\dot{x}_j(t) = -\alpha_j x_j(t) + \beta_{ij} x_i(t - \tau_{ij}) + C_j(t) \tag{4}$$

If $\beta_{ij} = 0$, then the list $r_i r_j$ cannot be learned, since a signal cannot pass from $v_i$ to $v_j$.

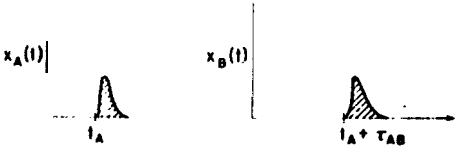


FIG. 3. Traces of sequential A-then-B presentation.

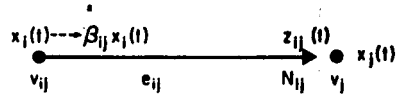


FIG. 4. Network interpretation of psychophysiological variables

## G. DIRECTED PATHS

The signal $\beta_{ij}x_i(t - \tau_{ij})$ from $v_i$ to $v_j$ in Eq. (4) is carried along some pathway at a finite velocity, or else the locality of the dynamics would be violated. Denote this pathway by $e_{ij}$. The pathways $e_{ij}$ and $e_{ji}$ are distinct because the lists $r_ir_j$ and $r_jr_i$ are distinct. To designate the direction of flow in $e_{ij}$, we draw $e_{ij}$ as an arrow from $v_i$ to $v_j$, whose arrowhead, $N_{ij}$, touches $v_j$, as in Fig. 4.

## H. BEFORE LEARNING

Consider the network before any learning occurs. If A leads only to B, then learning would have already occurred. The letter A must therefore also be able to lead to C, D, or some other letters. Thus the process of learning can be viewed as elimination of the incorrect pathways AC, AD, etc., while the correct pathway AB endures, or is strengthened. In other words, the connections between all vertices cannot be constant through time if learning occurs. The constant connections $\beta_{ij}$ must be supplemented by time-varying connection coefficients. (This is a "connectionist" theory in the broadest sense, but, it turns out, not a traditional one.)

## I. DISTINGUISHING ORDER

How does $\mathfrak{M}$ know that AB and not AC is being learned? By Fig. 3, practicing AB means that $x_A$ and then $x_B$ become large several times. Saying A alone, or B alone, or neither A nor B should yield no learning. We seek a mathematical way to distinguish the event "A occurs and then B occurs" from all other possibilities. This can be done most simply as follows. If AB occurs with a time spacing of $w$, then the product $x_A(t - w)x_B(t)$ is large at suitable times $t \cong t_A^{(i)} + w$, $i = 1, 2, \ldots, N_A$. If either A or B does not occur, then the product will be small. We therefore seek a process $\mathfrak{M}$ that can compute products of past $x_A(v)$ values $(v < t)$ and present $x_B(t)$ values. Denote this process by $z_{AB}(t)$. Note that $z_{AB} \neq z_{BA}$.

Where in $\mathfrak{M}$ do past values of $x_A(v)$ and present values of $x_B(t)$ come together, so that $z_{AB}(t)$ can compute them? (Locality again!) By Fig. 4, this happens only in the arrowhead $N_{AB}$. Thus $z_{AB}(t)$ takes place in $N_{AB}$. But then the past $x_A(v)$ value received by $N_{AB}$ at time $t$ is the signal $\beta_{AB}x_A(t - \tau_{AB})$. The most linear and continuous way to express this rule



for $z_{AB}(t)$ is the following:

$$\dot{z}_{AB}(t) = -\gamma_{AB}z_{AB}(t) + \delta_{AB}x_A(t - \tau_{AB})x_B(t)$$

with $\gamma_{AB}$ a positive constant, and $\delta_{AB}$ a nonnegative constant that is positive only if $\beta_{AB}$ is positive. More generally, for list $r_ir_j$ we find in $N_{ij}$ the process

$$\dot{z}_{ij}(t) = -\gamma_{ij}z_{ij}(t) + \delta_{ij}x_i(t - \tau_{ij})x_j(t) \tag{5}$$

Thus $z_{ij}$ depends linearly on the crucial product $x_i(t - \tau_{ij})x_j(t)$. Note that the reaction time, $\tau_{ij}$, replaces $w$ in Eq. (5), owing to locality. This fact adumbrates a connection between reaction times and presentation rates optimal for learning.

## J. GATING OUTPUTS

The $z_{ij}(t)$ function can distinguish whether or not $r_ir_j$ is practiced. But more is desired. Namely, if $r_ir_j$ is practiced, presenting $r_i$ should yield a delayed output from $v_j$. If $r_ir_j$ is not practiced, presenting $r_i$ should not yield an output from $v_j$. And even if $r_ir_j$ is practiced, no output from $v_j$ should occur if $r_i$ is not presented. In other words, $x_j(t)$ should become large only if $x_i(t - \tau_{ij})$ *and* $z_{ij}(t)$ are large. Again a product is called for, and Eq. (4) is changed to

$$\dot{x}_j(t) = -\alpha_jx_j(t) + x_i(t - \tau_{ij})\beta_{ij}z_{ij}(t) + C_j(t) \tag{6}$$

Thus $z_{ij}$'s location in $N_{ij}$ allows $z_{ij}$ both to compute products of past signals from $v_i$ and present values at $v_j$, and to gate the signal from $v_i$ before it reaches $v_j$.

## K. INDEPENDENCE OF LISTS IN FIRST APPROXIMATION

If B is not presented to $\mathfrak{M}$, then in first approximation CA should be learnable without interference from B. (Not so in second approximation, since a signal could travel from C to B to A.) Similarly, if C is not presented to $\mathfrak{M}$, then BA should be learnable without interference from C, in first approximation. In other words, it should be possible to practice particular skills without activating the entire embedded vocabulary of behavioral units. Mathematically speaking, this means that all signals to each $v_j$ combine additively at $v_j$. Thus Eq. (6) becomes

$$\dot{x}_j(t) = -\alpha_jx_j(t) + \sum_{i=1}^{n} x_i(t - \tau_{ij})\beta_{ij}z_{ij}(t) + C_j(t) \tag{7}$$

The system (5) and (7) is a mathematically well-defined proposal for a

learning machine that uses only such general notions as linearity, continuity, and locality, and a mathematical analysis of how a machine can learn to predict B given A on the basis of practicing AB. Note that this system is nonlinear, notwithstanding our efforts to keep it as linear as possible, owing to constraints (I) and (J).

### L. Thresholds

One further modification of systems (5) and (7) is needed—namely, the introduction of signal thresholds. Here we introduce this modification directly to keep background noise down. A more fundamental analysis would introduce it by first analyzing the need in complex learning situations for inhibitory interactions—they shut off competing or irrelevant channels, among other tasks (Grossberg, 1970a, 1972a,b). Learning becomes difficult, if not impossible, without signal thresholds if inhibitory interactions exist, since the signs of all functions $x_i$ and $z_{jk}$ begin to oscillate in uncontrollable ways.

The modification can also be motivated by the following possible difficulty in (5) and (7). Small signals can possibly be carried around and around the network, thereby building up background noise and interfering with the processing of behaviorally important inputs. We therefore seek to eliminate the production of signals in response to small $x_i(t)$ values, in the most linear possible way. Thresholds do this for us. Letting $[\xi]^+ = \max(\xi, 0)$, we replace (5) and (7) by

$$\dot{x}_i(t) = -\alpha_i x_i(t) + \sum_{m=1}^{n} [x_m(t - \tau_{mi}) - \Gamma_{mi}]^+ \beta_{mi} z_{mi}(t) + C_i(t) \quad (8)$$

and

$$\dot{z}_{jk}(t) = -\gamma_{jk} z_{jk}(t) + \delta_{jk}[x_j(t - \tau_{jk}) - \Gamma_{jk}]^+ x_k(t) \quad (9)$$

where all $\Gamma_{mi}$ are nonnegative (usually positive) thresholds, and $i, j, k = 1, 2, \ldots, n$.

## III. Psychophysiological Interpretation

### A. Psychological Variables

The function $x_i(t)$ is called the $i$th *stimulus trace*: it responds to the *stimulus* $C_i(t)$. The function $z_{jk}(t)$ is called the $(j, k)$th *memory trace*: it records the pairing of successive events $r_j$ and $r_k$. Alternatively, $x_i(t)$ is called the $i$th *short-term memory trace*: it represents brief activation of the state $v_i$ either by inputs $C_i(t)$ or by signals from other states $v_j$. Similarly, $z_{jk}(t)$ is called the $(j, k)$th *long-term memory trace*: its record of past events can endure long after the short-term memory traces have decayed. *Transfer* from short-term memory to long-term memory denotes the operation whereby the $z_{jk}$'s are altered by the distribution of $x_i$'s. *Activation* of short-term memories via long-term memories denotes the operation whereby signals from a given set of $v_j$'s, modulated in the pathways $e_{jk}$ by the $z_{jk}$'s, activate a given pattern of $x_k$'s—for example, as in the subliminal activation of learned predispositions, or "psychological sets," in response to particular sensory cues.



The term $\Gamma_{jk}$ is the $(j, k)$th *signal threshold*: no signal is emitted by $v_j$ into $e_{jk}$ at time $t$ unless $x_j(t) > \Gamma_{jk}$. The vertex $v_j$ is said to *sample* $v_k$ at time $t$ if the signal received at $N_{jk}$ from $v_j$ at time $t$ is positive. The *signal strength* at $N_{jk}$ at time $t$ is defined by $B_{jk}(t) = [x_j(t - \tau_{jk}) - \Gamma_{jk}]^+ \beta_{jk}$. The constant $\beta_{jk}$ is a structural parameter called the *path strength* of $e_{jk}$. The $n \times n$ matrix $\beta = \| \beta_{jk} \|$ determines which directed paths between vertices exist, and how strong they are. Otherwise expressed, $\beta$ determines the "anatomy" of connections between all vertices.

### B. Neural Variables

A natural neurological interpretation of these variables is readily noticed. This interpretation does not claim uniqueness, however, because there exist only two kinds of variables, $x_i$'s and $z_{jk}$'s, at this level of theorizing, and these variables can at best represent averages of finer physiological or biochemical variables. The anatomical interpretation seems unambiguous: $v_i$ is a cell body (population), $e_{jk}$ is an axon (population), $N_{jk}$ is a synaptic knob (population), and the gap between $N_{jk}$ and $v_k$ is a (population of) synapse(s). Part of the physiological interpretation also seems inevitable: $x_i(t)$ is an average potential taken over all units in $v_i$ and over a brief time interval. The signal $B_{jk}(t)$ should correspondingly represent an average over individual signals in the axon(s) $e_{jk}$; it is therefore assumed to be proportional to the spiking frequency in $e_{jk}$. The interpretation of $z_{jk}(t)$ is more speculative. The process $z_{jk}(t)$ exists either in, or adjacent to, the synaptic knobs $N_{jk}$, and, by Eq. (8), $z_{jk}(t)$—coupled to the spiking frequency $B_{jk}(t)$—determines the signal from $N_{jk}$ to $v_k$. Thus it is natural to let $z_{jk}(t)$ correspond to the rate of transmitter production in $N_{jk}$, or to the sensitivity of postsynaptic sites at $v_k$ to fixed amounts of transmitter. The former interpretation is accepted herein for definiteness. Then Eq. (9) becomes a statistical law for transmitter production. Section IX shows that, even if $z_{jk}(t)$ is a presynaptic process, it is coupled to postsynaptic processes in $v_k$.

## IV. Outstars

### A. Pavlovian Choices

This section studies the smallest anatomy that can learn a choice by Pavlovian conditioning (Grossberg, 1968a, 1969b, 1970b). The anatomy is shown in Fig. 5. Figure 5*a* shows the smallest anatomy that can possibly learn AB, as opposed to the lists AC, AD, etc.; that is, it can learn the choice B given A, as opposed to C given A, D given A, etc. Figure 5*b* interprets the same anatomy using the Pavlovian concepts CS, UCS, and UCR. Figure 5*c* replaces these particularized notations by a purely abstract labeling of states using indices. The cell population with cell body $v_1$ emits an axon which breaks up into axon collaterals whose synaptic knobs appose the UCS-activated cells $\mathfrak{B} = \{v_i;\ i = 2, 3, \ldots, n\}$. Figure 5*d* represents this system in a more symmetric fashion, which suggests the name *outstar* for it. Here $v_1$ is called the *source* of the outstar. Each $v_i$, $i \neq 1$, is called a *sink* of the outstar, and the set $\mathfrak{B}$ of all sinks is called the *border* of the outstar.

The outstar equations can readily be derived from Eqs. (8) and (9). The main constraint is that only $v_1$ can send signals to other cells $v_i$. Hence $\beta_{jk} = 0$ unless $j = 1$ and $k \neq 1$. We find the equations

$$\dot{x}_1(t) = -\alpha_1 x_1(t) + C_1(t) \tag{10}$$

$$\dot{x}_i(t) = -\alpha_i x_i(t) + \beta_{1i}[x_1(t - \tau_{1i}) - \Gamma_{1i}]^+ z_{1i}(t) + C_i(t) \tag{11}$$

and

$$\dot{z}_{1i}(t) = -\gamma_{1i} z_{1i}(t) + \delta_{1i}[x_1(t - \tau_{1i}) - \Gamma_{1i}]^+ x_i(t) \tag{12}$$

where $i = 2, 3, \ldots, n$.

### B. Unbiased Outstars

First we consider outstars in which no choice $r_i$, $i \neq 1$, is preferred above any others because of asymmetric choices of system parameters. In other words, we make the following restrictions on these parameters: (1) set all time lags $\tau_{1i}$ equal to $\tau$; (2) set all thresholds $\Gamma_{1i}$ equal to $\Gamma$; (3) set all decays rates $\alpha_i(\gamma_{1i})$ equal to $\alpha(\gamma)$; and (4) set all interaction weights $\beta_{1i}(\delta_{1i})$ equal to $\beta(\delta)$. The unbiased outstar therefore satisfies the Eqs. (10),

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta[x_1(t - \tau) - \Gamma]^+ z_{1i}(t) + C_i(t) \tag{13}$$

and

$$\dot{z}_{1i}(t) = -\gamma z_{1i}(t) + \delta[x_1(t - \tau) - \Gamma]^+ x_i(t) \tag{14}$$

where $i = 2, 3, \ldots, n$.



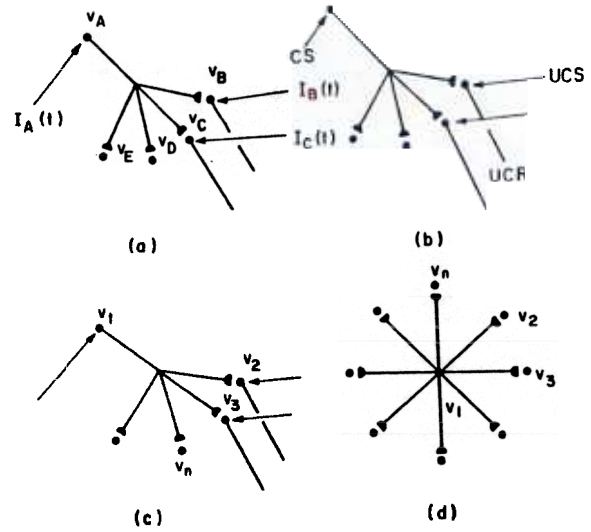


Fig. 5. Outstar: minimal network capable of classical conditioning.

Inspection of this system readily shows that it is a linear system of equations with variable coefficients. Indeed, integration of Eq. (10) yields

$$x_1(t) = x_1(0)e^{-\alpha_1 t} + \int_0^t C_1(v)e^{-\alpha_1(t-v)}\,dv$$

Hence the term $[x_1(t - \tau) - \Gamma]^+$ in Eqs. (13) and (14) is a known function of time. System (13) and (14) is therefore a special case of the following more general system of equations:

$$\dot{x}_i(t) = A(t)x_i(t) + B(t)z_{1i}(t) + C_i(t) \tag{15}$$

and

$$\dot{z}_{1i}(t) = D(t)z_{1i}(t) + E(t)x_i(t) \tag{16}$$

where $A(t)$, $B(t)$, $D(t)$, and $E(t)$ are continuous functionals of $t$, and moreover $B(t)$ and $E(t)$ are nonnegative. (A functional is a mapping from functions to real numbers. A functional can depend on system variables, evaluated at past times, in a complicated way.) A rigorous mathematical analysis of this class of systems has been carried out. Below we list in intuitive terminology some of the formal properties that have been found.

### C. Spatial Pattern Learning

What is the most general UCS whose UCR can be reproduced by a CS after Pavlovian conditioning in an unbiased outstar? The answer is a

FIG. 6. Classical conditioning of a spatial pattern on a grid.

*spatial pattern* (or *picture*). This we define to be any UCS of the form $C_i(t) = \theta_i C(t)$, $i = 2, 3, \ldots, n$, such that $\theta_i \geq 0$ and $\sum_{k=2}^{n} \theta_k = 1$. In other words, $\theta = \{\theta_i: i = 2, 3, \ldots, n\}$ is a fixed, but otherwise arbitrary, probability distribution, and $C(t)$ is a nonnegative and continuous function of $t$. A spatial pattern is the unit of long-term memory in an embedding field.

The intuitive meaning of this definition will be illustrated by an example. Actually the concept of spatial pattern will arise in more varied circumstances than this example might suggest. Consider Fig. 6. Suppose that an arbitrary picture in shades of black, white, and gray is shown on the region $\mathfrak{R}$. We want $\mathfrak{M}$ to be able to reproduce this picture on $\mathfrak{R}$, with an arbitrarily good spatial resolution, by Pavlovian conditioning.

How is the spatial resolution prescribed? Suppose that $m$ cells of $\mathfrak{M}$ are embedded in $\mathfrak{R}$, and that each cell receives an input proportional to the intensity of the picture at its position. For definiteness, imagine that the $m$ cells are arranged in a rectangular grid in $\mathfrak{R}$. As $m$ is increased to ever larger values, the density of the cells in $\mathfrak{R}$ increases, as does the accuracy with which the picture is represented in $\mathfrak{M}$ by these cells. We shall let these cells be the border, $\mathfrak{B}$, of an outstar. The mathematical results on learning by outstars hold for any $n = m + 1$, or any spatial resolution.

Now imagine a fixed picture, such as the Mona Lisa in shades of gray, shown on $\mathfrak{R}$. We can vary the *total* intensity of the light which illuminates the picture without changing the picture itself. The total intensity can be steady (and bright or dim), or can flicker between broad physiological limits, without changing our impression that the Mona Lisa is still being presented. In other words, the *relative* intensity of light, not its absolute intensity, characterizes the picture. Only the relative intensity of the picture is constant through time. The constant relative intensity at $v_i$ is denoted by $\theta_i$. The total intensity, which can fluctuate in time, is $C(t)$. In other words, the fact that outstars can learn the weight, $\theta_i$, means that they can pick out the "relative figure to ground" of an input pattern. The outstar can learn such a pattern no matter how we interpret the border, $\mathfrak{B}$, to which it is attached. For example, the border can consist of motor control cells, interneurons, cells in any sensory cortex, etc.

These assertions are made precise by studying the relative traces, or



*pattern variables*, $X_i = x_i(\sum_{k=2}^{n} x_k)^{-1}$ and $Z_{1i} = z_{1i}(\sum_{k=2}^{n} z_{1k})^{-1}$. Mathematical analysis shows that the pattern weight $\theta_i$ attracts $X_i$, while $X_i$ and $Z_{1i}$ mutually attract each other. Consequently, the relative memory trace $Z_{1i}$ is attracted toward ("encodes") the pattern weight $\theta_i$. On recall trials, an input to $v_1$ ("presenting A") creates an equal signal to each $N_{1i}$. In $N_{1i}$, the signal is multiplied by $z_{1i}$, which is proportional to $\theta_i$. Thus the input to $v_i$ is proportional to $\theta_i$. The learned pattern is hereby reproduced on the border $\mathfrak{B}$ by an input to the source cell $v_1$.

Spatial pattern learning by an outstar has the following properties (Grossberg, 1968a, 1969b, 1970b).

### 1. *Practice Makes Perfect*

The more $r_1 r_i$ is practiced, the better can $\mathfrak{M}$ predict $r_i$ in response to $r_1$. This learning can be "all-or-none"—occurring in one trial—or "gradual"—requiring several trials. In an outstar, learning rate is determined by CS and UCS input rate, intensity, relative timing, the number of response alternatives, and related factors. These factors influence both the rate with which $Z_{1i}$ approaches $\theta_i$ and the size of $z_{1i}$. In more general anatomies, the learning rate of a given item in a list of events depends on list position, or more generally on the context of other events in which the item occurs (cf. Section VII). For example, in serial learning of a long list presented at a rapid rate, the items at the two ends of the list might be quickly learned, whereas the items near the middle of the list might not be learned at all on the first few trials.

### 2. *Overt Practice Unnecessary*

The machine $\mathfrak{M}$ can remember without overt practice. The potentials and thus the outputs from $\mathfrak{M}$ can be zero during memory intervals without destroying the memory; that is, each $Z_{1i}$ remains constant. In fact, positive potentials (in particular, "reverberations" among the vertices) can destroy the memory in certain anatomies (Grossberg, 1968b). One must also distinguish perfect memory of pattern weights $Z_{1i}$ from perfect performance. For example, in Eq. (14), $z_{1i}$ can exponentially decay even if $Z_{1i}$ remains constant. If $z_{1i}$ decays to the level of network noise, then the memory is essentially zero.

### 3. *Recall Preserves Memory*

Item $r_i$ can be recalled in response to $r_1$ as often as one pleases without destroying the memory of $r_1 r_i$; that is, $Z_{1i}$ remains constant during recall trials. In fact, recall of $r_i$, given $r_1$, can "potentiate" the memory of

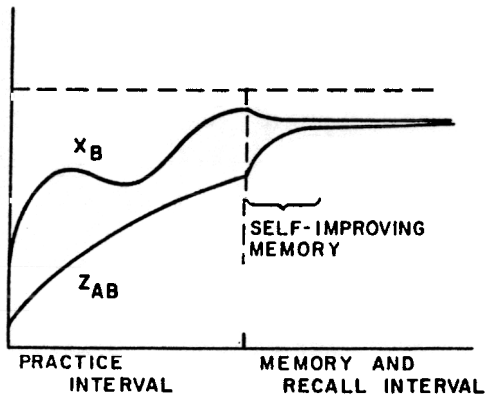


FIG. 7. Self-improving memory due to coupling of STM and LTM traces.

$r_1r_i$; that is, $z_{1i}$ can grow. There exist anatomies in which this is false: the very act of recall tends to destroy the memory (Grossberg, 1968b); these anatomies usually reverberate signals in closed loops.

### 4. *Self-Improving Memory*

Self-improving memory, or *reminiscence*, exists. For example, let two outstars, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, practice $r_1r_i$ the same number of times. Let recall occur $T_i$ time units after practice ceases in $\mathfrak{M}_i$, $i = 1, 2$. For certain choices of $T_1 < T_2$, recall is better in $\mathfrak{M}_2$ than in $\mathfrak{M}_1$. See Osgood (1953, pp. 509–513) for a discussion of an analogous experimental phenomenon. Figure 7 illustrates this phenomenon. It is due to a coupling between STM traces and LTM traces.

### 5. *Contour Enhancement*

After practice of a spatial pattern ceases, the memory of dark (bright) regions of the pattern can become darker (brighter). The mechanisms for contour enhancement and reminiscence are the same in an outstar. See Grossberg (1972b) and Ratliff (1965) for a discussion of contour enhancement due to lateral inhibition.

### 6. *Error Correction*

All errors can be corrected. Even after $r_1r_i$ is learned, $r_1r_j$ can be learned instead. The rate of learning $r_1r_j$ can depend on such factors as the prior level of $r_1r_i$ performance and on the total number of response alternatives.



### 7. *Several Memory and Recall Modes*

For suitable choices of the coefficients $A(t)$, $B(t)$, $D(t)$, and $E(t)$ in Eqs. (15) and (16), different properties of memory or recall can be achieved. Each choice has a distinct physiological interpretation. One can, for example, achieve:

($a$) Perfect memory, even during recall trials (Grossberg, 1970b); let

$$B(t) = \beta[x_1(t - \tau) - \Gamma]^+\left[\sum_{k=2}^{n} z_{1k}(t)\right]^{-1} \tag{17}$$

This describes a purely "interference" theory of forgetting: All forgetting is due to active relearning of new sequences.

This choice of $B(t)$ does *not* yield an interference theory in all anatomies. For example, if each $v_i$ sends signals to all vertices $v_k$, or to all vertices but itself, then a "phase transition" can occur (Grossberg, 1968b, 1969c). Given suitable numerical parameters, a learned pattern will be forgotten; given other parameters, it will be remembered. Exactly what is forgotten depends on the anatomy of the network. One can pass from the forgetting phase to the remembering phase by (say) speeding up axonal signals at a critical time; such an operation can "imprint" the pattern that exists at the critical time. See Section VI,G. This example dramatizes the fact that one cannot generally infer the global properties of a network from its local properties.

($b$) Exponential decay of memory, at any prescribed rate (Grossberg, 1970b); let $C(t) = -\gamma$. Even though $z_{1i}$ can spontaneously decay, the relative traces $Z_{1i}$ are changed only by "interference" due to new learning, or by reminiscence, as in property (4). The net decay rate of $z_{1i}$ itself is not always $\gamma$. This rate can be slowed down, or even reversed, by recall trials, by "spontaneous" rhythmic inputs to $v_1$ during memory intervals, by reminiscence effects, etc. Again, a local property—this time a decay rate—is not necessarily the global one.

($c$) Perfect memory until recall trials, followed by possible extinction of memory during recall if the prediction is not rewarded or retrained (Grossberg, 1970b); let $D(t) = -\gamma E(t)$. Again an interference theory of forgetting holds for the relative traces $Z_{1i}$, but not necessarily in all anatomies.

These examples point out that important properties of learning are invariant under changes that allow many variations in the details of learning and performance. Speaking mathematically, the pattern variables $X_i(t)$ and $Z_{1i}(t)$ have the same limiting and oscillatory possibilities given various choices of the coefficients $A(t)$, $B(t)$, $D(t)$, and $E(t)$. These coefficients determine the transient motions of the system, including learning rates.

### 8. *Stimulus Sampling*

Stimulus sampling theory is a purely behavioral theory that has successfully described various learning data using probability models (Atkinson and Estes, 1963). A physiological mechanism of stimulus sampling and a physiological interpretation of stimulus sampling probabilities exist in embedding fields. The relative memory traces $Z_1 = (Z_{12}, Z_{13}, \ldots, Z_{1n})$ are attracted toward the pattern weights $\theta = (\theta_2, \theta_3, \ldots, \theta_n)$ only at times when the synaptic knobs, $N_{1i}$, receive CS-activated spikes from $v_1$. This is the property of "stimulus sampling" in an outstar: $v_1$ samples the patterns playing on $\mathcal{B}$ by emitting signals at prescribed times. The relative memory traces, $Z_1$, which form a probability distribution at each time $t$, are the "stimulus sampling probabilities" of an outstar (Grossberg, 1970b). Whenever $v_1$ samples $\mathcal{B}$, the memory traces in its synaptic knobs begin to learn the spatial pattern playing on $\mathcal{B}$ at this time. If a sequence of patterns (that is, a space–time pattern) plays on $\mathcal{B}$ while $v_1$ is sampling, then $v_1$'s synaptic knobs learn a weighted average of all the patterns, rather than any single spatial pattern. Thus if an outstar samples $\mathcal{B}$ while a long sequence of spatial patterns reaches $\mathcal{B}$, then after sampling terminates, the sampling probabilities, $Z_1$, can be different from any one of the spatial patterns. On recall trials, a CS input to $v_1$ creates equal signals in the axons $e_{1i}$. These signals flow down to the $N_{1i}$. In $N_{1i}$, the signal interacts with the memory trace $z_{1i}$ to reproduce at the cell $v_i$ an output proportional to $Z_{1i}$. In this way, recall trials reproduce at $\mathcal{B}$ the weighted average of sampled patterns that was encoded on learning trials.

### 9. *Oscillatory Inputs and Monotonic Response*

When $r_1 r_i$ is practiced on successive trials, the inputs $C_1(t)$ and $C_i(t)$ are highly oscillatory in time. Yet increased practice yields the impression of a steady increase in learning (see Fig. 8). The probabilities $Z_{1i}$ bridge the gap between oscillatory inputs and monotonic learned response.

### 10. *Eidetic Memory*

An outstar is capable of *eidetic* memory. This remarkable phenomenon has been tested by using human subjects in the following ingenious way. Two pictures are constructed by computer from 10,000 randomly distributed black and white dots. These pictures conceal a figure in depth that can be seen only when the pictures are viewed binocularly (Julesz, 1964). An eidetic woman studies the first picture with one eye on day 1 of the experiment and returns the next day to study the second picture with the other eye. She then identifies the concealed figure (B. Julesz, personal



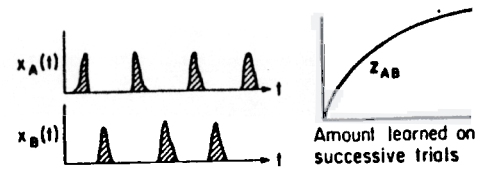


FIG. 8. Oscillatory inputs due to repetitive A-then-B presentation and monotonic response of stimulus sampling probability.

communication, 1970)! To accomplish this, she must presumably be able to conjure up in her mind's eye an almost perfect replica of the 10,000 dots shown on the previous day. In short, textural memory with an enormous storage capacity is possible.

A single cell in our networks can do this formally. Let an outstar (or a cluster of outstars that fire in unison) send axon collaterals to the correct visual representation area. If the network can activate the source cell(s) at will, then it can learn the first picture to an arbitrary degree of accuracy on day 1. On day 2, if it again activates the source cell, the internally produced representation of the first picture will interact with the externally produced representation of the second picture to produce the binocular effect of a figure in depth (see Fig. 9). Several properties of this mechanism are of interest.

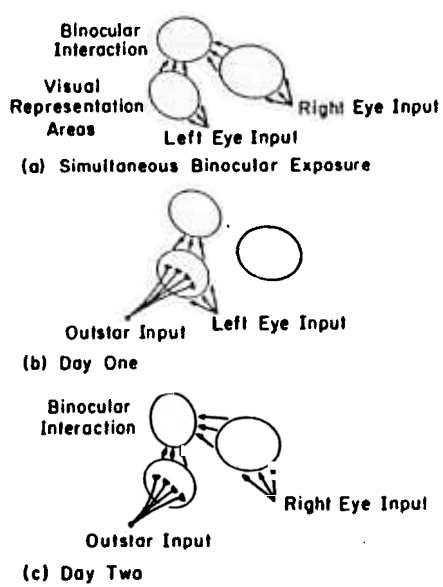


FIG. 9. Eidetic memory using outstar sampling.

(*a*) To learn $10^m$ pictures playing on $10^n$ visual cells, one needs no more than $10^m + 10^n (\leq 2 \times 10^{\max(m,n)})$ cells, not $10^{m+n}$ cells, as is occasionally claimed. In fact, it will later be shown that to learn $10^m$ moving pictures (space–time patterns, such as piano sonatas or dances) playing on $10^n$ cells of any kind, one needs no more than $10^m + 10^n$ cells. In principle, one could learn $10^{11}$ dances playing on $10^{11}$ motor control cells without using as many as $10^{12}$ cells, and our brains are thought to contain no fewer than $10^{12}$ cells. One could also learn a new pattern playing on $10^6$ cells every second for sixty years using fewer than $2 \times 10^{12}$ encoding cells. The networks contain more than enough storage capacity. None of us seems to know how to do $10^{11}$ complex acts, however. Hence we must ask what the extra cells are doing. Section V will begin to address this question.

(*b*) To learn eidetically as in Fig. 9, there must exist cells—other than cells leading progressively from the retina itself—that send axons to the visual representation areas. Further evidence for the existence of such cells has been acquired by studying epileptics (Penfield, 1958). An electrode in the temporal lobe of man can vividly activate a sequence of perhaps very old memories, including visual and auditory memories. Discontinuing electrode current while the sequence is being recalled can stop recall. Reapplying current at the same point can reinitiate recall of the same sequence. These data suggest that the cells being sought might project from the temporal lobes to visual and/or auditory representation areas.

(*c*) The blessing of eidetic memory also carries with it a possible liability. Suppose that the visual representation areas received a continual barrage of nonvisual inputs which were not synchronized with visual inputs to produce functionally useful results. Then hallucinations and other internal visual experiences could continually pop into our minds against our will. To prevent this, such cells should be forbidden from firing to the visual areas unless there exist functionally desirable stimulus conditions for visual learning or recall. The ability to fire the source cell at will seems to be a more remarkable phenomenon than the existence of textural memory per se, since it resembles the ability to hallucinate specific subject matter at will.

### 11. *Response Generalization: Variable Performance Velocities*

Suppose that the UCR sends signals to muscles which contract at a rate proportional to the signal. Let the UCR be a spatial pattern; that is, the UCR creates fixed relative contraction rates of the various muscle groups. An outstar that learns to contract these muscles at a given total velocity can also contract them—in the same pattern—at many other total velocities. This form of "response generalization" is the output version of the "stimulus generalization" property of being able to learn the "relative figure to ground" in an input pattern. Not all motor patterns are spatial patterns, however, and this property is modified when more complicated motor tasks are imposed, as in Section V.



## V. Avalanches

### A. Ritualistic Learning of Space–Time Patterns

This section studies the following question: What is the minimal number of cells needed to encode the memory of an arbitrarily complicated space–time pattern, such as a piano sonata or a dance? The answer is: One! What could be more "minimal"? Yet this answer creates a paradox. If one cell can encode a whole dance, and our brains contain at least $10^{12}$ cells, why doesn't anyone know $10^{12}$ (or even $10^4$) dances? What do the extra cells do?

Encoding a space–time pattern with one cell has a severe limitation: Performance is ritualistic, or by rote. Once performance of the "dance" begins, the entire dance must be completed, even if the stage on which the dance is being performed is consumed by flames as the dance progresses. In other words, such a system is insensitive to environmental feedback; it cannot adapt to changing environmental demands once the performance of an act begins. Once we note how to encode a space–time pattern without feedback, we shall also readily see how to begin construction of systems that are sensitive to feedback. Such systems will require many more than one cell to encode the entire pattern.

Study of systems that perform with little feedback is not of purely academic interest, however. There exist examples of such performance throughout the phylogenetic kingdom. For example, the seagoing mollusk Tritonia has individual, large cells, with extensively branched axons, whose direct electrical stimulation causes a well-organized swimming escape response (Willows, 1968). Clearly, given such individual cells, it is crucial that they fire only at appropriate times. For example, Tritonia would starve if it "escaped" whenever it approached a source of food. Nonetheless, Tritonia can escape from predators, such as starfish, with considerable reliability. Thus certain characteristic stimuli at Tritonia's periphery can create inputs to its swimming escape cells, but inappropriate inputs cannot. Such facts motivate the construction of networks that can selectively filter environmental inputs on their way to prescribed control cells (Grossberg, 1970a, 1972a,b, Hubel and Wiesel, 1968).

Other organisms also have individual cells capable of controlling well-organized behavioral acts. These include insects (Dethier, 1968, p. 8), and crayfish (Kennedy, 1968). On a higher level, the ring dove performs a

ritualistic sequence of acts during its reproductive cycle (Lehrman, 1965). Successive stages of this sequence are triggered by the previous stage and a well-defined combination of exteroceptive and interoceptive stimuli. The maternal behavior of the rat also involves a characteristic sequence of ritualistically organized acts (Thomas *et al.*, 1968, pp. 265–273). Even man is capable of performing complex sequences of acts without the benefit of continuous feedback. For example, a cadenza can be played by a skilled pianist so rapidly that motor feedback cannot possibly determine the next note to be played (Grossberg, 1969b, 1970b; Lashley, 1951). On the other hand, as one ascends the phylogenetic ladder, one finds that ever more subtle types of feedback can influence behavior. For example, the pianist can try to escape from a concert hall before it burns down, and can modify his performance of a piece in exquisitely subtle ways.

### B. Sequential Sampling

Given a finite collection of cells $v_i, i \in I$ ($I$ some set of integers), suppose that an arbitrary nonnegative and continuous input, $C_i(t)$, perturbs $v_i$. Consider the weights $\theta_i(t) = C_i(t)[\sum_{k \in I} C_k(t)]^{-1}$ as they fluctuate in time. Can we learn these weights to an arbitrary degree of accuracy? We can do so by using a collection of sequentially activated outstars if we invoke three mechanisms: (1) stimulus sampling; (2) brief signals from the CS-activated cell body; and (3) an anatomy in which each CS-activated outstar sends an axon collateral to each UCS-activated cell, $v_i$.

To see this, first note that $\theta_i(t)$, as a continuous function of $t$, can be arbitrarily well approximated by the discrete sequence

$$\{\theta_i(0), \theta_i(\zeta), \theta_i(2\zeta), \theta_i(3\zeta), \ldots, \theta_i(N_\zeta - 1)\}$$

of its values, if the positive number $\zeta$ is chosen sufficiently small; that is, the "moving picture" is replaced by a sequence of $N_\zeta$ "still pictures." Suppose that a sequence, $\mathfrak{M}_j$, of outstars is given, $j = 1, 2, \ldots, N_\zeta$, such that (1) each outstar sends one axon collateral to each cell $v_i$, $i \in I$, and (2) the synaptic knobs of $\mathfrak{M}_j$'s axon collaterals are active only during an interval of time $[(j-1)\zeta, (j-1)\zeta + \Delta\zeta]$. If $\Delta\zeta$ is sufficiently small, then the pattern weights, $\theta_i(t)$, change arbitrarily little from their values $\theta_i[(j-1)\zeta]$ during the time interval $[(j-1)\zeta, (j-1)\zeta + \Delta\zeta]$. Hence $\mathfrak{M}_j$ can learn the spatial pattern with weights $\theta_i[(j-1)\zeta]$ to an arbitrary degree of accuracy. The outstar $\mathfrak{M}_j$ samples only this pattern, by the property of stimulus sampling.

How can these sampling intervals be guaranteed? Simply let a cell body, $v_1$, send out a long axon, and attach the outstar $\mathfrak{M}_j$ at the axonal position which is excited by a signal emitted from $v_1$ at $(j-1)\zeta - \tau$ time units earlier, where $\tau$ is the time needed for signals to travel from $\mathfrak{M}_j$ to any $v_i$ (see Fig. 10). Such a system is called an *outstar avalanche*, or an *avalanche*, by analogy with avalanche conduction in the parallel fibers of the cerebellum (Grossberg, 1969b, 1970b). Physiologically, it is a cell whose axon emits sequential clusters of axon collaterals which converge on the common cells $v_i$, $i \in I$. Performance of the pattern is elicited by a signal from $v_1$, which successively activates the outstar-encoded spatial pattern approximations to the space–time pattern on the cells $v_i$ every $\zeta$ time units.



Note that the avalanche has the minimal number of formal degrees of freedom needed to learn the pattern perfectly, given a prescribed spatial and temporal resolution of the inputs: the number $|I|$ of cells $v_i$ determines the spatial resolution of the inputs $C_i(t)$, and the number $N_\zeta$ of time intervals determines the temporal resolution in memory that is desired. The minimal number of formal degrees of freedom is $|I| N_\zeta$, which is also the number of axon collaterals in the avalanche.

A sample set of equations for an avalanche is stated below. Let $x_1$ be the potential of $v_1$, and let $x_i$ be the potential of $v_i$, $i \in I$. Let $z_{ji}$ be the transmitter in the axon leading from the $j$th outstar to the cell $v_i$. Then system (8) and (9) becomes

$$\dot{x}_1 = -\alpha_1 x_1 + C_1 \tag{18}$$

$$\dot{x}_i = -\alpha x_i + \beta \sum_{k=1}^{N_\zeta} [x_1(t - (k-1)\zeta) - \Gamma]^+ z_{ki} + C_i \tag{19}$$

and

$$\dot{z}_{ji} = -\gamma z_{ji} + \delta[x_1(t - (j-1)\zeta) - \Gamma]^+ x_i \tag{20}$$

where $i \in I$, $j = 1, 2, \ldots, N_\zeta$. Suppose that $[x_1(t) - \Gamma]^+$ is positive in an interval whose duration is shorter than $\zeta$. Then at every time $t$, at most one term in the sum

$$\sum_{k=1}^{N_\zeta} [x_1(t - (k-1)\zeta) - \Gamma]^+ z_{ki}$$

is positive. At times when the positive term corresponds to $k = K$, then

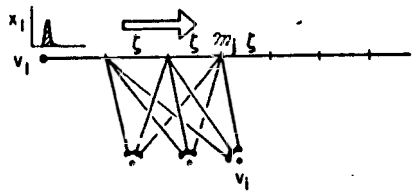


Fig. 10. Sequential sampling of a space–time pattern by an avalanche, or command cell.

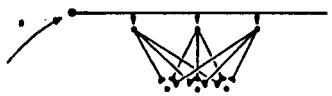

FIG. 11. A command cell that sequentially activates outstar interneurons.

Eq. (19) becomes

$$\dot{x}_i = -\alpha x_i + \beta[x_1(t - (K-1)\zeta) - \Gamma]^+ z_{ki} + C_i \qquad (21)$$

and when $k = K$, the system (18), (20), and (21) is an outstar.

There exist variations on this theme. For example, a single cell, $v_i$, can give off sequential axon collaterals to a series of outstars. Figure 10 is then replaced by Fig. 11, in which the outstars are interneurons between $v_1$ and $v_i$, $i \in I$. In this anatomy, several different command cells can sample the same outstar. Perhaps the most abstract anatomical arrangement is that given by Fig. 12, which shows that the local anatomy alone of the system does not necessarily disclose its function. In Fig. 12*a*, a parallel series of axons gives off regular axon collaterals to a rectangular lattice of cells. What this system learns depends entirely on what inputs are sent to it. For example, in Fig. 12*b*, synchronized CS inputs reach the first three sampling cell bodies, and (perhaps differently) synchronized CS inputs reach the next three cell bodies. Figure 12*c* draws the equivalent avalanches for this case. Next one must determine the distribution of UCS inputs. If, for

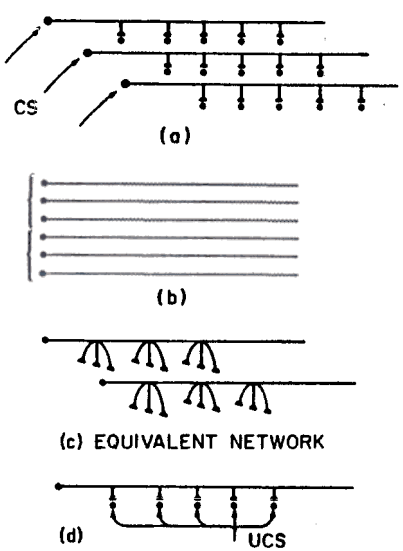


FIG. 12. An anatomy in which input symmetries determine equivalent avalanches.



example, all cells in a row parallel to the axons receive the same input, then the system of Fig. 12*d* can learn space–time patterns much as the system of Fig. 10 does. If, however, the UCS inputs are the same in each row perpendicular to the axons, then one learns only redundant copies of a sequence of perhaps uncorrelated events. In other words, the *symmetries of the input mechanisms* determine what the equivalent learning network is; the local anatomy itself need not reveal these symmetries. Various other anatomical variations are considered in Grossberg (1970b).

Avalanches of avalanches, or avalanches of avalanches of avalanches, etc., can readily be constructed. For example, a given cell population can control motions of a finger, a higher cell population in the hierarchy can control motions of all fingers in a hand, a still higher cell population can control motions of both hands, etc. Inputs can, in principle, enter this hierarchy at any level to activate a prescribed population.

## C. Sensitivity to Feedback: Command Cells as Arousal Sources

How can an avalanche be modified so that sequential performance can be stopped and switched to more urgent behavioral modes? Clearly this cannot be done in Fig. 10 because the signal propagates down the entire axon once it is emitted by $v_1$. To prevent this, successive outstars can be separated by interpolated cells, as in Fig. 13. Immediately we have gone from one encoding cell to $N_i$ such cells. These extra cells will provide no advantage unless a given cell, $v_j$, requires more than a signal from $v_{j-1}$ in order to fire. Namely, it might also require a simultaneous input from another part of the network which designates that sequential performance of the given act is still desirable—for example, an "arousal" or "positive incentive motivational" input (cf. Section VIII). The cell $v_j$ should also be unable to fire if it receives a $v_{j-1}$ signal along with an inhibitory signal from elsewhere in the network that designates the undesirability of continued

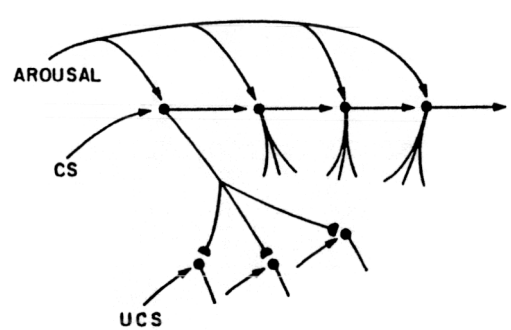


FIG. 13. A command cell as a nonspecific arousal source supporting sequential sampling.

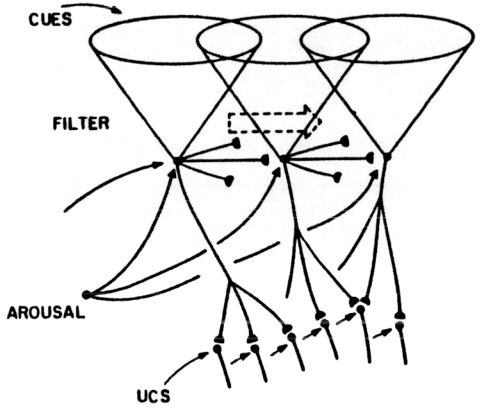


FIG. 14. An avalanche whose successive sampling sources are selected by cues and then learned.

sequential performance—for example, a "suppressor" or "negative incentive motivational" input; cf. crayfish swimmerets (Stein, 1971, p. 317).

Not every competing stimulus should be able to terminate ongoing performance. Only those inputs that have greater "significance" to the network should have this privilege. What are "significant" inputs? This question naturally leads one to discuss the question: What are rewarding or punishing inputs? In short, what is reinforcement? How does it influence the network's arousal level (Grossberg, 1971b, 1972c,d)?

Note that performance rate can be varied in Fig. 13. Each successive cell, $v_j$, can have its reaction time (that is, the time lag between input onset and onset of output signal) decreased, either by increasing the signal from $v_{j-1}$ or by increasing the arousal input. Performance rate can thus be continuously modified by continuously varying the arousal level of the machine. That is, even if the avalanche-like anatomy encodes the same sequence of events (the same "information") on successive performance trials, nonetheless the arousal level of the machine (its "energy" level) can modify details of performance. The same argument holds if no learning occurs at the synaptic knobs, and the avalanche anatomy merely controls the performance of a sequence of motions. Note that modifying the arousal level does not require feedback from the avalanche outputs. Successive outstars can be sampled much faster—and at variable rates—than feedback need permit.

Until this point, we have considered avalanches whose successive outstars are predetermined by the network anatomy. In general, this need not be true. Successive links can also be determined by sensory and motor cues, including feedback cues. Then one is led to ask: How are these cues filtered through cell populations with selective response profiles to sequentially activate particular outstar source cells? If the sequential activation of outstar sources is not innately determined by the anatomy, one must also ask: How does the $j$th outstar (population) in the avalanche form sequential connections with the $(j + 1)$st outstar (population) in the avalanche? (See Fig. 14.) Other questions readily suggest themselves. How can brief sampling pulses be guaranteed in the avalanche in response to possibly temporally prolonged sensory cues? Such pulses are needed to achieve accurate sampling of spatial approximants to a space–time pattern, as well as precise performance (Grossberg, 1970a). How can more than one sampling pulse be prevented from passing down the avalanche at any given time, again to achieve accurate sampling and performance? This requires the introduction of inhibitory signals, activated by the outstars, and descending toward the input sources. In short, the expansion of ritualistic avalanches to achieve responsiveness to environmental feedback imposes a definite teleology on our later constructions. Some of these constructions yield mechanisms of pattern discrimination, and in particular an analysis of various uses for nonspecific inhibitory interneurons (Grossberg, 1970a, 1972a, 1973).



As learning and performance become less ritualistic in an avalanche, the complexity of the total input to each of its outstar sources increases. The total input can be a sum of a rapidly fluctuating arousal input, an input from a complex hierarchy of sensory filters, an input from a previous outstar source that was itself perturbed by a complex input, etc. Thus we seek assurances that learning can occur even if the source is perturbed by very general inputs. The next section provides such assurances in a rigorous mathematical setting. Holographic theories of memory, which depend on the existence of precisely regulated periodic sampling sources, depart heavily from the spirit of this discussion.

## VI. Arbitrary Anatomies and Generalized Physiological Laws

### A. One Level in a Hierarchy

When an avalanche is modified to permit feedback adaptations, the cells $v_i$, $i \in I$, can be sampled by many cells $v_j$, $j = 1, 2, \ldots, N_i$. Below we therefore study the following question: Under what circumstances can a collection of cells $\mathcal{A} = \{v_j, j \in J\}$ sample a collection of cells $\mathcal{B} = \{v_i, i \in I\}$ in such a fashion that simultaneous sampling of $\mathcal{B}$ by different cells in $\mathcal{A}$ does not irrevocably bias what these cells learn? We shall find that this is possible, given any finite number of cells $\mathcal{A}$ and $\mathcal{B}$, under very weak conditions. The relevant theorems (Grossberg, 1969d, 1971c, 1972b) hold even if

the cells $\mathfrak{A}$ fire out of phase and in response to wildly oscillatory, and mutually uncorrelated, inputs. Thus the inputs to cells $\mathfrak{A}$ can be constructed from the outputs emitted by cells at a previous stage of learning or other preprocessing, and the outputs from $\mathfrak{B}$ can be used to construct inputs to a later stage of cells. In this way, a hierarchy of learning cells can be constructed. The theorems study one level in such a hierarchy in detail. If such a mechanism evolved at a given time, it could be adapted to any later specialization.

### B. A General Class of Systems

The equations that govern one level of this hierarchy can be substantially generalized beyond Eqs. (8) and (9) by weakening some linearities in these equations without changing their general form. These equations are defined by

$$\dot{x}_i = A_i x_i + \sum_{k \in J} B_{ki} z_{ki} + C_i \tag{22}$$

and

$$\dot{z}_{ji} = D_{ji} z_{ji} + E_{ji} x_i \tag{23}$$

$i \in I, j \in J$, where $A_i$, $B_{ji}$, $D_{ji}$, and $E_{ji}$ denote continuous functionals, not necessarily linear, with all $B_{ji}$ and $E_{ji}$ nonnegative. The input functions and initial data are chosen nonnegative and continuous. Mathematical analysis of Eqs. (22) and (23) shows that the classification of limiting and oscillatory possibilities for the pattern variables of these systems is invariant under broad changes in functionals, much as in the study of Eqs. (15) and (16). As in that situation, transient motions of the systems can be altered by changes in functionals, and a proper choice of functionals (including anatomy) must be made to guarantee efficient real-time learning of particular tasks. The invariance properties show that the systems are very stable and can be adapted to many particular situations. Below are reviewed some physically relevant choices of these functionals.

As in the case of Eqs. (15) and (16), the long-term memory decay functional, $D_{ji}$, can be chosen to guarantee a variety of forgetting possibilities. The choice of performance functional $B_{ji}$, as in Eq. (17), can also influence how decay due to $D_{ji}$ shows up in network response to inputs. Other useful choices of these functionals are listed below.

#### 1. *Now Print Signals of Shunting Type*

Suppose that a sequence of spatial patterns perturbs the cells $\mathfrak{B}$. There exist mechanisms that can quickly accelerate learning of the patterns which arrive during prescribed time intervals. These intervals can heuristically be called Now Print intervals (Livingston, 1967, p. 132). Such mechanisms



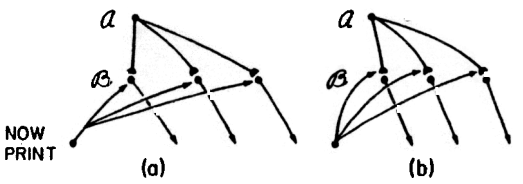


Fig. 15. Nonspecific arousal as a shunt of potentials or signals.

can be activated by arousal inputs that are turned on by the occurrence of significant events. The first mechanism works by sending synchronized signals to all cells in $\mathfrak{B}$. These signals then interact multiplicatively with (or "shunt") the potentials $x_i$ (see Fig. 15*a*).

Consider, for example, the system

$$\dot{x}_i = [-\alpha + f(t)]x_i + \theta_i C$$

where $C$ is a constant, $0 \leq f(t) < \alpha$, and $x_i(0) = 0$. Let $f(t)$ be constant in the interval $[0, T]$. Then

$$x_i(t) = \theta_i \left[ \frac{C}{\alpha - f} (1 - e^{-(\alpha - f)t}) \right]$$

for $t \in [0, T]$. The function

$$g_t(w) = \frac{1}{w} (1 - e^{-wt})$$

is, for fixed $t \geq 0$, a monotone decreasing function of $w \geq 0$. Thus, given ever-increasing values of $f \in [0, \alpha]$, $x_i(t)$ increases as well; the "shunt" $f$ has amplified the input intensity $C$. This multiplicative form of Now Print mechanism is not, for some purposes, as satisfactory as the additive mechanism that will be introduced in Section VIII.

Alternatively, the nonspecific shunting signal can act directly on the synaptic knobs that deliver the inputs $C_i$ to $v_i$ (see Fig. 15*b*). This would have the effect of directly amplifying the inputs, as in

$$\dot{x}_i = -\alpha x_i + \theta_i f(t) C$$

The same synaptic knob shunt can influence the memory traces by amplifying the presynaptic signals that perturb the knobs. For example, let

$$\dot{z}_{ji} = -\gamma_i z_{ji} + \delta_j f(t)[x_j(t - \tau_j) - \Gamma_j]^+ x_i$$

or let

$$\dot{z}_{ji} = -\gamma_i z_{ji} + \delta_j [x_j(t - \tau_j) - \Gamma_j(f(t))]^+ x_i$$

where $f(t)$ is a nonnegative, monotone increasing function of arousal level,

and $\Gamma(s)$ is a monotone decreasing function of $s = f(t)$. These laws mix exponential memory decay with a cross-correlator that can be shut on or off at will. Perfect memory until recall can also be modified in a similar fashion by letting

$$\dot{z}_{ji} = f(t)[x_j(t - \tau_j) - \Gamma_j]^+(-\gamma_j z_{ji} + \delta_j x_i)$$

or

$$\dot{z}_{ji} = [x_j(t - \tau_j) - \Gamma_j(f(t))]^+(-\gamma_j z_{ji} + \delta_j x_i)$$

Both sampling and Now Print must here be active as a precursor to learning or forgetting. Various other formal possibilities are special cases of our analysis; for example, shutting off the Now Print mechanism can prevent all memory change, whereas turning it on can permit exponential memory decay and/or new learning, as in the equation

$$\dot{z}_{ji} = f(t)\{-\gamma_j z_{ji} + \delta_j[x_j(t - \tau_j) - \Gamma_j]^+ x_i\}$$

2. *Local Flow*

The signal terms $\beta_{jk}[x_j(t - \tau_{jk}) - \Gamma_{jk}]^+$ and $\delta_{jk}[x_j(t - \tau_{jk}) - \Gamma_{jk}]^+$ in Eqs. (8) and (9), respectively, can be replaced, say, by

$$B_{jk}(t) = \beta_{jk}(t)[x_j(t - \tau_{jk}(t)) - \Gamma_{jk}(t)]^+$$

and

$$E_{jk}(t) = \delta_{jk}(t)[x_j(t - \sigma_{jk}(t)) - \Omega_{jk}(t)]^+$$

which permit different, and variable, time lags, thresholds, and path strengths in the two signal strength functionals. This includes the possibility of coupling a Now Print mechanism to these functionals, through either the variable path strengths or the thresholds. Functional $E_{jk}(t)$ describes the effect of the signal from $v_j$ on the cross-correlational process within $N_{jk}$ that determines $z_{jk}$. Functional $B_{jk}$ describes the net signal from $v_j$ that ultimately influences $v_k$ after being processed in $N_{jk}$. It is therefore natural to physically expect that $\Gamma_{jk} \geq \Omega_{jk}$. This *local flow condition* says little more than that the signal from $v_j$ passes through $N_{jk}$ on its way to $v_k$. Such a condition is, in fact, needed to guarantee that many cells can simultaneously sample a given pattern without creating asymptotic biases in their memory (Grossberg, 1971c, 1972b). This condition has an easily realized physical interpretation, given the assumption that the process $z_{jk}$ occurs in the synaptic knob or at postsynaptic membrane sites. Various other interpretations for $z_{jk}$ do not yield a physical basis for the local flow condition, and could not realize the possibility of simultaneous sampling by many input channels. The local flow condition provides examples of systems that can learn patterns without performing them until later, but cannot perform old patterns without also learning new patterns that are imposed during performance.



The functionals $B_{jk}$ and $E_{jk}$ permit more complicated possibilities as well. For example, *in vivo*, after a signal is generated in $e_{jk}$, it is impossible to generate another signal for a short time afterward (absolute refractory period) and harder to generate another signal for a short time after the absolute refractory period (relative refractory period). Also, some cells emit signals in complicated bursts. Intricate preprocessing of input signals can occur in the dendrites of cells before the transformed inputs influence the cell body. All such continuous variations are, in principle, covered by our theorems, which say that, whereas such variations can influence transient motions of the system, the classification of limits and oscillatory possibilities is unchanged by them. Given that weak constraints such as local flow hold, what is learned depends on which cells sample what patterns, and how intensely, no matter how complicated the rules are for determining when a cell will sample.

It is physically interesting that those terms, such as $B_{jk}$ and $E_{jk}$, which describe processes that act over a distance (such as signals flowing along $e_{jk}$) are the terms in Eqs. (22) and (23) that permit the most nonlinear distortion without destroying learning properties. The term $x_i$ in Eq. (23) is not of this type. This term is computed in $N_{ji}$ from the value $x_i$ in the contiguous vertex $v_i$.

## C. Local Symmetry Axes

In their final form, the theorems show that unbiased pattern learning can occur in systems with arbitrary positive path weights $\beta_{ji}$ from $j \in J$ to $i \in I$. This is achieved by first restricting attention to systems of the form

$$\dot{x}_i = Ax_i + \sum_{k \in J} B_k z_{ki} + C_i \tag{24}$$

and

$$\dot{z}_{ji} = D_j z_{ji} + E_j x_i \tag{25}$$

where $i \in I$ and $j \in J$. That is, all functionals $A_i$, $B_{ji}$, $D_{ji}$, and $E_{ji}$ are chosen independent of $i \in I$, and the anatomy is constrained to make this possible. These constraints mean that all cells $\mathcal{B} = \{v_i : i \in I\}$ are sampled by a given cell, $v_j$, in $\mathcal{A} = \{v_j : j \in J\}$ without biases due to system parameters ($B_{ji} = B_j$, $D_{ji} = D_j$, $E_{ji} = E_j$), and that the inputs to all cells $\mathcal{B}$ are averaged by their cell potentials without biases due to averaging rates ($A_i = A$) (see Fig. 16a). Systems (24) and (25) allow each cell to have a different time lag, threshold, and axon weight, as in

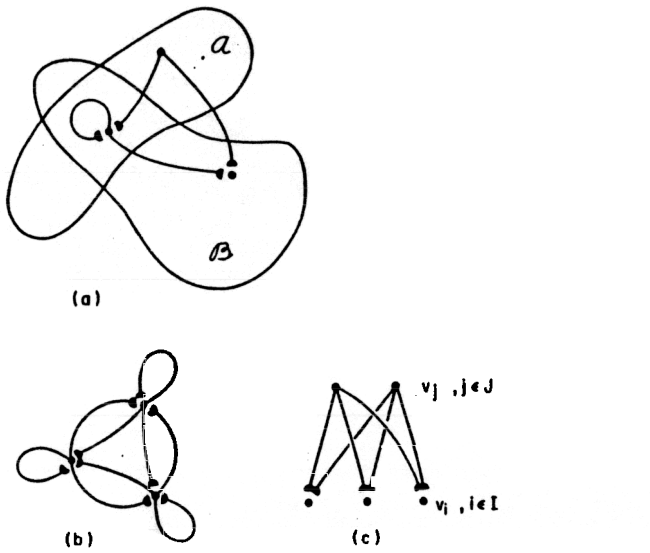


FIG. 16. Constraining an arbitrarily large set of sampling cells by imposing local symmetry axes.

$B_j(t) = \beta_j[x_j(t - \tau_j) - \Gamma_j]^+$. Even if all cells interact, as in Fig. 16*b*, no biases in asymptotic learning need occur due to these asymmetries in signal transfer among possibly billions of cells.

Figure 16, *b* and *c*, illustrates two extremal anatomies, the completely recurrent ($I = J$) and the completely nonrecurrent ($I \cap J = \phi$) cases. Generalizations of Fig. 16*a* are also possible. In these generalizations, $\mathcal{A}$ and $\mathcal{B}$ are replaced by sets $\{\mathcal{A}_k\}$ and $\{\mathcal{B}_k\}$ of subsets such that each cell in a given $\mathcal{B}_k$ is sampled by all cells in $\mathcal{A}_k$. One seeks the maximal subsets $\mathcal{B}_k$ for which this decomposition exists. For some purposes, a fixed set $\{\mathcal{B}_k\}$ is determined by structural considerations; for example, each $\mathcal{B}_k$ controls a different motor effector. It is then sometimes profitable to introduce fictitious cells into the sampling cells $\mathcal{A}$ if some cells in $\mathcal{A}$ sample two or more subsets $\mathcal{B}_k$. For example, if cell $v_i$ in $\mathcal{A}$ samples $\mathcal{B}_1$ and $\mathcal{B}_2$, replace $v_i$ by two cells, $v_{i1}$ and $v_{i2}$, such that $v_{ij}$ samples only $\mathcal{B}_j$, $j = 1, 2$, and each $v_{ij}$ receives the same inputs, and has the same parameters and initial data, as the original cell, $v_i$, had. Otherwise expressed, suppose that a given cell (population) can sample motor controllers of both hands, but that only the left hand is used to learn a given task. We then want to study the pattern variables associated with the left hand only, not both hands. The decomposition exhibits the system in a form suitable to this analysis. The mathematical analysis of systems (24) and (25) can be found in Grossberg (1969d, 1971c, 1972b).



### D. Unbiased Learning with Arbitrary Positive Axon Weights Using Chemical Transmission and Action Potentials

Let Eq. (24) be replaced by

$$\dot{x}_i = Ax_i + \sum_{k \in J} B_k \beta_{ki} z_{ki} + C_i \tag{26}$$

that is, let the path weights, $\beta_{ji}$, from $v_j$ to $v_i$ be arbitrary positive numbers. Can we transform Eq. (25) analogously so that learning and performance of spatial patterns is unimpaired? The answer is "Yes."

We want the pattern variables

$$Z_{ji}^{(\beta)} = \beta_{ji} z_{ji} \Big( \sum_{k \in I} \beta_{jk} z_{jk} \Big)^{-1}$$

to converge to $\theta_i$ after sufficient practice. This will happen if Eq. (25) is replaced by

$$\dot{z}_{ji} = D_j z_{ji} + E_j \beta_{ji}^{-1} x_i \tag{27}$$

since letting $w_{ji} = \beta_{ji} z_{ji}$, Eqs. (26) and (27) yield

$$\dot{x}_i = Ax_i + \sum_{k \in J} B_k w_{ki} + C_i$$

and

$$\dot{w}_{ji} = D_j w_{ji} + E_j x_i$$

which are again of the form of Eqs. (24) and (25). A mathematical analysis shows that our goal could *not* be achieved by replacing Eq. (25) with

$$\dot{z}_{ji} = D_j z_{ji} + E_j \beta_{ji} x_i$$

which would be the natural thing to do if we supposed that $E_j \beta_{ji}$ is determined wholly by spiking frequency (Grossberg, 1972b).

How can the $\beta_{ji}$'s in Eqs. (26) and (27) be interpreted? Suppose that $\beta_{ji} = \lambda_j R_{ji}$, where $\lambda_j > 0$ and $R_{ji}$ is the circumference of the cylindrical axon, $e_{ji}$. Let the signal in $e_{ji}$ [for example, the action potential (Ruch *et al.*, 1971)] propagate along the circumference of the axon to its synaptic knob. Let the signal disperse throughout the cross-sectional area of the knob [for example, as ionic fluxes (Katz, 1966)]. Let local chemical transmitter production in the knob be proportional to the local signal density. Finally, let the effect of the signal on the postsynaptic cell be proportional

to the product of local signal density and local transmitter density and the cross-sectional area of the knob.

These laws generate Eqs. (26) and (27) as follows. Signal strength is proportional to $R_{ji}$ or to $\beta_{ji}$. The cross-sectional area of the knob is proportional to $R_{ji}^2$. Hence signal density in the knob is proportional to $R_{ji}R_{ji}^{-2} = R_{ji}^{-1}$, or to $\beta_{ji}^{-1}$, as in Eq. (27). Thus (signal density) $\times$ (transmitter density) $\times$ (area of knob) $\cong R_{ji}^{-1}z_{ji}R_{ji}^2 = R_{ji}z_{ji} \cong \beta_{ji}z_{ji}$, as in Eq. (26).

By contrast, a mechanism whereby signals propagate throughout the cross-sectional area of the axon could not produce unbiased learning given arbitrary axon connection strengths, or at least such a mechanism is still elusive. The difficulty here is that signal strength is proportional to $R_{ji}^2$, signal density is proportional to one, and local transmitter production rate is then proportional to one. The postsynaptic signal is proportional to (signal density) $\times$ (transmitter density) $\times$ (area of knob) $\cong \beta_{ji}^2 z_{ji}$. Thus we are led to the system

$$\dot{x}_i = Ax_i + \sum_{k \in J} B_k \beta_{ki}^2 z_{ki} + C_i$$

and

$$\dot{z}_{ji} = D_j z_{ji} + E_j x_i$$

which can be written as

$$\dot{x}_i = Ax_i + \sum_{k \in J} B_k \beta_{ki} w_{ki} + C_i$$

and

$$\dot{w}_{ji} = D_j w_{ji} + E_j \beta_{ji} x_i$$

in terms of the variables $w_{ji} = \beta_{ji}z_{ji}$. This system has unpleasant mathematical properties (Grossberg, 1972b).

These observations suggest that the action potential not only guarantees faithful signal transmission over long cellular distances, as is well known, but also executes a subtle transformation of signal densities into transmitter production rates that compensates for differences in axon diameter. Note also that this transformation seems to require the chemical transmitter step. Purely electrical synapses presumably could not execute it. Thus our laws for transmitter production (and/or related processes) not only guarantee that learning occurs, but also that unbiased learning occurs, under very weak anatomical constraints. Section IX suggests another way in which the action potential contributes to unbiased learning on the level of individual cells.

The next two sections illustrate some phenomena that occur in networks with specific anatomies.



## E. Threshold-Dependent Phase Transitions in Recurrent Networks

Consider Figs. 16 and 17. Figure 16*b* is a *recurrent* network: the cells send signals to each other. Figure 16*c* is a *nonrecurrent* network: the cells send signals only to different cells. Not surprisingly, under certain circumstances, the memory of recurrent and nonrecurrent networks can differ dramatically. Less intuitively, a recurrent network can sometimes behave like a nonrecurrent network. Moreover, an anatomist could not tell the difference between a recurrent network which behaves recurrently from one which behaves nonrecurrently

Figure 17 illustrates what is involved in making this distinction. Figure 17 depicts a recurrent network whose recurrent signals are carried by interneurons between the signal generating cells. Let the threshold for signals to leave the cells be $\Gamma_1$, and let the threshold of the interneurons be $\Gamma_2$. Suppose that $\Gamma_1 = \Gamma_2 = 0$. Then any input to a cell $v_j$ will create outputs and signals to other cells $v_i$. These signals will, in turn, create outputs from $v_j$ and feedback signals to $v_i$, and so on. As a consequence, recall trials can destroy the memory of this system. Suppose, however, that $\Gamma_2 \gg 0$. Then an output from a cell can again create signals to other cells. These signals can in turn, create outputs from these cells *without* causing feedback signals. Such a network has a nonrecurrent kind of memory: Recall need not destroy the memory of the system. During recall, each cell and its interneurons behaves like an outstar embedded in a larger, but functionally passive, anatomy in this case. The thresholds thus serve to localize the memory trace, and to provide a kind of localized "context" which a given input can activate. Whereas this argument holds during recall of a spatial pattern or during slow recall of a space–time pattern, Section VII shows that it need not hold during rapid recall of a space–time pattern.

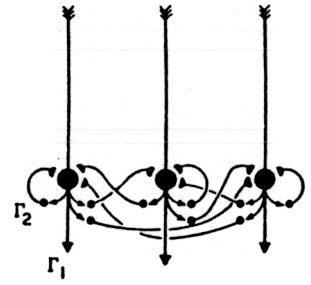


Fig. 17. Influence of interneuronal thresholds on whether a recurrent anatomy behaves recurrently or nonrecurrently.

Small inputs to the network of Fig. 17 can make it behave like a nonrecurrent network; even slightly larger inputs can make it behave recurrently, by creating signals that are sufficiently large to exceed the feedback thresholds. For example, varying the overall arousal level of the system can change its behavior in response to fixed externally controlled inputs from nonrecurrent to recurrent or conversely. The asymptotic behavior of the system is a discontinuous function of input and threshold perturbations: There is a "phase transition" at critical values of these parameters. Given this possibility, one can argue in the reverse direction. Suppose that a nonrecurrent type of memory is desired at all times. How can the total input to the cells be "normalized" so that the feedback thresholds are never exceeded? Various arrangements of nonspecific inhibitory interneurons can accomplish this task (Grossberg, 1970a, 1972d, 1973).

## F. Pattern Completion and Mass Action

In Fig. 16*c*, suppose that any fraction of sampling cells is excised away. The remaining sampling cells can reproduce an entire learned pattern on the sampled cells if some of the remaining sampling cells were active when that pattern was being learned ("pattern completion"). In Fig. 16*b*, *each* vertex, $v_i$, can encode and perform a *different* spatial pattern at all the vertices, if the dynamics of the network are nonrecurrent in the sense of the previous section. By contrast, suppose that sampling cells can sample only a fixed fraction of sampled cells, and that the sampled cells are chosen randomly. Then, on the average, excising ever greater numbers of sampling cells will create a proportional deficit in the ability of the remaining sampling cells to reproduce a previously learned pattern spread across all sampled cells ("mass action").

## G. Imprinting and Irreversibility

Mathematical analysis of systems (24) and (25) shows that, once these systems are factored into pattern variables and total energy variables, different choices of functionals influence transient motions of pattern variables, but not the possible oscillations of these variables. In particular, different functionals, or different values of fixed functionals due to particular choices of inputs, can determine different numerical limits of the pattern variables as $t \to \infty$. This section summarizes some results concerning these limits which have been proved for a particular choice of functionals, but which should hold for many other functionals chosen in the same anatomies (Grossberg, 1968b, 1969c).



This choice of functionals determines an interference theory of forgetting in the nonrecurrent outstar anatomy; for example, let $(B_j z_{ji})(t) = \beta_j[x_j(t - \tau_j) - \Gamma_j]^+ Z_{ji}(t)$ in Eq. (24). In various recurrent anatomies, however, these functionals do not determine an interference theory. Instead, there exists a phase transition in memory, such that one type of memory prevails if the network's numerical parameters have certain values, whereas a distinct type of memory prevails if the parameters take on the remaining values. Consider Fig. 18. Given the anatomy of Fig. 18*a*, there exists an example of the following type. The numerical values of the network parameters—such as $\alpha$, $\beta$, $\gamma$, $\tau$, $\Gamma$ in Eqs. (8) and (9)—form two exhaustive and nonoverlapping sets, A and B. If the parameter values fall in A, then the network can remember everything; if the parameter values fall in B, then the network cannot remember anything. Thus, spontaneous forgetting occurs if parameter values fall in B, even though, speaking locally, the interaction terms describe an interference theory of forgetting. The global anatomy determines this forgetting effect. In Fig. 18*b*, if the parameter values fall in A, then the network can remember everything; if the parameter values fall in B, then the network can remember spatial patterns. For example, given A, the network can remember lists, or space–time patterns. Given B, the network forgets temporal discriminations, and its memory seeks the spatial pattern closest to what it has learned. Thus the global recurrent anatomy not only determines that two phases exist, but also what the memory characteristics of each phase will be.

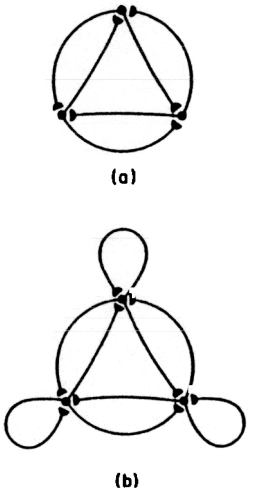


Fig. 18. Imprinting due to a phase transition in memory.

By varying network parameters, network dynamics can be transformed from phase B to phase A. Any mechanism that does this will "imprint" the memory of the input pattern that perturbs the network at the time this transition takes place. The transition from B to A can be effected, for example, by increasing the velocity of signals in the network axons. Given this formal observation, we now note various possible analogs of this phenomenon *in vivo*.

Signal velocity can be increased *in vivo* by laying down an axonal sheath around unmyelinated axons. Such a sheath can cause signals to jump along the axon in a saltatory fashion. Various strategies for imprinting a pattern of axonal connections in a particular subnetwork of a total network hereby suggest themselves. A nonspecific command signal (for example, a hormone) to this subnetwork to lay down sheaths on all subnetwork axons would suffice. Alternatively, one could imprint a pattern in the axons of particular nerves as they became active by coupling the activity of the sheath-producing cells to that of the nerves (cf. Orkand *et al.*, 1966). The *order* in which various cells imprinted patterns could be determined by such a mechanism. The interaction between external inputs and the total network anatomy could establish this order by determining which cells would reach the critical activity levels for sheath production first; cf. Grossberg (1969f, Section 19) in the light of Section IX below. Although the order in which particular nerves or subnetworks are imprinted can be developmentally predetermined by such a mechanism, the actual patterns that are imprinted depends on the choice of external inputs. If given cells do not pass from phase B to phase A, then they retain a plastic memory which can continue to spontaneously forget old patterns.

Grossberg (1969c) shows that these systems also have various properties that are of interest from the statistical mechanical point of view. For example, before such a network is probed by experimental inputs, its output might be linear, locally reversible ($z_{jk} = z_{kj}$), and globally reversible ($Z_{jk} = Z_{kj}$). An experimental input can make the output nonlinear, globally irreversible ($Z_{jk} \neq Z_{kj}$), but still locally reversible. After the effect of the input wears off, the output can become linear again. Whether the output again becomes globally reversible or not, however, depends on the sign of a function of network parameters that cannot be easily measured by an input–output analysis. Thus the (non)linearity of the system can be decoupled from its global (ir)reversibility. The decision whether the system will be become globally reversible or will remain globally irreversible after inputs cease depends on whether the network parameters fall into B or A. In all cases where this system is eventually free from inputs, its asymptotic behavior approaches that of a stationary Markov chain. Network dynamics provide a real-time description of the transient nonstationary behavior of the system as it approaches its stationary asymptote.



## VII. Serial Learning

### A. Qualitative Data

This section discusses the response of a recurrent network to a particular type of space–time pattern–namely, a list, or sequence of spatial patterns, in which only one component of each spatial pattern is positive. Section VI pointed out that a recurrent network can behave nonrecurrently in response to a spatial pattern if signals from a given vertex do not create feedback signals to that vertex. Even if parameters are chosen to guarantee this, the response of the network to a space–time pattern, in particular to a list of length $n$, can differ significantly from that of $n$ independent outstars to $n$ spatial patterns.

There exists a large body of data on list learning. Some of the themes in these data are sketched below. Our analysis of these data will be heuristic and will focus only on the effects that arise in the minimal anatomies that are capable of learning a list. Proofs and extensions of these assertions are found in Grossberg (1969e) and Grossberg and Pepe (1971). A more complete phenomenological analysis of the data on a neural level would study how list items, and sequences of items, are coded by hierarchically organized fields of cells with selective response profiles, and in particular of how the field activity is sustained by short-term memory mechanisms while it is transformed and transferred to long-term memory (cf. Atkinson and Shiffrin, 1968; Grossberg, 1973). This section studies one level of recurrent interactions in such a hierarchy. The goal is to better understand the hierarchical case by first gaining insight into various one-level cases. Once this is accomplished, hierarchical anatomies can be more readily synthesized.

#### 1. *Backward Learning*

Suppose that the list AB is sequentially presented several times to a learning subject $\mathcal{O}$. Let B alone be presented to the subject on recall trials. Other things being equal, prior practice of AB increases the probability of guessing A given B. That is, practicing AB yields at least partial learning of BA. Relative to the time scale of external events, which flows forward from A to B, learning both AB and BA, given practice of AB alone, means

that the internal dynamics of $\mathcal{O}$ flow both forward (AB) and backward (BA) in time.

2. *Global Arrow in Time*

Now suppose that the list ABC is practiced with a time lag of $w$ time units between successive presentations of each letter. After B has been presented to $\mathcal{O}$, and before C is presented, $\mathcal{O}$ has received only the list AB, and thus the association from B to A begins to form. We know, however, that ultimately ABC can be learned. Thus the forward association BC is stronger than the backward association BA, and can therefore inhibit it to yield a global arrow in time from A to B to C. In this sense, "time" is flowing both forward and backward within $\mathcal{O}$, but the forward flow is stronger and ultimately enables $\mathcal{O}$ to imitate the direction in time of external events.

3. *Bowing*

The same theme is illustrated by the phenomenon of bowing, which means that the middle of a serially learned list is harder to learn than either end, or, more familiarly, that we can often remember how a sequence of events began and ended but forget many intermediate details. If internal events in $\mathcal{O}$ flowed only forward in time, we might expect the plot of mean number of recall errors as a function of list position to be monotone nondecreasing, since at list positions ever deeper within the list, more response interference can accumulate from previously presented list items. In actuality, however, list positions near the list's middle are hardest to learn, which illustrates that the nonoccurrence of items after the last list item has somehow made items near the end of the list, which were presented earlier in time, easier to learn.

4. *Skewing*

A closely related phenomenon is skewing, which means that the list position that is hardest to learn often occurs nearer to the end than to the beginning of the list. This fact recalls the fact that learning in the forward direction (AB) is stronger than learning in the backward direction.

5. *Intratrial versus Intertrial Interval*

Many parametric studies of learning difficulty at various list positions have been reported. The intratrial interval (denoted by $w$) is the time between presentation of successive list items. The intertrial interval (denoted by $W$) is the time between two successive presentations of the



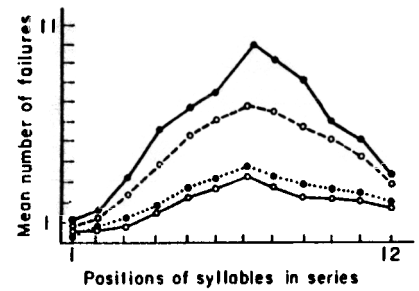


FIG. 19. Influence of intratrial interval and intertrial interval on degree of bowing. From Hovland (Osgood, 1953, p. 506). ●——● $W$ = 6 seconds, $w$ = 2 seconds; O- - -O $W$ = 2 minutes 6 seconds, $w$ = 2 seconds; ●· · ·● $W$ = 6 seconds, $w$ = 4 seconds; O——O $W$ = 2 minutes 6 seconds, $w$ = 4 seconds.

list—that is, the time between successive learning trials. Figure 19 illustrates the influence on bowing of varying $w$ and $W$. Note that increasing $w$ from 2 seconds to 4 seconds can substantially flatten the bowed curve, and that, once the curve is flattened in this fashion, increasing $W$ has little influence on the rate of learning. Slowing the presentation rate is an example of "distributing practice." Figure 19 shows that distributing practice reduces the number of learning errors.

When the list is presented rapidly (for example, $w$ = 2 seconds), increasing $W$ substantially reduces the number of errors in the middle of the list. In short, increasing the rest interval after the practice trial has simplified learning of the entire list, especially at its middle. This effect also illustrates the existence of backward learning effects. Increasing $W$ much beyond the 2-minute 6-second value does not reduce the number of errors substantially in these data.

Note that the dictum "Distributing practice improves learning" must be interpreted with caution. Letting $w$ approach 24 hours certainly distributes practice, but makes learning of the list quite unlikely. Thus we shall seek a list presentation speed, much less than $w$ = 24 hours but greater than $w$ = 0, that optimizes the benefits of distributing practice.

6. *Response Oscillation and Generalization*

This phenomenon is closely related to bowing (see Fig. 20). It says that the gap between the first correct guess and the last error is largest near the middle of the list. More list intrusions interfere with the correct association near the middle of the list than at its ends. In fact, a generalization gradient exists at each list position such that the probability of guessing an item, given presentation of a fixed item, decreases as a function of the number of intervening items presented on a single trial. The shape of this generaliza-

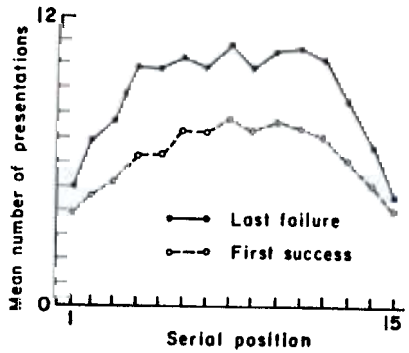


FIG. 20. Response oscillation. From Hull *et al.* (Osgood, 1953, p. 503).

tion gradient depends on list position. Given a sufficiently large intertrial interval, the gradient is skewed forward near the beginning of the list, backward near the end of the list, and in both directions near the list's middle, often with a broader span near the middle, and an advantage given to anticipactory rather than perseverative errors (Osgood, 1953), presumably as a manifestation of stronger forward than backward associations.

### 7. *Anchoring*

This phenomenon describes the order in which list items are learned (Atkinson and Shiffrin, 1968). Items are often learned both in the forward direction and in the backward direction around the "anchor" stimulus, A. For example, AB, then YZ, then BC and CD, then XY might be the first associations to be learned, and in the given order.

### 8. *Chaining*

By putting the learned fragments around the anchor together, we see that list items are often learned in growing chains around the anchor stimulus. These chains propagate from the anchor in both forward and backward directions, toward the middle of the list, and can gradually reduce the number of competing items that contribute to response oscillation at the list's middle.

### 9. *Chunking*

Suppose that a chain has formed. The chain can be performed—*as a unit*—given presentation of the anchor stimulus and persistent arousal, if it has an avalanche structure, in which each unit of the chain excites its motor representation as well as the next unit of the chain (Fig. 14). This



yields performance of each item, in its proper order, via successive excitation of the entire chain, unless arousal is withdrawn at an intermediate point. In this sense, starting with independent list items A, B, C, . . . , Z, practicing the alphabet (ABC . . . XYZ) can create new items, such as subsequences (AB), (ABC), (BCDE), . . . , etc., of the list. These new items can eventually be performed as effortlessly as the original items were. Composite list units, or "chunks" (Miller, 1956), are presumably being continually formed and reaggregated as practice continues until perfect learning is achieved. Here an analysis of hierarchical coding is appropriate, and in particular of whether there are sampling cells that are excited only by particular subsequences of the list.

### 10. *Primacy versus Recency*

Typically, the beginning of a serially learned list is easier to learn than the end, as in Fig. 19; that is, the *primacy* effect is stronger than the *recency* effect, or "primary dominates recency." In the minimal network, increasing the arousal level to high values can reverse this effect. Is there a corresponding phenomenon *in vivo*?

### 11. *Inverted* U *in Learning*

Either too little motivation (or arousal), or too much, can hamper performance. Figure 21 illustrates this typical result in general terms. It is well described in Hebb (1955). Analogous difficulties occur in the network below. Given underarousal, there is too little energy to drive the learning process. Given overarousal, there is ample energy to drive learning, but a high level of response interference is produced by incorrect associations that are similar either in time of presentation or in meaning to the correct associations. In other words, overarousal produces "fuzzy response sets," and by impairing the network's ability to focus on the correct association interferes with "paying attention."

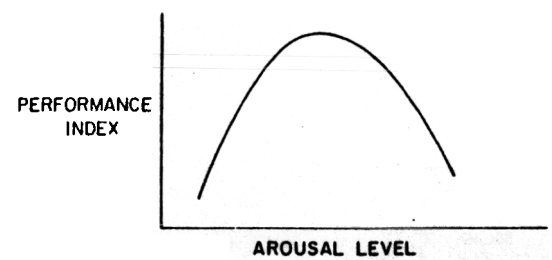


FIG. 21. Inverted U in learning.

Recent experiments (Kornetaky and Eliasson, 1969; Phillips and Bradley, 1970) have investigated the hypothesis that paying attention can be impaired by overarousal. Kornetsky and Eliasson varied the excitation level of white rats during a sustained attention task. The task chosen was for the rats to press the lever on presentation of a specific auditory stimulus. The experimenters noted any impairment in performance as a function of excitation level. High excitation was produced by electrically stimulating the rats' reticular formation. Low excitation was produced by administering a dose of chlorpromazine. Both electrical stimulation and chlorpromazine interfered with attention. The two treatments together resulted in performance indistinguishable from that seen after injections of saline alone. Presumably electrical stimulation and chlorpromazine antagonize each other and return the rat to a normal arousal level.

### 12. *Overarousal and Punning*

There exist networks in which overarousal weakens the strength of correct associations at the list's beginning by forcing them to compete with incorrect associations formed with later list items. Suppose that the list is a sentence. By the time the entire sentence has been presented to such an overaroused network, the earlier portions of the sentence have been washed away by a flood of competing associations. The meaning of the sentence is similarly lost. Only the last few list items survive the flood, and only these can therefore influence responses to the sentence. Structurally similar words, such as rhymes or puns, can be expected, rather than meaningful replies. Maher (1968) has discussed a phenomenon of punning in certain schizophrenics who are presumed to be in a continual state of overarousal. Various manics also pun. Chlorpromazine can improve the performance of schizophrenics at tasks that require sustained attention, presumably by lowering their arousal level. Lithium presumably has a similar effect in manics (Dally, 1967). Inspection of the networks shows that different mechanisms can produce similar symptoms of overarousal. For example, unduly large inputs from a nonspecific arousal source, such as reticular formation, can cause overarousal. Alternatively, pathological changes in the binding of ions (for example, $Ca^{++}$) at network cells, none of which is necessarily a nonspecific arousal source, can cause overarousal by amplifying all signals in the network. Presumably cures for similar difficulties in paying attention needed by different patients can be quite dissimilar.

Different network anatomies respond to fluctuations in arousal level in different ways. For example, in networks that describe the interaction of rewarding events with internal drives, either under- or overarousal can produce "emotional depression" by reducing the incentive motivational



response of the network to emotionally charged cues (Grossberg, 1972c). The underaroused network responds "irritably" to sufficiently large increases in such cues, whereas the overaroused network is "indifferent" to these cues. In networks describing recurrent on-center off-surround interactions of shunting type, low arousal can help the network "choose" among many response alternatives, whereas high arousal tends to store many cues in short-term memory (Grossberg, 1972b). Grossberg (1975) combines such mechanisms to analyze various attentional and discrimination learning data.

The remainder of this section qualitatively describes some formal network mechanisms that behave analogously to psychological data such as that above.

### B. Backward Learning

Consider the minimal anatomy that can learn AB or BA (that is, $\beta_{AB} > 0$ and $\beta_{BA} > 0$), as well as related response alternatives such as AA, AC, BB, or BC. Suppose that the network parameters are unbiased and that no association is preferred initially. The very possibility of learning BA in this context will imply that BA will be at least partially learned when AB is practiced. Thus backward learning effects can arise simply because choices exist. The greater learning of AB than of BA will be due to the existence of better cross-correlations between signals and potentials in the forward direction than in the backward direction.

Let the network be represented by the following equations for definiteness. More general functionals can also be used.

$$\dot{x}_A = -\alpha x_A + \beta[x_A(t-\tau) - \Gamma]^+ z_{AA} + \beta[x_B(t-\tau) - \Gamma]^+ z_{BA} + C_A$$

$$\dot{x}_B = -\alpha x_B + \beta[x_B(t-\tau) - \Gamma]^+ z_{BB} + \beta[x_A(t-\tau) - \Gamma]^+ z_{AB} + C_B$$

$$\dot{x}_C = -\alpha x_C + \beta[x_A(t-\tau) - \Gamma]^+ z_{AC} + \beta[x_B(t-\tau) - \Gamma]^+ z_{BC} + C_C$$

and

$$\dot{z}_{ij} = -\gamma z_{ij} + \delta[x_i(t-\tau) - \Gamma]^+ x_j$$

where $(i, j) = (A, A), (B, B), (A, B), (B, A), (A, C)$, or $(B, C)$. Present the serial list once with an intratrial interval of $w$. Then $C_A(t) = C_B(t + w)$, and $C_C(t) \equiv 0$.

A particular, but noncrucial, choice of $w$ will be made to emphasize the main effects. To maximize the possibility of learning AB, let the signal from $v_A$ to $N_{AB}$ arrive at $N_{AB}$ as the input $C_B(t)$ to $v_B$ arrives; that is, let the sampling delay from the onset time of the input $C_A$, namely,

$$D(\tau, \Gamma) = \tau + \min\{t: x_A(t) = \Gamma, \dot{x}_A(t) > 0\}$$

satisfy the identity

$$D(\tau, \Gamma) = w \tag{28}$$

This yields maximal overlap of the signal $\beta[x_A(t-\tau) - \Gamma]^+$ and the potential $x_B(t)$ for purposes of cross-correlation by $z_{AB}(t)$. All knobs $N_{AA}$, $N_{AB}$, and $N_{AC}$ receive equal signals from $v_A$. The signal from $N_{AC}$ to $v_C$ is dominated at $v_B$ by the signal from $N_{AB}$ and the input $C_B(t)$. Thus, after learning begins, $Z_{AB} > Z_{AC}$, where $Z_{ij} = z_{ij}(\sum z_{ik})^{-1}$. The vertex $v_A$ also receives two inputs—namely, the signal from $N_{AA}$ and $C_A(t)$. Nonetheless, the correlation between the $N_{AA}$ signal and $C_A(t)$ is not as good as the correlation between the $N_{AB}$ signal and $C_B(t)$. Thus, $Z_{AB} > Z_{AA} > Z_{AC}$. A similar argument shows that $Z_{BA} > Z_{BC}$ after sampling begins at the knobs $N_{BA}$ and $N_{BC}$. The correlation between the $N_{BA}$ signal and $C_A(t)$ is not as good as that between the $N_{AB}$ signal and $C_B(t)$. Choosing between the inequalities $Z_{BA} > Z_{BB}$ and $Z_{BA} \leq Z_{BB}$ requires a study of network parameters. This is because $N_{BA}$ samples the decaying input $C_A(t)$ boosted by self-excitation via $N_{AA}$, whereas $N_{BB}$ samples the decaying input $C_B(t)$ boosted by its own self-excitation.

### C. Optimal Learning Speeds

Consider the following network anatomies for definiteness.

1. *Complete* n-*Graph without Loops.* This is the minimal anatomy that can learn any list, $r_i r_j$, of length 2 with distinct entries (see Fig. 22*a*).

2. *Complete* n-*Graph with Loops.* This is the minimal anatomy that can learn any list of length 2 (see Fig. 22*b*).

3. *Two-Layer Graph with Completely Nonrecurrent Sampling.* Each input $C_i(t)$ is delivered to two vertices, $v_{1i}$ and $v_{2i}$. Each vertex $v_{1i}$ can sample all the vertices $v_{2k}$ (see Fig. 22*c*); e.g., each $v_{1i}$ is a command population excited by a subsequence of the list at a uniform rate.

We shall denote a particular network corresponding to a given alphabet $\mathfrak{A} = \{r_1, r_2, \ldots, r_n\}$ of behavioral units by $\mathfrak{N}(\mathfrak{A})$. The graphs in Fig. 22 will be assumed to be unbiased for definiteness; that is, all vertices or edges of a given type possess the same parameters. For an example of an unbiased complete $n$-graph without loops consider

$$\dot{x}_i = -\alpha x_i + \beta \sum_{k \neq i} [x_k(t-\tau) - \Gamma]^+ z_{ki} + C_i \tag{29}$$

$$\dot{z}_{jk} = -\gamma z_{jk} + \delta[x_j(t-\tau) - \Gamma]^+ x_k, \qquad j \neq k \tag{30}$$

and

$$z_{jj} = 0 \tag{31}$$

where $i, j, k = 1, 2, \ldots, n$.



Let a long list $r_1 r_2 \ldots r_L$ be serially presented to an unbiased complete $n$-graph without loops, for definiteness. Thus $C_1(t) = C_2(t+w) = \cdots = C_L[t + (L-1)w]$. The stimulus sampling probabilities of such a network are defined by $Z_{jk} = z_{jk}(\sum_{m \neq j} z_{jm})^{-1}$. Suppose initially that the network is at rest and that all associations are equally strong; that is, $x_i(t) = 0$ and $Z_{jk}(0) = 1/(n-1)$, for $i = 1, 2, \ldots, n$, $j \neq k$, and $t \leq 0$.

Even if the inputs $C_i(t)$ arrive through independent input channels, no learning occurs if $w = 0$, since then all inputs are equal and, by symmetry, the memory traces remain uniformly distributed.

Suppose by contrast that $w \gg D(\tau, \Gamma)$. Then $v_i$ begins to sample $D(\tau, \Gamma)$ time units after it is perturbed by $C_i(t)$. After $C_i(t)$ becomes zero again, these sampling signals gradually decay to zero. Only after sampling ceases does $C_{i+1}(t)$ become positive. Hence $[x_i(t-\tau) - \Gamma]^+ x_j(t) \cong 0$ for all $i, j$, and no learning occurs.

No learning occurs if $w = 0$ because the potentials are uniformly distributed, and therefore indistinguishable from each other. No learning

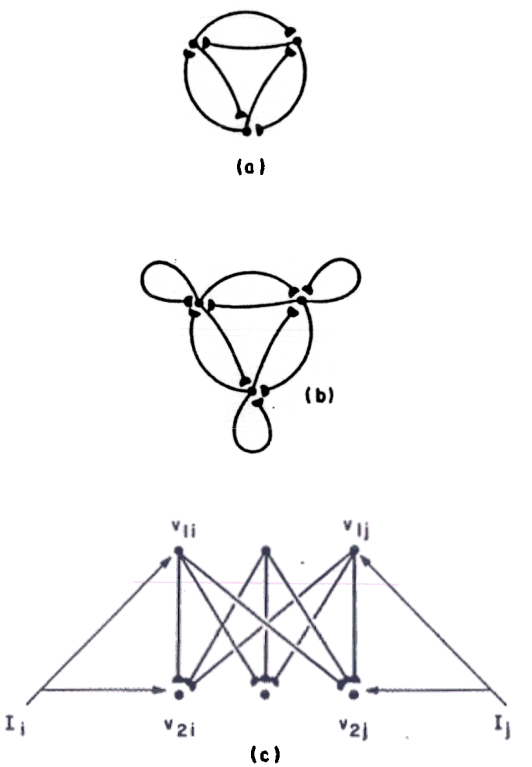


Fig. 22. Some networks in which bowing can occur.

occurs if $w \gg D(\tau, \Gamma)$ because the cross-correlations are poor. Learning is best given intratrial intervals $w$ such that $w \simeq D(\tau, \Gamma)$, at which good distinguishability *and* good cross-correlations prevail.

### D. Bare Field

The description of bowing can be approached in several stages. First, suitable anatomies must be chosen. The networks of Fig. 22, given unbiased parameters, are suitable examples. When a long serial list is presented to these graphs, bowing occurs. Thus, the analysis in Section II of the "two-body problem" of learning AB implies the existence of phenomena, such as bowing, which occur when "$n$-bodies" such as the alphabet ABC . . . XYZ, interact. For definiteness, we shall restrict attention to Fig. 22*a* using the simplest possible functions as in Eqs. (29)–(31). Let the inputs $C_i(t)$ be presented with intratrial interval $w$ and intertrial interval $W$ on $N$ trials $\varepsilon_1(w, W; L), \ldots, \varepsilon_N(w, W; L)$ of the list $r_1 r_2 \ldots r_L$. Thus

$$C_i(t) = \sum_{m=0}^{N-1} J[t - (i-1)w - (L-1)mw - mW], \qquad i = 1, 2, \ldots, L \tag{32}$$

and

$$C_i(t) \equiv 0, \qquad i = L+1, L+2, \ldots, n \tag{33}$$

where $J(t)$ is an input pulse that is positive in the interval $(0, \lambda)$.

We seek a closed formula for $Z_{jk} = z_{jk}(\sum_{m=1}^{n} z_{jm})^{-1}$ as a functional of the serial inputs $C_i$. Such a formula is not available for the system (29)–(31), but one can be derived for a closely related system that embodies the main effect of the serial inputs on the sampling probabilities $Z_{jk}$. This system, called the *bare field* of (29)–(31), ignores the influence of the nonlinear interaction term $\beta \sum_{k \neq i} [x_k(t-\tau) - \Gamma]^+ z_{ki}$ in (29), which tends to preserve learned pattern weights except for a certain amount of smoothing when several vertices are simultaneously active, and the decay term, $-\gamma z_{jk}$, in (30), which does not change the equations for pattern variables. The bare field of a complete $n$-graph without loops is therefore defined by

$$\dot{x}_i = -\alpha x_i + C_i \tag{34}$$

$$\dot{z}_{jk} = \delta[x_j(t-\tau) - \Gamma]^+ x_k, \qquad j \neq k \tag{35}$$

and (31), subjected to the inputs (32) and (33). Thus, bowing can be derived from three properties taken together: (1) exponentially averaged serial inputs, from (34), (2) delayed cross-correlations of the averaged inputs, from (35), and (3) the influence of competing associations $r_j \to r_m$, $m \neq k$, on $r_j \to r_k$, from the definition of $Z_{jk} = z_{jk}(z_{jk} + \sum_{m \neq k} z_{jm})^{-1}$.



Rather than state theorems about the bare field, we first present an intuitive argument that clarifies the main effects.

### E. Accumulation Sets

At what times does learning occur from $r_j$ to $r_k$? That is, at what times does $Z_{jk}(t)$ grow? $Z_{jk}(t)$ grows if $z_{jk}(t)$ grows and the competing terms $z_{jm}(t)$, $m \neq k$, do not grow commensurately. By (35) this means that $Z_{jk}(t)$ grows if

$$[x_j(t-\tau) - \Gamma]^+ x_k(t) \gg [x_j(t-\tau) - \Gamma]^+ x_m(t), \qquad m \neq k \tag{36}$$

Equation (36), in turn, can be achieved if

$$[x_j(t-\tau) - \Gamma]^+ \gg 0, \qquad x_k(t) \gg 0, \quad \text{and} \quad x_m(t) \cong 0, \qquad m \neq k \tag{37}$$

Equation (37) shows that the growth of $Z_{jk}(t)$ will be influenced by the number of STM traces, $x_m(t)$, that are large at any given time—that is, by the distinguishability of the correct association. We therefore seek a simple way to count how many, and which, STM traces are large at any time. For simplicity, we shall constrain the input pulse, $J(t)$, from which the inputs $C_i(t)$ are constructed by the following conditions:

(1) $J(t)$ is positive only in $(0, \lambda)$, where $\lambda < D(\tau, \Gamma)$. That is, the duration of each input is less than the time needed for sampling at $v_j$ to be induced by an input to $v_i$.

(2) $J(t)$ increases monotonically to a finite maximum and then decreases monotonically to zero.

(3) $w \leq W$.

Given these conditions, the following proposition holds.

*Proposition* 1. Suppose that $x_i(0) = 0$, $i = 1, 2, \ldots, n$. Then

(*a*)
$$x_i(t) = x_1[t - (i-1)w], \qquad i = 1, 2, \ldots, L \tag{38}$$

where

(*b*)
$$x_1(t) = \begin{cases} 0 & t < 0 \\ \int_0^t e^{-\alpha(t-v)} J(v)\, dv, & 0 \leq t \leq \lambda \\ Ke^{-\alpha t} & \lambda < \ \leq (L-1)w + W \end{cases} \tag{39}$$

with $K = \int_0^\lambda e^{\alpha v} J(v)\, dv$. Moreover

(c) $x_1$ rises monotonically to its maximum in $(0, \lambda)$ and thereafter decays to zero, at the exponential rate $\alpha$ for $t > \lambda$.

To count how many potentials are large at time $t$, introduce a criterion $\epsilon > 0$ of largeness. For fixed $\epsilon > 0$, let $A_\epsilon(w, W, L; t)$ denote the collection of indices $i$ such that $x_i(t) \geq \epsilon$. That is $A_\epsilon(w, W, L; t)$ tells us which $x_i$ are at least as large as $\epsilon$ at time $t$. For simplicity, we write $A_\epsilon(w, W, L; t)$ as $A_\epsilon(t)$, and also let $| A_\epsilon(t) |$ be the number of indices in $A_\epsilon(t)$. The set $A_\epsilon(t)$ is called the *$\epsilon$-accumulation set at time $t$*, since it contains the indices of all vertices, $v_i$, which have accumulated at least an amount $\epsilon$ of potential at time $t$. We always suppose in the following that $\epsilon$ is fixed in such a way that $0 < \epsilon < \max\{x_1(t) : t \geq 0\}$ to avoid trivialities. The following basic facts concerning $A_\epsilon(t)$ on the first trial $\mathcal{E}_1(w, W; L)$ are easy consequences of proposition 1.

The function $| A_\epsilon(t) |$ remains zero until the first time $t = t_\epsilon$ at which $x_1(t) = \epsilon$. Then $| A_\epsilon(t) | = 1$. The index 1 remains in $A_\epsilon(t)$ until the time $t = T_\epsilon$ at which $x_1(t) = \epsilon$ for the last time, since by proposition 1 we can also assert that $x_1(t) \geq \epsilon$ for all $t$ in $[t_\epsilon, T_\epsilon]$. Since $x_i(t) = x_1[t - (i-1)w]$ for all $i = 1, 2, \ldots, L$, the index 2 enters $A_\epsilon(t)$ at time $t = t_\epsilon + w$, the index 3 enters $A_\epsilon(t)$ at time $t = t_\epsilon + 2w$, and in general the index $i$ enters $A_\epsilon(t)$ at time $t = t_\epsilon + (i-1)w$, $i = 1, 2, \ldots, L$. Each of these indices remains in $A_\epsilon(t)$ for $T_\epsilon - t_\epsilon$ time units, and none of the indices $i = L + 1, \ldots, n$ ever enters $A_\epsilon(t)$.

The overall behavior of $A_\epsilon(t)$ as $t$ varies within $[0, Lw]$ depends on two factors, for fixed $w$, $W$, and $L$. These are the amount of time $S_\epsilon = T_\epsilon - t_\epsilon$ that a single index remains in $A_\epsilon(t)$, and the number of new indices that are added to $A_\epsilon(t)$ during this time. To describe the interplay of these quantities in a precise way, we introduce the following notation.

For any $u \geq 0$, let $[u]$ be the greatest integer that does not exceed $u$. Now let $G_\epsilon(w) = [S_\epsilon/w]$. The term $G_\epsilon(w)$ measures the number of new indices that can be added to $A_\epsilon(t)$ before an old index drops out. Since $S_\epsilon$ is independent of $w$, $G_\epsilon(w)$ is a *monotone decreasing* function of $w$. We shall find that the existence or nonexistence of a bowing effect in $\mathfrak{M}(\mathfrak{U})$ during the learning of a given list $\mathfrak{L} = r_1 r_2 \ldots r_L$ can be qualitatively decided by examining the absolute and relative sizes of $G_\epsilon(w)$, $L$, and $W$. To do this, we must distinguish two cases.

*Case* 1 $[G_\epsilon(w) < L - 1]$: In this case, $A_\epsilon(t)$ accumulates the indices $1, 2, \ldots, G_\epsilon(w) + 1$ at a linear rate at the times $t = t_\epsilon, t_\epsilon + w, \ldots, t_\epsilon + wG_\epsilon(w)$. In particular, $| A_\epsilon(t) |$ jumps by 1 every $w$ time units from its initial value 0 until it reaches $G_\epsilon(w) + 1$. After time $t = T_\epsilon$, the "old" index 1 drops out of $A_\epsilon(t)$, but at time $t = w[G_\epsilon(w) + 1]$ the "new" index $G_\epsilon(w) + 2$ enters $A_\epsilon(t)$. Thereafter, one old index leaves $A_\epsilon(t)$ and one new index enters $A_\epsilon(t)$ every $w$ time units until all indices $i = 1, 2, \ldots, L$ have entered $A_\epsilon(t)$. Thus, after $| A_\epsilon(t) |$ climbs at a linear rate to its maximum value $G_\epsilon(w) + 1$, it thereafter oscillates with period $w$ between $G_\epsilon(w) + 1$ and $G_\epsilon(w)$ until $t = Lw$.

*Case* 2 $[G_\epsilon(w) \geq L - 1]$: This case can be treated just as case 1 was with the following difference: The list indices $i = 1, 2, \ldots, L$ all enter $A_\epsilon(t)$ as $| A_\epsilon(t) |$ climbs to the value $L$. Thus, $| A_\epsilon(t) |$ climbs at a linear rate to the maximum value $L$, and there is no steady-state oscillatory behavior with period $w$.

Cases 1 and 2 exhaust all the possibilities for $t$ in $[0, Lw]$, so that in all cases $A_\epsilon(t)$ is a *connected* set of indices of the form

$$A_\epsilon(t) = \{k_\epsilon(t), k_\epsilon(t) + 1, \ldots, k_\epsilon(t) + r_\epsilon(t)\}$$

where $k_\epsilon(t)$ and $r_\epsilon(t)$ depend on $w$ and $L$. We summarize these facts in the following proposition.

*Proposition* 2. On trial $\mathcal{E}_1(w, W, L)$, for $t$ in $[0, Lw]$, $A_\epsilon(t)$ is a connected set of indices such that $(a)$ indices are added to $A_\epsilon(t)$ in chronological order at times $t = t_\epsilon, t_\epsilon + w, \ldots, t_\epsilon + (L-1)w$, where $t_\epsilon = \min\{t: x_1(t) = \epsilon\}$, and $(b)$ each index remains in $A_\epsilon(t)$ for $T_\epsilon - t_\epsilon$ time units, where $T_\epsilon = \max\{t: x_1(t) = \epsilon\}$. In particular, letting $G_\epsilon(w) = [(T_\epsilon - t_\epsilon)/w]$, if $G_\epsilon(w) < L - 1$, then $| A_\epsilon(t) |$ increases in unit steps every $w$ time units until $| A_\epsilon(t) | = G_\epsilon(w) + 1$. Thereafter $| A_\epsilon(t) |$ oscillates between $G_\epsilon(w) + 1$ and $G_\epsilon(w)$ with period $w$, whereas if $G_\epsilon(w) \geq L - 1$, then $| A_\epsilon(t) |$ increases in unit steps every $w$ time units until $| A_\epsilon(t) | = L$.

We now use proposition 2 to study how changes in $w$ and $L$ produce changes in the associational strengths $Z_{jk}(t)$ through time.

## F. Massed versus Distributed Practice

The association $Z_{jk}(t)$ from $r_j$ to $r_k$ grows quickly at times $t$ for which (37) holds. This means that $j$ is in $A_\epsilon(t - \tau)$, $k$ is in $A_\epsilon(t)$, and all $m \neq j, k$ are not in $A_\epsilon(t)$, for some sufficiently large $\epsilon$ which we fix once and for all. In particular, $| A_\epsilon(t) |$ is a small number, since only $j$ and $k$ can be in $A_\epsilon(t)$.

How can we guarantee that $| A_\epsilon(t) |$ be a small number? By proposition 2, the maximum of $| A_\epsilon(t) |$ in $[0, Lw]$ is $G_\epsilon(w) + 1$. We need therefore merely require that $G_\epsilon(w)$ be small. But $G_\epsilon(w) = [S_\epsilon/w]$, which is monotone decreasing in $w$. Therefore $| A_\epsilon(t) |$ will remain small for all $t$ in $[0, Lw]$ if $w$ is taken sufficiently large. One way of speeding up learning in $\mathfrak{M}(\mathfrak{U})$ is thus to slow down the rate with which list symbols are presented—that is, to "distribute practice."

Conversely, if the presentation rate of the list is fast, $G_\epsilon(w)$ will be large, and there will exist times $t$ in $[0, Lw]$ at which $|A_\epsilon(t)|$ is large. Indices that are in $A_\epsilon(t)$ at these times will not be incorporated rapidly into new associations. Thus "massed practice" can slow the learning rate.

Distributing practice will not always facilitate learning. Choosing $w \gg D(\tau, \Gamma)$ yields bad learning even though such a choice of $w$ certainly "distributes practice." The good or bad effects on learning of increasing $w$ correspond to two different factors:

1. *Distinguishability.* Increasing $w$ decreases $G_\epsilon(w)$ and thereby keeps $|A_\epsilon(t)|$ small in $[0, Lw]$. Thus only a few $x_i(t)$ are large at any time, and these can easily be distinguished from the many small $x_i(t)$ by the associational strengths.

2. *Correlations.* Choosing $w \gg D(\tau, \Gamma)$ means that all products $[x_j(t-\tau) - \Gamma]^+ x_k(t)$ with $k \neq j$ are always small, and thus all $Z_{jk}(t)$ remain approximately constant.

Distributing practice helps learning only if good distinguishability *and* good correlations prevail—that is, if $G_\epsilon(w)$ is small *and* $w \cong D(\tau, \Gamma)$. Since $G_\epsilon(w) = [S_\epsilon/w]$, $G_\epsilon(w)$ is small and $w \cong D(\tau, \Gamma)$ only if $S_\epsilon \cong w \cong D(\tau, \Gamma)$.

## G. Contiguity versus Connectedness

The above analysis shows that *contiguous* symbols, such as $r_{j-1}$, $r_j$, and $r_{j+1}$, are most likely to enter into associations with one another. This is because $A_\epsilon(t)$ is always a *connected* set. For example, in order that $Z_{jk}(t)$ grow rapidly, $x_j(t-\tau)$ and $x_k(t)$ must be large, and $|A_\epsilon(t)|$ must be small. Since this is best guaranteed when $S_\epsilon \cong w$, no index $i$ remains in $A_\epsilon(t)$ much longer than $w$ time units. By proposition 2, we also know that indices are added to $A_\epsilon(t)$ in chronological order. Since fast learning requires that $j$ be in $A_\epsilon(t-\tau)$ and $k$ be in $A_\epsilon(t)$, we conclude that $Z_{jk}(t)$ will grow fastest if $k \cong j+1$—that is, if $r_j$ and $r_k$ are contiguous.

When $w$ is small [and $G_\epsilon(w)$ is large], $A_\epsilon(t)$ is still a connected set, even though there exist times, $t$, when it contains many indices. Once again contiguous associations are the strongest ones, but only in the weak sense that associations form best at any time $t$ among the indices in the connected set $A_\epsilon(t)$. In particular, associations such as $Z_{j,j+2}(t)$, $Z_{j,j+3}(t)$, and $Z_{j,j-1}(t)$ can be of substantial size, thereby reducing the size of $Z_{j,j+1}(t)$.

These facts show that a decrease in $w$ can cause a smooth change from contiguity-type conditioning to field effects among closely related items. Such field effects are closer to a Gestaltist than to a contiguity theoretical viewpoint.



## H. The Beginning and the Middle of a List

We can now define the beginning and middle of a list in a way that takes into account various learning factors. For example, apply proposition 2 to the case in which $w$ and $L$ satisfy $L \gg 1$ and $G_\epsilon(w) < L - 1$. Then three cases can occur:

*Case* 1 [$G_\epsilon(w) \gg 0$ and $G_\epsilon(w) < L - 1$]: In this case, $A_\epsilon(t)$ goes through two phases:

(*a*) A *transient* phase at times $t$ corresponding to the monotonic increase of $|A_\epsilon(t)|$ from 0 to $G_\epsilon(w) + 1 \gg 0$; and

(*b*) A *steady-state* phase at times $t$ corresponding to the periodic oscillation of $|A_\epsilon(t)|$ between $G_\epsilon(w)$ and $G_\epsilon(w) + 1$.

*Case* 2 [$G_\epsilon(w) \gg 0$ and $G_\epsilon(w) \geq L - 1$]: In this case, $A_\epsilon(t)$ goes through only a transient phase, as $|A_\epsilon(t)|$ increases toward $L$.

*Case* 3 [$G_\epsilon(w) \cong 0$]: In this case, $A_\epsilon(t)$ goes through essentially only one phase, since $|A_\epsilon(t)|$ oscillates between $G_\epsilon(w) \cong 0$ and $G_\epsilon(w) + 1 \cong 1$ at *all* times $t$ in $[0, Lw]$.

Define the (*dynamical*) *beginning* of the list $r_1 r_2 \ldots r_L$, for fixed $w$ and $L$ at times $t$ in $[0, Lw]$, by the set of symbols $r_i$ whose indices $i$ are in the *same phase* of $A_\epsilon(t)$'s development *as the index* 1 is. The (*dynamical*) *middle* of the list is the set of symbols $r_i$ corresponding to the *second* phase of $A_\epsilon(t)$'s development, whenever this phase exists. We denote the set of symbols in the dynamical beginning by $B_\epsilon \equiv B_\epsilon(w, L)$, and those in the dynamical middle by $M_\epsilon \equiv M_\epsilon(w, L)$. When $G_\epsilon(w)$ is large, several symbols will be in both $B_\epsilon$ and $M_\epsilon$. This ambiguity is in the nature of the problem.

The above definitions of $B_\epsilon$ and $M_\epsilon$ have some unusual but informative consequences. For example, (*a*) the numerical length of a list's dynamical beginning is a function of $w$; and (*b*) there exist lists that have no dynamical middle, and all of whose symbols belong to the list's dynamical beginning. Various experimental bowing effects can be conveniently summarized in terms of these definitions. For example, (*c*) symbols in the list's dynamical middle are harder to learn than symbols in the dynamical beginning.

To see this, suppose that $i$ enters $A_\epsilon$ when $|A_\epsilon|$ is large. Thus $v_i$ begins to sample other $v_j$ when $|A_\epsilon|$ is large. In particular, $v_{i+1}$ is not readily distinguishable from the many other vertices with large potentials. Hence the association $r_i \rightarrow r_{i+1}$ will receive substantial competition from other associations $r_i \rightarrow r_k$, and $Z_{i,i+1}$ will grow slowly, if at all.

By contrast, even if $G_\epsilon(w)$ is large, $x_1(t)$ is large when $|A_\epsilon(t)|$ is small,

$x_2(t)$ is large when $|A_\epsilon(t)|$ is only slightly larger, and so on. Learning in $B_\epsilon$ is therefore faster than learning in $M_\epsilon$ if $G_\epsilon(w)$ is large.

If $G_\epsilon(w)$ is small, say $\cong 0$, learning is fast throughout $B_\epsilon$, which now includes all symbols $r_i$ in the list, since $|A_\epsilon(t)|$ remains small for all $t$ in $[0, Lw]$.

The assertion that learning is faster in $B_\epsilon$ than in $M_\epsilon$ confirms the first half of the experimental bowed learning curve of Fig. 19. The assertion that case $[G_\epsilon(w) \geq L-1 \gg 0]$ is transformed into case $[L-1 > G_\epsilon(w) \gg 0]$ and finally into case $[G_\epsilon(w) \cong 0]$ as $w$ increases agrees with the experimental fact that slowing the presentation rate flattens the first part of the bowed curve.

The definitions $B_\epsilon$ and $M_\epsilon$ illustrate the interplay of temporal and geometrical factors in determining ease of learning. The size of $|A_\epsilon(t)|$ when each $v_i$ is sampling is the crucial fact to determine. The rigorous treatment of this system also considers transfer from STM to LTM, and how this transfer depends on the interplay of presentation rate of list items, decay rate of STM traces, arousal level, and sampling times (Grossberg, 1969e; Grossberg and Pepe, 1971).

## I. Where Is the End of a List?

We have thus far considered the behavior of $A_\epsilon(t)$ only for $t$ in $[0, Lw]$ and have based our definitions of $B_\epsilon$ and $M_\epsilon$ on this behavior. *No dynamical end exists in the list before time $t = Lw$, even though the last list item was presented at time $t = (L-1)w$.* This is not surprising, since a learning subject cannot know that $r_L$ is the last list item until an extra intratrial interval transpires followed by no future items. The dynamical end of a list is created only *after* time $t = Lw$, and is due to the interactions of stimulus traces $x_i(t)$ and associations $Z_{jk}(t)$ before the list is presented for the second time. To see this, let us now consider $A_\epsilon(t)$ throughout trial $\mathcal{E}_1(w, W; L)$, where, as usual, $w \leq W$—that is, throughout the time interval $[0, (L-1)w + W]$. It suffices to consider the interval $[Lw, (L-1)w + W]$. This is readily done, since all $x_i(t) \equiv 0$, $i = L+1, \ldots, n$, by (34), and thus no new indices enter $A_\epsilon(t)$ for $t$ in $[Lw, (L-1)w + W]$. Old indices continue to drop out of $A_\epsilon(t)$, however, and consequently $|A_\epsilon(t)|$ decreases in unit steps every $w$ time units. We can distinguish two cases.

*Case* 1 $[G_\epsilon(w) \gg 0]$: The steady-state phase or peak of the first transient phase of $A_\epsilon(t)$ is followed by a second transient phase during the times $t$ at which $|A_\epsilon(t)|$ decreases in unit steps at the rate $w$.

*Case* 2 $[G_\epsilon(w) \cong 0]$: Since $|A_\epsilon(t)|$ is always small in $[0, Lw]$, any decrease in $|A_\epsilon(t)|$ due to an uncompensated dropping out of indices is negligible, and so once again $A_\epsilon(t)$ has essentially only one phase.



We now define the (*dynamical*) *end* of a list as the set of symbols, if any, whose indices appear in $A_\epsilon(t)$ during its *second* transient phase. Denote the symbols in the dynamical end by $E_\epsilon \equiv E_\epsilon(w, W, L)$. We immediately conclude that:

(*a*) Learning is faster in the dynamical end of a list than in its dynamical middle. The reasoning is the same as that which showed the advantage of the beginning over the middle. Thus one can show that the middle is harder to learn than the beginning and end simply by counting the number of large stimulus traces $x_i(t)$ which the associations $Z_{jk}(t)$ must distinguish and correlate with $x_j(t-\tau)$ at any time $t$.

The distinction between a list's dynamical beginning, middle, and end is ambiguous when $G_\epsilon(w) \cong 0$. This is because $A_\epsilon(t)$ goes through essentially only one phase, and all one can say heuristically is that *all* list symbols are either in the beginning, or the middle, or the end, and that learning is satisfactory if also $w \cong D(\tau, \Gamma)$. [This statement can be modified to take into account the numerical length, $L$, of the list when the interactions, and thus accumulating noise, are introduced into the network. It is also modified when sequential STM buffers having a maximal storage capacity are included in the discussion (Atkinson and Shiffrin, 1968)]. This ambiguity therefore implies that:

(*b*) The bowed curve flattens both at its beginning and its end as the intratrial interval increases, as also occurs in Fig. 19.

## J. The Dependence of a List's End on the Intertrial Interval and Associational Span

The intertrial interval, $W$, affects $E_\epsilon(w, W; L)$ because $|A_\epsilon(t)|$ has less opportunity to decrease when $W$ is small. For example, suppose that $W = w$. Then trial $\mathcal{E}_2(w, w; L)$ begins right after trial $\mathcal{E}_1(w, w; L)$ ends. $v_1$ receives its second input pulse, $J(t - Lw)$, at time $t = Lw$, and thus the index 1 enters $A_\epsilon(t)$ on trial $\mathcal{E}_2(w, w; L)$ not longer than $w$ time units after $L$ enters $A_\epsilon(t)$ on trial $\mathcal{E}_1(w, w; L)$. Since each $x_i(t)$ with $i = 1, 2, \ldots, L$ satisfies $x_i(t) = x_1[t - (i-1)w]$, the indices $1, 2, 3, \ldots, L$ enter $A_\epsilon(t)$ on trial $\mathcal{E}_2(w, w; L)$ in chronological order at rate $w$.

Consider the effect of increasing $W$ step by step when the list has a middle; for example, let $0 \ll G_\epsilon(w) < L-1$. If $W = 2w$, then $|A_\epsilon(t)|$ decreases by 1 after its steady-state phase on trial $\mathcal{E}_1(w, 2w; L)$. Trial $\mathcal{E}_2(w, 2w; L)$ then begins, and $|A_\epsilon(t)|$ quickly rises once again to its steady-state phase. The advantage to $Z_{12}(t)$ of trial $\mathcal{E}_1(w, 2w; L)$ is not entirely destroyed on trial $\mathcal{E}_2(w, 2w; L)$, but the advantage to $Z_{23}(t)$ on trial $\mathcal{E}_2(w, 2w; L)$ is slight.

Now increase $W$ step by step. For a fixed value of $W$, $|A_i(t)|$ decreases step by step at a rate $w$ for $t$ in $[Lw, (L-1)w + W]$ to a minimum value of $\max\{0, G_i(w) + 1 - [W/w]\}$. If $W$ is chosen so large that $[W/w] = G_i(w) + 1$, then $|A_i(t)|$ decreases to 0 before trial $\mathcal{E}_2(w, W; L)$ begins. Therefore $A_i(t)$ will have essentially the same phases on trials $\mathcal{E}_2(w, W; L)$ through $\mathcal{E}_N(w, W; L)$ as it had on trial $\mathcal{E}_1(w, W; L)$. In particular, symbols that are in $B_i$, $M_i$, or $E_i$ on one trial will be in the same dynamical part of the list on all trials [except for funneling effects (Grossberg, 1969e)], and the effects on associations which characterize a given list part will be cumulative as more and more trials occur.

These mechanisms suggest formal analogs of the major bowing effects of Fig. 19. For example:

(*a*) If $L \cong 2$ and $w \cong D(r, \Gamma)$, bowing does not occur, since $|A_i(t)|$ is always small.

(*b*) Bowing occurs when $0 \ll L \leq G_i(w)$ or $0 \ll G_i(w) < L$, since then $|A_i(t)|$ achieves large values.

(*c*) The bowed curve is flattened, but raised, when $0 \ll L \leq G_i(w)$ or $0 \ll G_i(w) < L$ if $W \cong w$, since then all list symbols are usually in $M_i$.

(*d*) If for fixed $W > w$ bowing does occur, then increasing $W$ lowers the bowed curve near its numerical middle by increasing the numerical length of $B_i$ and $E_i$.

(*e*) Increasing $W$ by a fixed amount has less of a lowering effect if $w$ is large than if $w$ is small, because $G_i(w)$ is monotone decreasing in $w$.

(*f*) For fixed $w$, increasing $W$ beyond a $W_0$ such that $[W_0/w] = G_i(w) + 1$ has little lowering effect on the bowed curve, since $|A_i(t)|$ decays to zero at the end of each trial for all such $W$.

(*g*) If bowing occurs but $1 \ll L - 1 < G_i(w)$, then increasing the list's numerical length $L$, while keeping $w$ and $W$ fixed, can decrease the skewness of the bowed curve by increasing the numerical length of $M_i$.

The list's associational span and intertrial interval interact to influence the bowed curve. Consider Fig. 23. The $i$th *associational span* is the interval of sampling by $v_i$—namely, the set $\{t: x_i(t - \tau) > \Gamma\}$. By (34), this interval is $(i\tau + T_1, i\tau + T_2)$, where

$$x_i(t) = \int_0^t e^{-\alpha(t-v)} J(v)\, dv > \Gamma \qquad \text{for} \quad t \in (T_1, T_2)$$

Only those list positions whose associational span includes times when $|A_i(t)|$ is in its second transient phase are influenced by an increase in $W$. In Fig. 23*a*, these indices include all indices greater than $j$. In Fig. 23*b*, these indices include indices in $B_i$, $M_i$, and $E_i$. Without all associational spans being known, the size of $|A_i|$ gives incomplete information concerning the way in which the bow changes as a function of $W$.



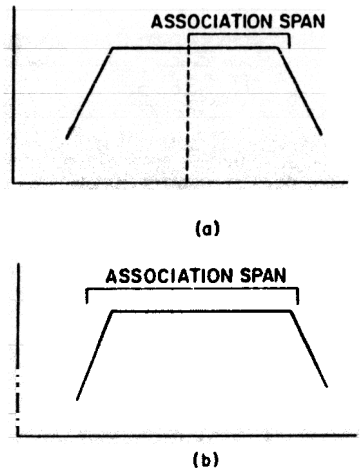


FIG. 23. Interaction of associational span and intertrial interval.

## K. Response Oscillation and Remoteness

Suppose that $1 \ll G_i(w) < L - 1$ and that $W$ is sufficiently large for some bowing to occur. Then at times $t$ when $|A_i(t)|$ is small, the formation of new associations will be restricted to a small number of indices. Thus learning will begin to show its effects faster in $B_i$ and $E_i$ than in $M_i$, and competing responses are restricted to a relatively small set of list symbols. By contrast, for $r_i$ in $M_i$, $i$ is in $A_i(t)$ when $|A_i(t)|$ is large. Competing response tendencies to a symbol in $M_i$ are therefore broadly distributed across the list. Learning therefore takes relatively long to show its effects in $M_i$, and a long time is needed to eliminate the large collection of competing response tendencies after learning begins. These are the main effects of Fig. 20. The analysis can be refined by studying the shape of the generalization gradients at each list position.

## L. Overarousal and Inverted U in Learning

Each $v_i$ can sample all $v_k$ with $k \leq i - 1$, but not necessarily any $v_k$ with $k \geq i + 1$ other than $v_{i+1}$. That is to say, when associations are being formed with $r_i$, different information is available in the network concerning the past than the future. In fact, if $J(t)$ is a rectangular input pulse of

intensity $J$ and duration $\lambda$, then the associational span has length

$$s = \lambda + \frac{1}{\alpha}\log\left[\left(\frac{J}{\alpha\Gamma} - 1\right)(1 - e^{-\alpha\lambda})\right] \tag{40}$$

which is monotone decreasing in the signal threshold $\Gamma$. As $\Gamma$ decreases, more forward associations, $r_i \to r_k$, $k > i + 1$, can form, thereby reducing the relative strength of $r_i \to r_{i+1}$. This does not mean, however, that increasing $\Gamma$ always improves learning of $r_i \to r_{i+1}$. If $\Gamma$ is too large, then, even though no forward associations can compete with $r_i \to r_{i+1}$, nonetheless $[x_i(t) - \Gamma]^+$ is usually zero or small in value, so that little learning of $r_i \to r_{i+1}$ occurs. Thus there exists an optimal region of threshold choice that reduces response interference without unduly diminishing the rate of learning. Alternatively expressed, this optimal region maximizes distinguishability of the correct association while providing enough energy to drive the learning process.

Notice that decreasing $J$ in Eq. (40) has the same qualitative effect as increasing $\Gamma$. Thus all our statements concerning threshold regulation given fixed levels of physiological excitation can be transformed into corresponding statements concerning variations in the level of excitation ("arousal") as it compares with the system's fixed threshold parameters.

### M. Skewing

The fact that the middle of the list is harder to learn than either end is the net result of two effects in the bare field of $\mathfrak{M}(\mathfrak{U})$. First, as list position $i$ increases, there always exist more backward associations, $r_i \to r_k$, $k < i$, that compete with $r_i \to r_{i+1}$, thereby increasing learning difficulty. Second, there exist fewer forward associations, $r_i \to r_{i+1}$, thereby decreasing learning difficulty. However, by varying the associational span, we can guarantee that no forward association ever competes with $r_i \to r_{i+1}$ for any $i$. For example, choose $\Gamma$ so large that $[x_i(t) - \Gamma]^+ = 0$ whenever $x_k(t) > 0$ and $k > i + 1$. Then the associations $r_i \to r_k$ never form, and consequently the major effect on the association $r_i \to r_{i+1}$ as $i$ increases is to increase response interference due to increasing numbers of backward response alternatives. Apart from such degenerate cases, however, it can be proved that bowing always occurs in the bare field. Indeed, letting

$$\mathfrak{B}(i, \Gamma) \equiv \lim_{t\to\infty} Z_{i,i+1}(t), \qquad i = 1, 2, \ldots, L - 1$$

one can prove that, for any fixed $\Gamma \geq 0$, $\mathfrak{B}(i, \Gamma)$ either first decreases and then increases as $i$ increases from 1 to $L$, or the degenerate case occurs in which $\mathfrak{B}(i, \Gamma)$ is monotone decreasing. By definition, for fixed $\Gamma$, the bow occurs at the list position $M(\Gamma)$ for which $\mathfrak{B}(i, \Gamma)$ is a minimum. If there exists more than one such position, we let $M(\Gamma)$ be the largest one, since in the presence of nonlinear interactions, background noise can only increase as more events are presented.



In the bare field, $M(\Gamma)$ is a monotone increasing function of $\Gamma$. Furthermore, $M(0) = \frac{1}{2}(L - 1)$ if $L$ is odd and $M(0) = \frac{1}{2}L$ if $L$ is even (Grossberg, 1969e). In the degenerate case above, $M(\Gamma) = L$ for sufficiently large $\Gamma$. Thus maximal difficulty in learning can occur at any list position greater than the list's numerical middle. Since "normal" learning requires a positive $\Gamma$, the bow will occur nearer to the end than to the beginning of the list, and the bowed curve will therefore be skewed.

At times $t < \infty$, let $\mathfrak{B}(i, \Gamma, t) = Z_{i,i+1}(t)$, and suppose that $\min_i \mathfrak{B}(i, \Gamma, t)$ occurs at list position $M(t, \Gamma)$ for every fixed $t$ and $\Gamma$. Then for fixed $\Gamma$, $M(t, \Gamma)$ ultimately decreases from $M(t, \Gamma) = L$ to $M(t, \Gamma) = M(\Gamma)$ as $t$ increases beyond the time at which $r_L$ is presented to infinity (Grossberg and Pepe, 1971). This happens because the *non*occurrence of the events $r_{L+1}, r_{L+2}, \ldots, r_n$ gradually decreases the relative amount of response interference to $r_{L-1} \to r_L$ growth, since the future associations $r_{L-1} \to r_k$, $k > L$, never form as $t$ increases. Thus skewing can depend both on $\Gamma$ and on the intertrial interval. If $\Gamma$ is very large, the intertrial interval effect will be negligible.

## VIII. Instrumental Conditioning

### A. Additional Postulates

The derivation of Section II can be supplemented by additional postulates that lead to mechanisms of reinforcement, drive, and incentive motivation. The first of these postulates are the following:

*Postulate* 1. Practice makes perfect.

*Postulate* 2. The time lags between CS and UCS on successive learning trials can differ.

*Postulate* 3. After learning has occurred, the UCR can be elicited by the CS alone on recall trials.

*Postulate* 4. A given CS can be conditioned to any of several drives (for example, bell → salivation if the UCS is food, or bell → fear if the UCS is a shock).

*Postulate* 5. Amount and/or rate of responding is influenced by the state of deprivation.

Postulate 1 is a truism that will be implemented in conjunction with postulate 2. Postulates 2 and 3 are observations about the Pavlovian condi-

tioning paradigm. Postulates 4 and 5 are obvious facts. Such trivialities would yield little directive in a theoretical vacuum. Applied to the theory already derived, however, they are powerful guides to constructive theorizing.

### B. UCS-Activated Nonspecific Arousal of CS-Activated Sampling Cells

Consider the typical situation in which a spatial pattern to be learned is embedded in a space–time pattern presented to $\mathcal{B}$, and the space–time pattern can be different on successive learning trials. Alternatively, one could let the UCS be the space–time pattern, and could consider the problem of learning a particular spatial pattern of the UCS perfectly by practicing the UCS several times. How is a particular event in a stream of events picked out as significant and learned? To simplify our notation, we suppose that the same space–time pattern is presented on each trial. Thus, on each trial a sequence $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots, \theta^{(N)}$ of spatial patterns with weights $\theta^{(k)} = \{\theta_i^{(k)} : i \in I\}$ is the UCS delivered to $\mathcal{B}$, $k = 1, 2, \ldots, N$. In this situation, an outstar anatomy does not suffice to achieve postulate 1 if postulate 2 also holds; that is, a given sampling cell, $v_j$, in $\mathcal{A}$ cannot learn a definite spatial pattern, $\theta^{(m)}$, chosen from the UCS sequence if the CS alone can fire $v_j$ on successive learning trials. To see this, consider sampling by $v_j$ of $\theta^{(1)}$ for definiteness. The sampling cell $v_j$ can learn $\theta^{(1)}$ only if $v_j$ fires briefly a fixed time before the onset of $\theta^{(1)}$ on *every* trial, and if the signals from $v_j$ reach $\mathcal{B}$ only when $\theta^{(1)}$ plays on $\mathcal{B}$. This will not happen if the CS alone can fire $v_j$ while postulate 2 holds, since signals from $v_j$ will reach $\mathcal{B}$ on successive trials while spatial patterns $\theta^{(k)}$ other than $\theta^{(1)}$ play on $\mathcal{B}$. Thus the stimulus sampling probabilities $Z_j = (Z_{ji} : i \in I)$ will learn a weighted average of the patterns $\theta^{(k)}$ rather than $\theta^{(1)}$.

To avoid noisy sampling, the outstar must be embedded in a larger network. The sampling cell $v_j$ must be prevented from firing unless it simultaneously receives a CS input *and* an input controlled by the UCS which signals that the UCS will arrive at $\mathcal{B}$ a fixed time interval later. This is accomplished in two steps: Let the UCS activate axons leading to $v_j$ that deliver an input to $v_j$ a fixed time before the UCS arrives at $\mathcal{B}$; and set the common spiking threshold, $\Gamma_j$, of all $v_j$'s axon collaterals so high that $v_j$ can fire only if it simultaneously receives large CS- and UCS-controlled inputs. Then, on every trial, $v_j$ can fire and begin to sample the spatial pattern $\theta^{(1)}$ as it arrives at $\mathcal{B}$, if also the CS has been presented. Grossberg (1970a) discusses an inhibitory mechanism that guarantees brief $v_j$ outputs in response to even prolonged CS plus UCS inputs; sampling can therefore terminate before $\theta^{(2)}$ occurs at $\mathcal{B}$.



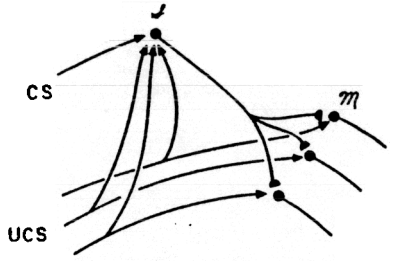


FIG. 24. UCS-activated nonspecific arousal of CS-activated sampling cells.

All cells in the network that can sample $\mathcal{B}$ receive UCS-activated axons, for the reasons given above. In other words, there exists a UCS-activated nonspecific arousal of CS-activated sampling cells. These cells are polyvalent cells, or cells that are influenced by more than one modality, such as the sound of a bell (CS) and the smell of food (UCS). The polyvalent cells fire only if the sum of CS and UCS inputs is sufficiently large. Grossberg (1971b) reviews physiological data relevant to this concept.

Some suggestive terminology is now introduced by denoting sampling cells $\mathcal{A}$ generically by $\mathcal{S}$, for "sensory cells" or "sensory representation," and sampled cells $\mathcal{B}$ by $\mathcal{M}$ for "motor cells" or "motor representation." This distinction has no absolute significance, of course, since both $\mathcal{A}$ and $\mathcal{B}$ contribute to sensory and motor processing. It is nonetheless convenient (see Fig. 24).

### C. Conditioned Reinforcers

Postulate 3 is invoked on recall trials. After learning has taken place, the CS alone can elicit performance on recall trials. Thus the CS alone can fire cells in $\mathcal{S}$ on recall trials. But $\mathcal{S}$ cells can fire only if inputs along two axon paths converge simultaneously on them. The UCS is not available on recall trials to activate one of these paths. Only the CS is available. How does CS–UCS pairing on learning trials enable the CS to gain control over the UCS $\rightarrow \mathcal{S}$ pathway on recall trials? This dilemma imposes the concept of "conditioned arousal," which will later be specialized as "conditioned incentive motivation." Namely, CS–UCS pairing during learning trials allows the CS to gain control over the nonspecific arousal channel via Pavlovian conditioning (that is, by cross-correlating presynaptic spiking frequencies and postsynaptic potentials at suitable synaptic knobs). Conditioning of nonspecific arousal at these synaptic knobs takes place while specific motor patterns are learned in the $\mathcal{S} \rightarrow \mathcal{M}$ synaptic knobs. Consequently, on recall trials, the CS can activate two input channels:

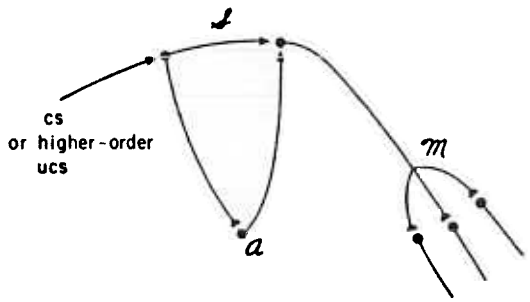


FIG. 25. Minimal nonrecurrent interaction between external cue and arousal source

unconditioned specific inputs to $\mathcal{S}$, and conditioned nonspecific arousal inputs to $\mathcal{S}$. At cells in $\mathcal{S}$ where these two inputs converge, the cell potential can be driven above its spiking threshold. These cells can fire, yielding signals along $\mathcal{S} \rightarrow \mathfrak{M}$ axons which activate the $\mathcal{S} \rightarrow \mathfrak{M}$ synaptic knobs and reproduce at $\mathfrak{M}$ the patterns encoded in these knobs. In this way, a CS can acquire UCS properties, and thus aspects of higher-order conditioning emerge as a consequence of postulates 2 and 3.

After a CS can activate the arousal pathway, it has UCS properties; it can serve as the UCS for a new CS in a later learning experiment. The transition from CS to UCS in these networks is effected by an alteration (not necessarily a strengthening!) of extant pathways, rather than by the creation of new pathways. Thus both CS and UCS inputs are processed in parallel pathways ("path equivalence"), except possibly the primary UCS input (for example, taste of food) on which a chain of conditioning experiments can be built. In particular, "higher-order" UCS inputs, as well as CS inputs, are delivered to $\mathcal{S}$.

## D. Arousal Cells

The cells $\mathcal{A}$ at which conditioning of arousal takes place are neither $\mathcal{S}$ cells nor $\mathfrak{M}$ cells. This is because the $\mathcal{S}$ cells must be aroused before they sample the activity of $\mathfrak{M}$ cells, and $\mathfrak{M}$ cell activation must await the onset of sampling—and thus prior firing—by $\mathcal{S}$ cells, or else $\theta^{(1)}$ cannot be learned. Similar arguments have been used to prove that at least two successive cell sites are needed in each sensory representation. The first site receives the CS input and thereupon sends signals to $\mathcal{A}$ and to the second site. The second site can fire to $\mathfrak{M}$ only if it also receives a feedback signal from $\mathcal{A}$ (see Fig. 25). Sensory representations with more than two cell sites are also possible, but the theory restricts itself to the construction of minimal anatomies. As new requirements are imposed, the anatomy can be expanded to include new properties.



The $\mathcal{A}$ cells can be interpreted as network analogs of hypothalamus, reticular formation, and related brain areas implicated in arousal and reinforcement tasks. Certainly $\mathcal{A}$ cells are at best rudimentary analogs of these neural regions. Nonetheless, the formal tasks that $\mathcal{A}$ cells perform will be seen to be strikingly reminiscent of facts known about their neural counterparts. Moreover, the interactions between $\mathcal{A}$ cells will become increasingly complex and realistic as the derivation continues.

## E. Existence of Several Drives

The $\mathcal{A}$ cells include drive-activated cells. For example, when a bell (CS) is conditioned to elicit salivation (UCR), it activates the $\mathcal{A}$ cells corresponding to hunger. Now invoke postulate 4. Postulate 4 directs us to further expand the minimal network to include several subsets of $\mathcal{A}$ cells, such that each subset subserves a different "drive." These $\mathcal{A}$ subsets can overlap if their corresponding drives are not mutually independent—for example, hunger and thirst. For convenience of representation, however, we draw them as individual points in Fig. 26. By postulate 4, a given sensory event can be conditioned to any of several drive contingencies. Thus, each $\mathcal{S}$ in the minimal construction will send axons to several subsets of $\mathcal{A}$ cells. Each $\mathcal{A}$ subset, in turn, sends axons nonspecifically to $\mathcal{S}$ cells; otherwise the several drives could not control nonspecific arousal signals from $\mathcal{A}$ to $\mathcal{S}$ capable of releasing signals in particular $\mathcal{S} \rightarrow \mathfrak{M}$ pathways (see Fig. 26).

## F. Drive Inputs

Postulate 5 imposes a new constraint on the firing of $\mathcal{A}$ cells. If an $\mathcal{A}$ cell can always fire in response to conditioned arousal inputs from $\mathcal{S}$ cells alone, then an $\mathcal{A}$ cell can always elicit (say) hunger-specific motor activity, even if $\mathcal{O}$ is not hungry, whenever food is presented. This property would kill $\mathcal{O}$. The difficulty is formally analogous to allowing an $\mathcal{S}$ cell to fire in the absence of its CS input. Maladaptive $\mathcal{A}$ cell firing of this kind can be easily

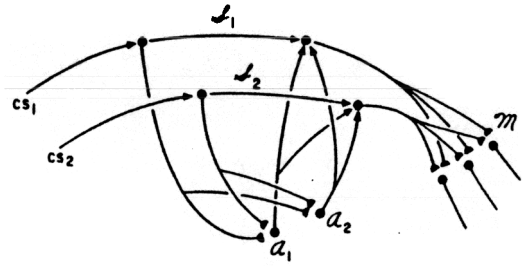


FIG. 26. Sampling of spatially distributed drive representations.

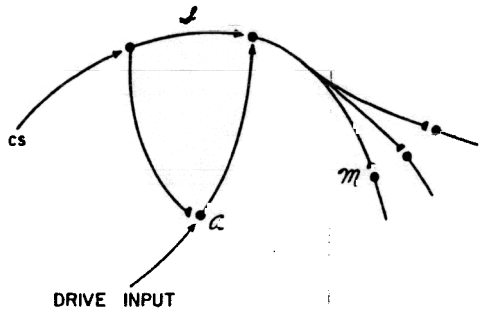


FIG. 27. Existence of drive inputs.

prevented, just as in the $\mathcal{S}$ cell case. In the $\mathcal{S}$ cell case, an $\mathcal{S}$ cell can fire to $\mathfrak{M}$ only if it simultaneously receives a nonspecific input from $\mathcal{A}$ and a specific sensory input. Require analogously that an $\mathcal{A}$ cell can fire only if it simultaneously receives a nonspecific input from $\mathcal{S}$ (for example, a conditioned input from $\mathcal{S}$ or a primary UCS input) and a specific sensory input. In the $\mathcal{A}$ cell case, the sensory input is interpreted to be a drive input whose source is within $\mathcal{O}$. The size of this input indicates the level of this drive in $\mathcal{O}$ through time. This restriction on $\mathcal{A}$ cell firing is achieved by setting the spiking threshold of $\mathcal{A} \rightarrow \mathcal{S}$ axons so high that only the sum of sufficiently large inputs from $\mathcal{S}$ and from internal drive sources can fire an $\mathcal{A}$ cell (see Fig. 27). Now $\mathcal{A}$ cells are also "sensory" cells, but their sensory inputs describe the internal state of $\mathcal{O}$ rather than the external state of the world.

Grossberg (1971b) develops those ideas and cites relevant data. Noteworthy is the possibility of learning to push a lever persistently to deliver electric shocks to a (consummatory) drive representation without reducing the internal drive input (no "drive reduction"), as Olds and his collaborators have reported (Olds, 1955).

Various psychological terms can be used to describe $\mathcal{A}$ cells. They supply "incentive motivation" in support of learned sensory-motor acts encoded in $\mathcal{S} \rightarrow \mathfrak{M}$ pathways. They resemble the "amplifier" elements of Estes (1969), the "Go" mechanism of Miller (1963), and the Now Print mechanism of Livingston (1967).

The foregoing construction is supported by rigorous mathematical theorems. For example, in Fig. 26, any number of cells in $\mathcal{S}$ can sample any number of cells in $\mathcal{A}$, where the $\mathcal{A}$ cells can receive primary UCS inputs, internal drive inputs, and/or conditioned inputs. This situation is covered by theorems in Grossberg (1969d, 1971c, 1972b) on nonrecurrent sampling. The same theorems cover the case of $\mathcal{S} \rightarrow \mathfrak{M}$ sampling. These are the only places in Fig. 26 where learning occurs. (Actually, learning in the feedback pathway $\mathcal{A} \rightarrow \mathcal{S}$ is needed in more advanced discussions.) It remains only



to guarantee that thresholds and other parameters can be set to restrict the times at which $\mathcal{S} \rightarrow \mathcal{A}$, $\mathcal{A} \rightarrow \mathcal{S}$, and $\mathcal{S} \rightarrow \mathfrak{M}$ signals occur. Some further network structure is needed, and is discussed in Grossberg (1972d).

## G. Suppression by Punishment

The previous discussion yields a network $\mathcal{O}$ which can learn and perform consummatory responses under suitable constraints. This construction does not suffice to prevent a consummatory response if environmental contingencies change so that the response yields aversive results. The construction will now be extended to include this crucial possibility. We shall consider the following situation for definiteness. Suppose that a CS (bell) which was once a cue for food is now a cue for shock. How does $\mathcal{O}$ prevent itself from inappropriately carrying out food-consummatory behavior in response to the CS and thereby getting shocked? To implement our construction we shall use the following postulate, which prevents $\mathcal{O}$ from indiscriminately learning unsuccessful responses.

*Postulate* 6. $\mathcal{O}$ does not (readily) learn escape responses that do not terminate shock.

The construction is, of course, constrained by the network that has already been derived, since the postulates from which this network emerged still hold. In Fig. 26, consummatory behavior is modifiable by two parallel conditioning processes: conditioning of nonspecific $\mathcal{A} \rightarrow \mathcal{S}$ arousal via the $\mathcal{S} \rightarrow \mathcal{A}$ synaptic knobs, and conditioning of specific motor patterns via the $\mathcal{S} \rightarrow \mathfrak{M}$ synaptic knobs. Which of these conditioning processes must be supplemented to fulfill postulate 6? We proceed by asking for the minimal possible change: Can $\mathcal{O}$ recondition the $\mathcal{S} \rightarrow \mathfrak{M}$ pathway without altering the $\mathcal{S} \rightarrow \mathcal{A}$ pathway? The answer will be "No" for the following reasons. The $\mathcal{S} \rightarrow \mathfrak{M}$ pathway can be reconditioned in two ways:

### 1. *Passive Extinction*

Prevent firing of the $\mathcal{S} \rightarrow \mathfrak{M}$ pathway for long time intervals. Then transmitter levels in $\mathcal{S} \rightarrow \mathfrak{M}$ synapses can slowly decay to the level of network random noise. This process takes too long, however, to prevent $\mathcal{O}$ from violating postulate 6, and there exist workable transmitter laws in which no passive extinction occurs; for example, laws such as

$$\dot{z}_{jk}(t) = \{-\delta_{jk} z_{jk}(t) + \epsilon_{jk} x_k(t)\}[x_j(t - \tau_{jk}) - \Gamma_{jk}]^+$$

in which perfect memory exists until practice or recall trials, or random bursts of presynaptic spiking, occur. Also, decay can be retarded or even

reversed if recall trials intermittently occur when $\mathcal{O}$ is hungry (cf. Section IV). Then the $\mathcal{S} \rightarrow \mathfrak{M}$ pathway is activated and the $\mathcal{S} \rightarrow \mathfrak{M}$ synaptic levels are restored to supranoise levels by transmitter potentiation, without destroying the encoded motor pattern ["post-tetanic potentiation" (Eccles, 1964)].

2. *Interference Theory of Forgetting* (Adams, 1967)

Let every occurrence of shock input generate a new UCR pattern at $\mathfrak{M}$, which is incompatible with eating. If the CS also occurs at these times, and $\mathcal{O}$ is hungry, then $\mathcal{S}$ will sample the new pattern at $\mathfrak{M}$, and the $\mathcal{S} \rightarrow \mathfrak{M}$ synaptic knobs will encode the new UCR pattern. Thereafter, whenever the bell rings and $\mathcal{O}$ is hungry, the new motor pattern will be released, rather than eating. This mechanism has severe faults during recall trials. First, $\mathcal{O}$ cannot learn specific avoidance tasks, since the shock—and not a specific avoidance response—controls the competing UCR at $\mathfrak{M}$. Second, $\mathcal{O}$ remains conditioned to the hunger $\mathcal{A}$ cells. Thus $\mathcal{O}$ will indulge in general (for example, autonomic) preparations for eating without being able to eat. Third, $\mathcal{O}$ is maladaptively fearless, since only positive consummatory drives are conditionable to the CS. Counterconditioning along a new $\mathcal{S} \rightarrow \mathcal{A}$ pathway is clearly needed. Denote the new subset of $\mathcal{A}$ cells by $\mathcal{A}_f$.

Let shock create an input at the subset $\mathcal{A}_f$. Let this input be a monotone increasing function of shock intensity. Again we are called upon to psychologically interpret a formal operation. In this case, associate activation of the cells $\mathcal{A}_f$ by shock with production within $\mathcal{O}$ of a comparable amount of fear. This interpretation introduces fear into the network using a minimum of network machinery. Given this interpretation, activating conditioned $\mathcal{S} \rightarrow \mathcal{A}_f$ synaptic knobs will yield a CER [conditioned emotional response (Estes, 1969)], both by eliciting fear in $\mathcal{O}$ and, perhaps, by activating autonomic expressions of fear through $\mathcal{A}_f$. Let $\mathcal{A}_h$ denote the subset of $\mathcal{A}$ cells that subserves hunger, and consider postulate 6 in this context.

Why is postulate 6 needed? Suppose that it does not hold. Then $\mathcal{O}$ can learn all unsuccessful escape responses. Efficient avoidance performance would therefore be unlikely, since mistakes are more likely than correct response during a period of frantic trial and error in a complex experimental chamber. At best, $\mathcal{O}$ would learn to execute the avoidance response as the terminal response in a long chain of previously learned incorrect responses. To prevent this from happening, $\mathcal{A}_h$ cells cannot be the only $\mathcal{A}$ cells that fire to $\mathcal{S}$ when the CS occurs and shock is on. For, if they were, not only could maladaptive consummatory responses be performed give the CS and sufficient hunger, but also all erroneous escape responses could be sampled and learned by $\mathcal{S} \rightarrow \mathfrak{M}$ synaptic knobs with the $\mathcal{A}_h$ cells as the arousal



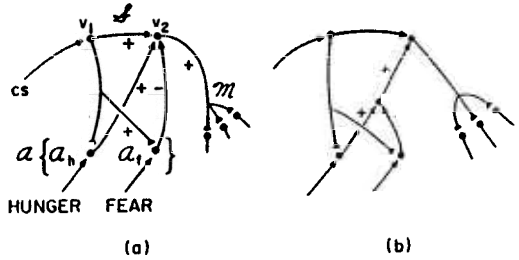


FIG. 28. Net incentive-motivational feedback.

source. The effect of $\mathcal{A}_h$ arousal on $\mathcal{S}$ must be inhibited while shock is on. The $\mathcal{A}_f$ cells are the minimal source of this inhibition. Hunger and fear arousal cells thus reciprocally inhibit each other, as Logan (1969) suggested in his discussion of net incentive motivation. Figure 28 displays two inhibitory mechanisms. Consider Fig. 28*a* when the synaptic knobs of $v_1$ are active. At these times, the sampling probabilities $Z_1(t)$ learn a weighted average of the spatial patterns $\theta(t) = [\theta_h(t), \theta_f(t)]$ that reach $\mathcal{A}_h$ and $\mathcal{A}_f$. Thus the probabilities learn the net balance of hunger and fear during times when $v_1$ samples $\mathcal{A}$. When $v_1$ fires and $\mathcal{O}$ is hungry, $\mathcal{A}_h$ sends excitatory feedback signals to $v_2$, whereas $\mathcal{A}_f$ sends inhibitory signals to $v_2$. Cell $v_2$ requires the sum of two excitatory inputs—one from $v_1$ and one from $\mathcal{A}_h$—in order to fire. As the relative strength of the inhibitory signal from $\mathcal{A}_f$ grows, it cancels the effect of the $\mathcal{A}_h$ input and prevents $v_2$ from firing. Thus $v_2$ cannot sample and learn the motor patterns reaching $\mathcal{A}$ at times when $\mathcal{A}_f$ feedback is active. This is true of every sensory representation.

Five conclusions follow: (1) An intense shock can suppress consummatory behavior by competing with $\mathcal{A}_h \rightarrow \mathcal{S}$ arousal via the inhibitory $\mathcal{A}_f \rightarrow \mathcal{S}$ pathway. (2) This suppression does not extinguish memory of the patterns already encoded in the $\mathcal{S} \rightarrow \mathfrak{M}$ synaptic knobs. (3) Suppression can take place faster than passive extinction. (4) An intense shock can prevent new $\mathcal{S} \rightarrow \mathfrak{M}$ associations from forming by inhibiting release of sampling signals from $\mathcal{S}$. (5) After $\mathcal{S} \rightarrow \mathcal{A}_f$ conditioning takes place, properties (1) through (4) can be elicited on recall trials wherever the CS input activates $\mathcal{S} \rightarrow \mathcal{A}_f$ synapses.

Similar qualitative properties hold for Fig. 28*b*. Here, however, the $\mathcal{A}_f$ and $\mathcal{A}_h$ signals compete with each other at a second stage of processing before a signal to $\mathcal{S}$ is emitted. It can be proved that only $\mathcal{A}_h$ can create an input (excitatory) to $\mathcal{S}$, and does so only if it emits a stronger signal than $\mathcal{A}_f$ does. The competitive mechanism is called a subtractive on-center off-surround field. Its mathematical properties have been discussed

(Grossberg, 1970a). Figure 28*b* requires half as many $\mathfrak{A} \rightarrow \mathfrak{S}$ axons as Fig. 28*a*. This represents a considerable saving of axons, since each $\mathfrak{A}$ subset projects nonspecifically to numerous $\mathfrak{S}$ cells. On the other hand, Fig. 28*a* requires fewer cellular processing stations.

## H. Avoidance: Heuristics

The following postulate is essentially a rewording of postulate 6.

*Postulate* 7. $\mathfrak{O}$ learns escape responses that do terminate shock faster than it learns escape responses that do not terminate shock.

This postulate also builds upon mechanisms that are already at our disposal. In particular, while shock is on, $\mathfrak{S} \rightarrow \mathfrak{M}$ sampling is prevented by $\mathfrak{A}_f \rightarrow \mathfrak{S}$ inhibition. Shock termination removes $\mathfrak{A}_f \rightarrow \mathfrak{S}$ inhibition, but $\mathfrak{S} \rightarrow \mathfrak{M}$ sampling remains impossible until *some* excitatory arousal source is activated. Postulate 7 can thus be reduced to the following question: What excitatory arousal source releases $\mathfrak{S} \rightarrow \mathfrak{M}$ sampling just after shock is turned off, and thereby establishes conditioned pathways from the sensory cues that are available when the avoidance response occurs to both the active arousal source and the motor controls of the avoidance response? Speaking heuristically, this arousal source provides the "motivational support" for learning the avoidance response. We suggest that an experimental analog of exciting this new arousal source is, other things being equal, an internally perceived "relief" from fear (Denny, 1971; Masterson, 1970; Reynierse and Rizley, 1970).

Denote by $\mathfrak{A}_f^-$ the arousal cells which are excited by termination of shock input to the cells $\mathfrak{A}_f$, which we henceforth denote by $\mathfrak{A}_f^+$. Some formal requirements must be imposed on $\mathfrak{A}_f^-$ and $\mathfrak{A}_f^+$ to ensure that the arousals work together effectively. First, require that excitation of $\mathfrak{A}_f^-$ by shock termination is transient. Transient response is needed to prevent irrelevant sensory-motor coordinations from being learned whenever shock is off. The cells $\mathfrak{A}_f^+$ are *on-cells*; they are turned on by shock, and they remain on until shock is shut off. The cells $\mathfrak{A}_f^-$ are *off-cells*; they are turned on temporarily by shock termination. On-cells and off-cells are familiar physiological components (Thompson, 1967, pp. 253 and 349). Second, require that the outputs from $\mathfrak{A}_f^+$ to $\mathfrak{A}_f^-$ reciprocally inhibit each other before they send signals to $\mathfrak{S}$. Thus these outputs interact to form a consensus between "fear" and "relief." A possible behavioral analog of this rebound from $\mathfrak{A}_f^+$ on-cells to $\mathfrak{A}_f^-$ off-cells is the rebound in behavioral effects reported to occur after electrical hypothalamic stimulation terminates (Cox *et al.*, 1969; Grastyan, 1968; Valenstein *et al.*, 1969). This



analogy receives further support from a chemical and anatomical analogy which is developed in Grossberg (1972d) between the twofold system $(\mathfrak{A}_f^+, \mathfrak{A}_f^-)$ and sites in the twofold system of ventromedial and lateral hypothalamus.

The network must be expanded once again to allow $\mathfrak{S}$ to become conditioned to the new arousal source. Thus let each sensory representation, $\mathfrak{S}$, send axons to $\mathfrak{A}_f^-$ as well as to $\mathfrak{A}_f^+$, $\mathfrak{A}_h$, and other $\mathfrak{A}$ cell clusters. At any time, the synaptic knobs of each $\mathfrak{S}$ encode a spatial pattern derived from the patterns $\theta(t) = [\theta_f^+(t), \theta_f^-(t), \theta_h(t), \ldots]$. This pattern describes the net balance of excitatory and inhibitory $\mathfrak{A} \rightarrow \mathfrak{S}$ feedback that this representation controls. It is determined by a weighted average of the spatial patterns $\theta(t)$ that reach $\mathfrak{A}$ when the given $\mathfrak{S}$ is sampling.

In summary, the classical notion that instrumental reinforcement is due to "drive reduction" when shock terminates is replaced by rebound from negative-incentive motivational on-cells to positive-incentive motivational off-cells when shock terminates. The balance of excitation of on-cells and off-cells can be classically conditioned, perhaps at different times, to any $\mathfrak{S}$ representations. The net $\mathfrak{A} \rightarrow \mathfrak{S}$ output, and thus $\mathfrak{S} \rightarrow \mathfrak{M}$ firing and performance on recall trials, is determined by *all* the $\mathfrak{S}$ sites that fire to $\mathfrak{A}$ at such times. Even if half of $\mathfrak{S}$ fires to $\mathfrak{A}_f^-$, no $\mathfrak{S} \rightarrow \mathfrak{M}$ channel need be activated by positive $\mathfrak{A} \rightarrow \mathfrak{S}$ feedback if the other half fires to $\mathfrak{A}_f^+$, since $\mathfrak{A}_f^-$ and $\mathfrak{A}_f^+$ will reciprocally inhibit each other's outputs. Similarly, shock termination yields little "relief" if it is antagonized by a switching-on of new $\mathfrak{S} \rightarrow \mathfrak{A}_f^+$, or "fear," channels. Shock termination per se is not necessarily "drive reducing."

Recent psychophysiological data and concepts can be qualitatively analyzed in terms of these network analogs (see Grossberg, 1972c). These concepts include aspects of the following: relaxation, or elicitation, theory, which claims that an unconditioned response of relief precedes reinforcement; the concept of "effective reinforcement," which notes that shock offset and onset of fearful situational cues can influence reward in opposite ways, as is illustrated by two-way avoidance tasks in which a rat escapes a chamber in which it is shocked by running into another chamber where it was previously shocked; classical and instrumental properties of a CS+ paired with shock, a CS− paired with no-shock, and feedback stimuli contingent on the avoidance response, including transfer of their effects from classical to instrumental conditioning experiments; autonomically nonchalant asymptotic avoidance performance originally motivated by fear; forced extinction of the CAR without fear extinction; response suppression without an avoidance response; relief without an avoidance response; opposite effects of contingent and noncontingent punishment on fear and suppression of consummatory responding; punishment hypothesis

of avoidance learning, describing rewarding effects of terminating proprioceptive cues that correspond to nonavoidance responses; response (or no-response) generalization from one shock level to a different level; and rewarding effects of response-contingent reduction in frequency of shock.

The argument leading to an explicit construction of the rebound mechanism falls into eight main stages.

1. *Existence of a Tonic Input*

When shock terminates, $a_f^-$ emits a transient output. Thus, by Eq. (22), the potentials $x_f^-$ of $a_f^-$ cells grow transiently to suprathreshold values. In Eq. (22), an input source is required to thusly perturb $x_f^-$. What input source does the job? (The concept of "input source" includes possible energy sources within the cells themselves.)

In these systems, shutting off one input (such as the shock input to $a_f^+$) does not provide energy for turning on another input (such as the one driving $a_f^-$ rebound). Terminating shock input can, however, unmask the effects of an *internally* driven input to $a_f^-$ whose influence is inhibited by shock. The internal source of $a_f^-$ input is therefore neither turned on nor off by shock offset. It is not turned off by shock onset, since then it would be off at shock offset, and could not drive $a_f^-$ rebound. Finally, if it is turned on by shock onset, or is unaffected by shock onset, then it is always on. The internal input is therefore *tonic.*

2. *Existence of Accumulation–Depletion*

Output from $a_f^-$ shuts off soon after it is turned on. How is this done? No externally driven input is available to do this. The $a_f^-$ output is depleted by its own activity. In other words, while shock is on, an accumulation process occurs at $a_f^-$. When shock is off, output from $a_f^-$ is a monotone increasing function of the amount accumulated at each given time. This amount is gradually depleted when shock is off, until the $a_f^-$ output vanishes. [The accumulation mechanism that is ultimately used is derived in Section IX, and is given by Eq. (44).]

3. *Consensus between Fear and Relief*

We suppose that at most one of the outputs from $a_f^+$ and $a_f^-$ is nonzero at any time. In other words, either fear or relief, but not both, can be "perceived" by the network at a given time. Thus the final state of processing in $a_f^+$ and $a_f^-$, before signals are sent to $S$, is the resultant of a competition between the $a_f^+$ and $a_f^-$ channels due to some form of mutual inhibition.



4. *Existence of a Parallel Accumulation Process in the Fear Channel*

When shock is off for a long time, outputs from both $a_f^+$ and $a_f^-$ to $S$ are zero. Thus the accumulation process at $a_f^-$, driven by its tonic input, is balanced by a process going on at $a_f^+$. The simplest idea is that a parallel process of accumulation–depletion, driven by its own tonic input which equals the $a_f^-$ input, takes place in the $a_f^+$ channel. When shock is on, the shock input summates with the tonic input in the $a_f^+$ channel.

This idea is strengthened by the next few arguments, which elucidate the basic question: What accumulates? Is it potential or is it transmitter? Several facts favor the latter alternative. Other possibilities have been discussed by Grossberg (1972d).

5. *The Rebound Is Slow*

It lasts at least seconds rather than milliseconds. It is a slow process compared to network fluctuation rates of cell potentials in response to input changes. After shock terminates, $a_f^+$ and $a_f^-$ receive no externally driven inputs. Their potentials presumably equalize rapidly. Output from $a_f^-$ nonetheless continues. Thus there exists a process slower than potential change that can bias output from $a_f^+$ and $a_f^-$ in favor of $a_f^-$ after shock terminates.

6. *Both Fear and Relief Are Increasing Functions of Shock Duration and Intensity*

Data on the effect of CS and UCS intensity on the CER and CAR have been reported. Thus both channels contain slowly varying processes which parametrically depend on shock intensity and duration, and which counterbalance each other when shock is off for long intervals.

7. *The Relative Balance of Accumulation Is Changed by Shock*

What causes the $a_f^-$ rebound to shut itself off? Is complete depletion of the accumulated product at $a_f^-$ responsible for this? Suppose that the answer is "Yes." Then the tonic input alone can deplete $a_f^-$. By symmetry, during shock, the shock input plus the tonic input to $a_f^+$ could surely deplete $a_f^+$. This does not occur, since then fear could not be maintained by a prolonged shock. A weaker conclusion is necessary: Shock shifts the relative balance of accumulation in the two channels by depleting the $a_f^+$ channel more than the $a_f^-$ channel.

8. *Signal Size Is a Joint Function of Input Size and Amount Accumulated*

This argument is crucial. During $\alpha_f^-$ rebound, both $\alpha_f^+$ and $\alpha_f^-$ receive equal tonic inputs which ultimately balance the amounts accumulated at $\alpha_f^+$ and $\alpha_f^-$, and thereby nullify $\alpha_f^+$ and $\alpha_f^-$ signals to $S$. Before this happens, $\alpha_f^-$ output exceeds $\alpha_f^+$ output because $\alpha_f^-$ accumulation exceeds $\alpha_f^+$ accumulation. In other words, given a fixed input size (the equal tonic inputs to $\alpha_f^+$ and $\alpha_f^-$), output is an increasing function of accumulation level (in the two channels, $\alpha_f^+$ and $\alpha_f^-$).

When shock is on, increasing shock intensity increases $\alpha_f^+$ output, since it causes an increase in fear. Increasing shock intensity also *decreases* the amount accumulated at $\alpha_f^+$; this is the basis of the rebound at $\alpha_f^-$ when shock is turned off. Thus, output is not a function of accumulation level alone, since then increasing shock intensity would decrease $\alpha_f^+$ output by decreasing the amount accumulated at $\alpha_f^+$. Output size is a joint function of input size and accumulation level.

The terms $B_{ji}z_{ji}$ in (22) shows that output size is the product of spiking frequency and transmitter level. Spiking frequency is an increasing function of potential, which is an increasing function of input size. This leaves transmitter accumulation level as the abstract accumulation level discussed above. This argument commits us to our formalism. We could not proceed further unless: (i) the amount of accumulated transmitter is a decreasing function of input size, and (ii) output size is nonetheless an increasing function of input size. Fortunately, both (i) and (ii) are true in embedding fields, and make a construction of the rebound mechanism possible in this context.

Grossberg (1972d) carries out this construction and rigorously analyzes the resulting mechanisms. These mechanisms include an analogy with adrenergic and cholinergic interactions in series with lateral and ventromedial hypothalamic sites, dependent on phasic sensory input and tonic reticular formation input. Mechanisms emerge for such phenomena as: the lesser rewarding effect of reducing $J$ units of shock to $J/2$ units than of reducing $J/2$ units to 0 unit; a relationship between the rewarding effect of reducing $J$ units of shock to $J/2$ units and the possibility of releasing a conditioned avoidance response in the presence of fearful cues; two kinds of depressed emotional affect—one due to overarousal, which can also be associated with massive associational confusions and poor paying attention, and one due to underarousal, which can also be associated with overreactive fear and relief responses; persistent nonspecific fear which biases interpretation of specific cues, and can "resist" new learning or "repress" old learning; different effects of gradual and abrupt shock on response supression; response generalization from one shock level to another; reduction of pain in the presence of loud noise (analgesia); influences of drugs such as carbachol, atropine, and scopolamine on conditioned emotional and avoidance responses, and on self-stimulation via implanted hypothalamic electrodes; sensory-drive heterarchy that allows changes in situational cues to release responses compatible with any of several nonprepotent drives; feedback inhibition of adrenergic transmitter production; potentiation of adrenergic production by presynaptic spiking, and by postsynaptic spiking via a feedback loop that controls higher-order instrumental conditioning; and learning at cholinergic synapses.



## IX. Possible Chemical Substrates of Network Processes

### A. Refinement of Spatiotemporal Scales

Equations (22) and (23) are derived from psychological postulates and yield an abstract network anatomy whose variables are interpreted as averages over physiological variables. This section illustrates a correspondence procedure whereby spatial and temporal scales in the network are expanded to reveal possible finer processes that are compatible with Eqs. (22) and (23). Further details of this procedure can be found in Grossberg (1969f), along with additional references to relevant data. Here we develop the interpretation of $z_{ji}$ as a transmitter variable, rather than as a measure of postsynaptic membrane sensitivity to fixed amounts of transmitter. Postsynaptic modifications nonetheless arise.

### B. Coupling of $K^+$ to ACh Release

Consider the term $F_{ji} \equiv B_{ji}z_{ji}$ in Eq. (22). The physiological interpretation given in Section III suggests a coupling between outward flux of $K^+$ and of ACh (acetylcholine) from synaptic knobs. Such a coupling has been experimentally reported (Hebb and Krnjevic, 1962; Hutter and Kostial, 1955; Liley, 1956). It is approached as follows: $B_{ji}$ increases with spiking frequency, and each spike is associated with an inward flux of $Na^+$ and an outward flux of $K^+$ (Katz, 1966). Hence an increase in $B_{ji}$ is associated, on a microscopic level, with an increased total outward flux of $K^+$. The term $z_{ji}$ describes the production of excitatory transmitter (say Ach) within $N_{ji}$. $F_{ji} = B_{ji}z_{ji}$ is proportional to the rate of excitatory transmitter released from $N_{ji}$. Hence, increasing the outward flux of $K^+$ increases the rate of transmitter release from $N_{ji}$.

The argument holds even if $B_{ji}$ is a functional of spiking frequency or spike size. This added generality is needed to interpret $B_{ji}$ if $x_j$ becomes large. Since $F_{ji}$ represents rate of transmitter release and $z_{ji}$ is proportional

to total transmitter, $B_{ji}$ must have a finite maximum as $x_j \to \infty$; for example:

$$B_{ji} = \frac{a_{ji}[x_j(t - \tau_{ji}) - \Gamma_{ji}]^+}{b_{ji} + [x_j(t - \tau_{ji}) - \Gamma_{ji}]^+}$$

The mathematical development discussed in Section VI includes this possibility, among many others.

### C. Two Pairs of Antagonistic Ions: ($Na^+$, $K^+$) and ($Ca^{++}$, $Mg^{++}$)

The above interpretation of network variables can be used to suggest the existence of more speculative couplings. These couplings are also compatible with various data, but direct confirmation of their existence seems to be lacking, if only because the necessary experiments would be very hard to perform. First note that, in the presence of inhibitory interactions, Eq. (23) is changed to

$$\dot{z}_{ji} = D_{ji}z_{ji} + E_{ji}[x_i]^+ \tag{41}$$

to prevent negative values of the potential $x_i$ from producing negative amounts of transmitter. How can the product $G_{ji} = E_{ji}[x_i]^+$ in Eq. (41) be interpreted? The term $E_{ji}$ is, along with $B_{ji}$, associated with spiking frequency. The most obvious participants in the spike are the antagonistic ions $Na^+$ and $K^+$. Hence we assume that increases in $E_{ji}$ correspond, on a microscopic level, to (a process in parallel with) an inward flux of $Na^+$ and an outward flux of $K^+$. This process will occur within $N_{ji}$ if we associate $z_{ji}$ with transmitter. The product $G_{ji}$ is then also computed within $z_{ji}$, since it determines the rate of transmitter production, by Eq. (41). The term $[x_i]^+$ in $F_{ji}$ corresponds, however, to a process in $v_i$. Thus there exists a transport of material from $v_i$ to $N_{ji}$, in an amount proportional to $[x_i]^+$, that enables $G_{ji}$ to be computed in $N_{ji}$. What is transported?

Product $G_{ji}$ is a result of two processes. Process $E_{ji}$ is in parallel with a pair of rapidly fluctuating antagonistic ion fluxes. The other process presumably occurs on a similar time scale, and involves chemical species that are known to interact with these ions. Also the two processes in $G_{ji}$ are treated symmetrically: $G_{ji}$ is a product of terms which, in the simplest cases, are both functionals of cell potentials cut off at a threshold (for example, $G_{ji} = \delta_{ji}[x_j(t - \tau_{ji}) - \Gamma_{ji}]^+[x_i]^+$), and it is known in the case of spike production that the threshold is produced by interaction between the pair $Na^+$ and $K^+$ of antagonistic ions. The simplest assumption is thus that $[x_i]^+$ also represents a process (in parallel with) a pair of antagonistic ion fluxes. This assumption turns out to be compatible with various data. In



the following discussion of these data, the phrase *in parallel with* a pair of antagonistic ions is critical. Indeed, our macroscopic theory can do little more than suggest the symmetries of microscopic interactions, so that the pairs being sought, in principle, need not be composed of ions at all (cf. amino acids). The formal structure of the argument seems to hold, no matter how we interpret these chemicals.

The pair of ions associated with $[x_i]^+$ cannot be ($Na^+$, $K^+$). If it were, increases in $[x_i]^+$ would correspond to an influx of $Na^+$ and an outflux of $K^+$ at $v_i$. The process $z_{ji}$ is, however, influenced only by those aspects of these fluxes that affect $N_{ji}$. These effects are a decrease in $Na^+$ and an increase in $K^+$. Process $E_{ji}$ involves the same ions and has the opposite effect when $E_{ji}$ increases. How then do these processes affect $z_{ji}$ in Eq. (41) only through their product? In particular, by Eq. (41), $z_{ji}$ cannot grow in response to even an enormous $E_{ji}$ value if $[x_i]^+ = 0$, even though $E_{ji}$ provides within $N_{ji}$ all the effects that $[x_i]^+$ can trigger. Thus, if $[x_i]^+$ is in parallel with a pair of antagonistic ions, it must be a pair other than ($Na^+$, $K^+$).

In many biochemical processes, the divalent ions $Ca^{++}$ and $Mg^{++}$ powerfully interact with $Na^{++}$ and $K^+$, and the pair ($Ca^{++}$, $Mg^{++}$) is mutually antagonistic (Dixon and Webb, 1958). We take this to be the pair being sought. In many reactions, $Na^+$ and $Ca^{++}$ act synergistically (Fruton and Simmonds, 1958). We therefore consider this possibility in the present context: Let an increase in $[x_i]^+$ correspond microscopically to an increase in $Ca^{++}$ and a decrease in $Mg^{++}$.

### D. Binding of $Na^+$ and $Ca^{++}$ as Synergistic Cofactors on Transmitter Production Sites

Now term $G_{ji}$ says that transmitter production sites are activated at a rate proportional to the product of (processes in parallel with) $Na^+$ and $Ca^{++}$ concentrations. In particular, we expect joint inward $Na^+$ and $Ca^{++}$ fluxes to be created by membrane excitation and to thereby stimulate transmitter production, whereas $K^+$ and $Mg^{++}$ antagonize $Na^+$ and $Ca^{++}$, respectively, in this role. Analogous fluxes have been experimentally reported (del Castillo and Engbaek, 1954; Harvey and MacIntosh, 1940; Hodgkin and Keynes, 1957). Just as inward fluxes of $Na^+$ and $Ca^{++}$ presumably facilitate transmitter production, it is natural to expect that such fluxes facilitate transmitter release, so as not to cancel out one process with another. If ACh is the transmitter, then reducing $Ca^{++}$ concentration around $N_{ji}$ would reduce ACh release, other things being equal. If $Mg^{++}$ is acting as a $Ca^{++}$ antagonist, then $Mg^{++}$ should antagonize $Ca^{++}$ in controlling the amount of ACh release. Compatible experimental reports are

found in del Castillo and Engbaek (1954), del Castillo and Katz (1954) Hubbard (1961), Jenkinson (1957), and Liley (1956).

### E. A Hierarchy of Intracellular Ionic Binding Strengths

By Eq. (41), new transmitter production sites are activated only when $G_{ji} > 0$—that is, only if supraequilibrium amounts of (quantities in parallel with) $Na^+$ and $Ca^{++}$ simultaneously reach these sites. When equilibrium is restored, $G_{ji} = 0$. The rate of change in $z_{ji}$ due to $G_{ji}$ is also zero during equilibrium; the sites remember how much transmitter to produce.

The following basic questions hereby arise. How can high concentrations of $Na^+$ and $Ca^{++}$ jointly activate a process that maintains its activity even after the concentrations of these ions are reduced at equilibrium? Otherwise expressed, what keeps $z_{ji}$ at the high values needed to produce a memory of past events even when the sources of these high values are removed as equilibrium is restored? In particular, why doesn't the high intra-end-bulb $K^+$ concentration at equilibrium reversibly inhibit $z_{ji}$ growth, just as $Na^+$ and $Ca^{++}$ excited $z_{ji}$ growth at nonequilibrium?

Since $z_{ji}$ does maintain the high values acquired during nonequilibrium, and joint coupling of $Na^+$ and $Ca^{++}$ causes these values, we are led into the following conclusion: The $Na^+$ and $Ca^{++}$ ions which activated the transmitter production sites are not removed from the end bulb when equilibrium is restored; a fraction of the free $Na^+$ and $Ca^{++}$ ions which enter the end bulb during excitation is bound on intra-end-bulb transmitter production sites, and this binding is so strong that it cannot be displaced by the return of a high intra-end-bulb $K^+$ concentration as equilibrium is restored. In particular, the intracellular $K^+$ ions are not so strongly bound. We are hereby led to expect that most of the intracellular $K^+$ exists in unbound form, whereas higher proportions of intracellular $Na^+$ and/or $Ca^{++}$ exist in bound form. These expectations have been experimentally reported (Brink, 1954; Ussing, 1960).

### F. The Control of Cellular Production Rates by Ions: Strength of Binding versus Ion Availability

The above remarks suggest a qualitative answer to a special case of the following general question: How do cells "know" how much of a given quantity to produce in response to external environmental demands?

Our point of departure is the hypothesis that ions such as $Na^+$ and $Ca^{++}$, which presumably activate intra-end-bulb sites (or enzymes) with considerable vigor, are kept substantially out of the end bulb during equilibrium. Only in nonequilibrium periods such that $x_j(t - \tau_{ji}) > \Gamma_{ji}$ and $x_i(t) > 0$ can these ions penetrate the membrane en masse to initiate higher levels of intra-end-bulb transmitter production. Since equilibrium time intervals can, in principle, exceed nonequilibrium time intervals by a very large numerical factor, the ions $Na^+$ and $Ca^{++}$, which bind most strongly, are available least frequently within the end bulb. In other words, the process of synergistic ($Na^+$, $Ca^{++}$) binding, having a limited opportunity to occur, is made effective by guaranteeing that, whenever the opportunity does occur, the process takes place vigorously and its effects are long-lasting (cf. Brink, 1954; Quastel, 1962).

These facts suggest the following general heuristic scheme for integrating equilibrium and nonequilibrium phases in the life of a cell, which subsumes the problem of rendering the cell responsive to fluctuations in its external environment. The argument can be broken into three main steps.



#### 1. *Coexistence of Equilibrium and Evolution*

An equilibrium phase of a cell can, in principle, be characterized by particular values of prescribed cellular parameters. For example, the equilibrium of a nerve cell can be characterized by the membrane concentrations of such parameters as $Na^+$ and $K^+$. Suppose that a cell exists whose equilibrium is characterized by particular values of all its parameters. Such a cell "forgets" all nonequilibrium values of its parameters when it returns to equilibrium. In particular, the equilibrium of such a cell cannot coexist with long-term responses of the cell to brief changes in its external environment. For convenience, we henceforth call such long-term responses evolutionary trends.

Certainly not all cells are of this type. Brains can learn! Henceforth we concern ourselves only with cells whose equilibrium phase can coexist with an evolutionary trend. We denote such a cell by $C$. By definition, the equilibrium phase of $C$ does not require a specification of values for all cellular parameters. It suffices to specify the values of a fraction of these parameters. We denote these equilibrium parameters collectively by $E$. A particular evolutionary trend in $C$ requires the specification of values for parameters which we denote by $N$. Since the parameters $N$ control an evolutionary trend, they need not always take on the same values when the parameters $E$ take on equilibrium values.

#### 2. *The External Environment Perturbs the Equilibrium Parameters*

The external environment communicates its demands upon $C$ by changing the values of parameters at $C$'s periphery, or membrane. These parameters are, however, often the parameters $E$, since equilibrium is a state of $C$ which is characterized by a particular choice of external environment. For

example, a nerve cell returns to equilibrium when all excitatory and inhibitory inputs are zero. We conclude that the external environment often induces an evolutionary trend in the parameters $N$ by perturbing the parameters $E$. The parameters $E$ therefore *faithfully* communicate to the parameters $N$ the demands of the external environment. We are hereby led to the following basic but merely ostensible paradox: If the parameters $E$ faithfully communicate to the parameters $N$ the external environmental demands that signal an evolutionary trend, then why don't the parameters $E$ also faithfully communicate to the parameters $N$ the external environmental demands that signal equilibrium, and thereby eradicate the evolutionary trend in $N$ whenever equilibrium is restored?

### 3. *The Equilibrium Values Compete with the Nonequilibrium Values of the Equilibrium Parameters*

Given the natural assumption that the parameters $E$ pass on faithfully to $N$ *all* states of the external environment, the following resolution of this paradox seems natural: The equilibrium values of $E$ do not eradicate the evolutionary trend in $N$ because they cannot dislocate from $N$ the nonequilibrium values of $E$ that induced the trend. In the case that the parameters $E$ are realized by ions, this means that a hierarchy of ionic binding strengths exists at the intracellular sites (or enzymes) which alter intracellular demands. The ions that are most available during equilibrium are bound least strongly to these sites. The ions introduced at these sites by the extracellular demands are strongly bound as synergistic cofactors to these sites, and thereby activate them. Proceeding in the reverse direction, suppose that the ions that bind most strongly to these sites are *not* substantially kept out of the cell during equilibrium, and are allowed freely to bind with these sites and thereby to activate them. Then essentially all sites will always be occupied, and the production rate at these sites will always be in a state of equilibrium, albeit a very active equilibrium. The evolutionary trend is hereby destroyed.

## G. The Mitochondrion and Ion Translocation

Given the hypothesis that $Na^+$ and $Ca^{++}$ are synergistic cofactors in the activation of sites that contribute to transmitter production, it is desirable to find candidates for these sites. A cellular system which has a strong affinity for $Na^+$ and $Ca^{++}$ is the mitochondrion, whose importance as the "power plant" of aerobic cells is well known. For example, Lehninger (1965, pp. 169–171) reports a striking increase during respiration in both the relative uptake of $Na^+$ over $K^+$ and of $Ca^{++}$ over $Mg^{++}$. To the extent that this fact is an example of our theoretical expectations, then ion translocation in neural mitochondria can be interpreted as a means for setting mitochondrial reaction rates at a level commensurate with the intensity and duration of a positively polarized nonequilibrium excitation phase. These rates endure long into the equilibrium phase.



## H. Provision of ATP for Synaptic Vesicles by Mitochondria

Suppose that ion translocation in the mitochondrion is indeed an example of the synergism between $Na^+$ and $Ca^{++}$ that contributes to transmitter production. Then mitochondria should be found clustered near regions of high transmitter density. Histological evidence suggests that transmitter is stored in synaptic vesicles, and that mitochondria can be found clustered near these vesicles (de Robertis, 1964, p. 32, and micrographs throughout the book). Perhaps the activated mitochondria supply the ATP needed to produce acetyl coenzyme A, which in turn presumably reacts with choline under the aegis of the enzyme choline acetylase to produce acetylcholine (Eccles, 1964; Fruton and Simmonds, 1958).

## I. Contiguity of Synaptic Vesicles and the Synaptic Cleft

The histological investigations (Eccles, 1964; de Robertis, 1964) which have revealed the existence of synaptic vesicles also show that these vesicles are often clustered most densely along the end-bulb surface which faces the synaptic cleft. This location is well chosen for a vesicle whose supposed role is to expeditiously release transmitter into the synaptic cleft to excite the postsynaptic membrane. Yet how does the vesicle know how to choose this useful location? Such knowledge will seem mysterious in any theory that holds that transmitter production depends only on the past excitation history of the presynaptic nerve which contains the transmitter, since the excitation of just this nerve does not provide information concerning the location of the synaptic cleft relative to the end-bulb membrane. Such a theory predicts that transmitter vesicles will be found uniformly throughout the end bulb, or closer to the presynaptic source of excitation than to the synaptic cleft, or at best with uniform density along all end-bulb surfaces.

The preferential location of synaptic vesicles near the synaptic cleft is qualitatively easily understood in a theory in which transmitter production depends on both presynaptic and postsynaptic influences. Presumably the postsynaptic influence is carried over the synaptic cleft to the presynaptic end bulb, so that the region most likely to have all the ingredients needed for transmitter production lies nearest to the synaptic cleft. The postsynaptic ionic influence does not spread evenly throughout the presynaptic end bulb because the $Ca^{++}$ influence near the synaptic cleft is presumably

bound within the end bulb as soon as it reaches an appropriate site, and the amount of $Ca^{++}$ entering the cell cannot be so large as to uniformly saturate all sites within the end bulb, or else the desired evolutionary trend will be destroyed. Indeed, one way to turn a knob capable of learning into a knob incapable of learning is to open the tight junctions for the transport from $v_i$ to $N_{ji}$, and thereby bathe the presynaptic end bulb in an ionic atmosphere that is not driven by postsynaptic events.

## J. Binding of $Mg^{++}$ by RNA in the Cell Body

The $Ca^{++}$ needed for synergistic binding of $Na^{++}$ and $Ca^{++}$ in $N_{ji}$ are released into the synaptic cleft facing $N_{ji}$ when the postsynaptic cell, $v_i$, is sufficiently excited. Otherwise, much of the $Ca^{++}$ in the synaptic cleft is presumably reabsorbed into $v_i$.

This argument fails completely if $N_{ji}$ can provide as much $Ca^{++}$ as $v_i$, given a fixed level of excitation, since then $E_{ji}$ would stand for essentially the same ionic fluxes as $[x_i]^+$, and the coupling $F_{ji}$ could not be realized. Since $v_i$ presumably can supply more $Ca^{++}$ than $N_{ji}$, we must find a rationale for this fact.

Given that $[x_i]^+$ represents an antagonism between $Ca^{++}$ and $Mg^{++}$, the fact that $Ca^{++}$ is released when $v_i$ is excited means that $Mg^{++}$ is needed by $v_i$ during excitation. A structure therefore exists within $v_i$—which is not found in $N_{ji}$—which selectively binds $Mg^{++}$ ions when $v_i$ is active and whose binding with $Mg^{++}$ is preferred to (or antagonized by) binding with $Ca^{++}$. This argument does not mean that no $Ca^{++}$ is provided by $N_{ji}$, but only that more $Ca^{++}$ is provided by $v_i$. In a similar fashion, the fact that presynaptic excitation at $N_{ji}$ induces coupled $Na^+$ and $K^+$ fluxes does not imply that such fluxes are absent from postsynaptic excitation at $v_i$.

The cell body $v_i$ certainly has at least one prominent structure which the end bulb $N_{ji}$ does not have—namely, the cell nucleus. If this is the structure being sought, then the cell nucleus, or processes sustained by the nucleus, ought to selectively bind $Mg^{++}$ ions when the cell body is activated. Among the most plentiful cell body constituents of this type are the RNA's. It is also known that RNA activity depends sensitively on $Mg^{++}$ concentration (Boedtker, 1960; Spirin, 1964; Watson, 1965).

## K. Interaction of Neural Excitation and RNA

Suppose indeed, that the RNA's are among the structures that we are seeking to bind $Mg^{++}$. Then learning will be associated with systematic variations in the RNA's. Such variations have been reported experimentally (Hamberger and Hydén, 1963; Hydén, 1962; Koenig, 1964).



Once experiments were produced demonstrating variations in RNA activity in learning situations, it was proposed that individual RNA strands coded the content of the learning in some fashion, and that one could, in principle, recover the content of whole segments of learned experiences in such a strand if one but had the key for decoding its structure. This view seems unnecessary from the present perspective. The RNA's seem to be needed merely to keep the cell at production levels appropriate to the metabolic drains placed on the cell by the levels of excitation imposed from the external environment. Indeed, if a spatial pattern is the unit of long-term memory, then an individual cell does not have enough information to know what is being learned. Nonetheless, the cross-correlational processes presumed to occur at the cellular level do provide enough information for the cell to discriminate whether a learning type of process is occurring or not.

## L. Transport down the Axon

The hypothesis that $Mg^{++}$ is bound to nucleus-related processes is further strengthened by the following observation.

Figure 29 schematically represents a presynaptic nerve cell with nucleus $N_j$ whose excitatory end bulb, $N_{ji}$, impinges upon the postsynaptic nerve cell $v_i$ with nucleus $N_i$. Suppose that $N_i$ selectively binds $Mg^{++}$ in order to free $Ca^{++}$ for binding within $N_{ji}$ when both $N_{ji}$ and $v_i$ are vigorously excited. If $v_i$ and $v_j$ are of the same cell type, then $Mg^{++}$ will also be selectively bound by $N_j$ when $v_j$ is vigorously excited. Since $v_j$ is connected to $N_{ji}$ by the axon $e_{ji}$, we must prevent most of the molecules that bind $Mg^{++}$ within $v_j$ from flowing down the axon to $N_{ji}$, or else $N_{ji}$ will have too many $Mg^{++}$-binding molecules. Thus at least part of the $Mg^{++}$ must be bound within $v_j$ to structures that are so large or so well cemented within $v_j$ that they are never carried down the axon to the end bulb. Macromolecules within $N_j$, such as the RNA's, are plausible candidates for such a role.

On the other hand, whenever $v_j$ is excited to suprathreshold values, then the axon $e_{ji}$ and the end bulb $N_{ji}$ are also excited. The axon and the end bulb must be able to recover from this excitation. The postulated mecha-

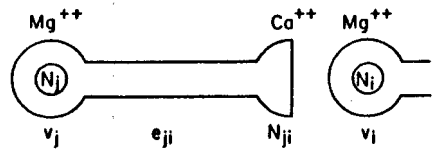


Fig. 29. Interacting chemical dipoles.

nism of recovery is activation by (processes in parallel to) $Mg^{++}$ of the RNA's during excitation, leading to higher rates of protein synthesis, etc. However, the RNA's are substantially localized within the cell body $v_j$. Thus the molecules produced by RNA activation, after being produced in $v_j$, must be able to travel down the axon to the end bulb where they will be needed to guarantee recovery from excitation. These molecules therefore might well be lighter than the more immobile RNA's, and they might well be bound to less $Mg^{++}$ than is bound to the activated RNA's. A transport of material from the cell body along the axon to the end bulb exists (Friede, 1959; Ochs and Burger, 1958; Waelsch and Lajtha, 1960; Weiss and Hiscoe, 1946). Various details concerning this formal transport process are considered in Grossberg (1969f).

## M. Why Aren't Nerve Cells Spherical? An Intimate Bond between Neural Geometry and Neural Dynamics

It is practically a truism that the simplest geometrical objects are as homogeneous and as symmetric as possible. Thus, among the simplest three-dimensional and finite bodies are the spheres, and it is useful to think of the complexity of a three-dimensional and finite body—such as a nerve cell—in terms of its deviations from sphericity. It is also natural to suppose that a finite system in nature will assume the simplest shape that is compatible with its function. We are then readily led to ask: What features of a nerve cell's functions require that it be nonspherical?

Our speculations suggest that the role of nerve cells as mechanisms of learning requires their nonspherical shape. We link a nerve cell's ability to learn with the existence of different chemical affinities at two opposite poles of the nerve cell—namely, near the cell body and end bulbs; that is, the nerve cell is presumed to be a chemical dipole. Were the nerve cell spherical in all ways, in particular with a spherical nucleus in its center, then symmetry arguments would imply that this chemical dipole could not be realized.

Given the need for a dipole shape, the nerve cell is then confronted with the formidable problem of carrying signals from its external environment reliably from one end of the dipole to the other. This problem is formidable because the functional biases caused by the dipole might well be expected to distort the signal as it travels along the cell. The cell has solved this problem in an ingenious, but intuitively simple, way. The signals from the external environment, which first perturb the boundary, or membrane, of the cell, are transmitted reliably from one end of the dipole to the other along this boundary, whereas the chemical dipole properties of the cell are



safely ensconced well within the cellular interior, where they can secondarily benefit from external environmental news without profoundly distorting the transmission of this news along the entire cell. Note by Section VI that this constraint aiming at unbiased signal transfer on the individual cell level also seems to create unbiased learning on the network level.

## N. Two Main Steps in Transmitter Production

We now show that the single variable, $z_{ji}$, represents two processes taking place at two different rates. These two processes are the following ones.

1. *Slowly Varying Transmitter Production Rates.* Long-term memories of past network events are contained in the $z_{ji}$ functions. These functions therefore vary more slowly than the events themselves. In particular, if $G_{ji} = \delta_{ji}[x_j(t - \tau_{ji}) - \Gamma_{ji}]^+[x_i]^+$, then $z_{ji}$ varies more slowly than $x_j$ and $x_i$.

2. *Rapidly Varying Transmitter Release.* Suppose that

$$F_{ji} = \beta_{ji}[x_j(t - \tau_{ji}) - \Gamma_{ji}]^+ z_{ji}$$

for definiteness. At suprathreshold values, $F_{ji}$ is a linear function of $x_j(t - \tau_{ji})$, and is therefore rapidly varying compared to $z_{ji}$.

The physical interpretation of $F_{ji}$ leads to the two processes represented by $z_{ji}$. The function $F_{ji}$ is proportional to the rate of transmitter release from $N_{ji}$, and $z_{ji}$ is the total amount of transmitter in $N_{ji}$. Why, then, doesn't the law (41) for $z_{ji}$ read as follows?

$$\dot{z}_{ji} = D_{ji} z_{ji} + E_{ji}[x_i]^+ - F_{ji}$$

That is, shouldn't the total amount of transmitter in $N_{ji}$ be reduced by the amount of transmitter that is released from $N_{ji}$? On formal grounds, this subtraction procedure is inadmissible; then $z_{ji}$ would be drastically reduced in size whenever the presynaptic spiking frequency became large, and the "memory" represented by $z_{ji}$ would quickly be destroyed. A conceptual distinction clearly must be made between $z_{ji}$ as "memory" and $z_{ji}$ as "releasable transmitter."

Two problems must simultaneously be resolved:

1. Distinguish $z_{ji}$, the rate of transmitter production, from $y_{ji}$, the amount of transmitter.

2. Show how $z_{ji}$ can represent both variables in the macroscopic psychological picture; that is, show that, on the average,

$$y_{ji}(t) \cong \epsilon_{ji} z_{ji}(t) \tag{42}$$

where $\epsilon_{ji}$ is a positive constant. The relation (42) can hold at all times $t$

only if the transmitter lost from $N_{ji}$ is instantly replenished until it reaches a level proportional to $z_{ji}$. This happens only if the rate of replenishment is infinite. This rate only seems to be infinite on the time scale of psychological events because replenishment is a rapid process on this scale. We now refine this time scale by assuming that the replenishment rate is finite, but otherwise do not change our equations. As usual, we seek the most linear way to express our intuitive ideas, while realizing that there exist variations on the linear theme.

By (42), at times when no transmitter is released from $N_{ji}$, $y_{ji}$ seeks a level proportional to $z_{ji}$. Hence, at these times,

$$\dot{y}_{ji}(t) = \zeta_{ji}[\epsilon_{ji}z_{ji}(t) - y_{ji}(t)] \tag{43}$$

where $0 \leq y_{ji} < \epsilon_{ji}z_{ji}$ and $\zeta_{ji} > 0$.

If transmitter is released from $N_{ji}$ at a rate $H_{ji}$, then $\dot{y}_{ji}$ in (43) is reduced by this amount. Thus, in general,

$$\dot{y}_{ji} = \zeta_{ji}(\epsilon_{ji}z_{ji} - y_{ji}) - H_{ji} \tag{44}$$

The term $H_{ji}$ cannot be identified with $F_{ji}$ because $z_{ji}$ no longer represents the amount of transmitter. Guided by the definition of $F_{ji} = B_{ji}z_{ji}$ and (42), we let

$$H_{ji} = \eta_{ji}B_{ji}y_{ji}, \qquad \eta_{ji} = \epsilon_{ji}^{-1}$$

Thus Eq. (44) merely replaces a process with an infinite reaction rate by a qualitatively identical process with a finite reaction rate. In the special case that the transmitter is ACh, a possible interpretation of these variables is

$z_{ji}$ = total amount of available ACh in $N_{ji}$

and

$y_{ji}$ = total activity of the choline acetylase (ChAc) system which controls ACh production (Fruton and Simmonds, 1958; Krnjević, 1965; Sumner and Somers, 1953).

## O. Feedback Inhibition

Equation (44) has the following chemical interpretation. Write (44) as a sum of three terms:

$$\dot{y}_{ji} = U_{ji} + V_{ji} + W_{ji} \tag{45}$$

where

$$U_{ji} = \zeta_{ji}\epsilon_{ji}z_{ji} \tag{46}$$

$$V_{ji} = -\zeta_{ji}y_{ji} \tag{47}$$



and

$$W_{ji} = -\eta_{ji}B_{ji}y_{ji} \tag{48}$$

The term (46) says that transmitter production rate is proportional to the number of active transmitter producing sites. Term (47) says that transmitter production rate is diminished by an amount proportional to the amount of transmitter—that is, by a feedback inhibition by the transmitter end product of a prior stage of transmitter production. This inhibition cannot influence those transmitter-producing sites which are activated by extracellular demands without destroying the cellular memory of these demands. Hence a later, or intermediate, stage of transmitter production is inhibited (cf. Fruton and Simmonds, 1958; Wyatt, 1964). Term (48) implies that feedback inhibition is reduced by release of transmitter from $N_{ji}$.

It is interesting to compute the response of Eqs. (45)–(48) to a spiking frequency that is switched to a steady-state level $B > 0$ at time $t = 0$ after a long internal of zero spiking. One finds three major effects:

1. A transient overshoot in transmitter release.
2. A progressive *decrease* in the asymptotic total available transmitter $y_{ji}(\infty)$ as a function of increasing $B$.
3. A progressive *increase* in the asymptotic rate of transmitter release $H_{ji}(\infty)$ as a function of increasing $B$.

Thus the total amount of transmitter in $N_{ji}$ and the amount of transmitter that is released from $N_{ji}$ do not covary as function of E (cf.). This fact makes it possible to construct the rebound mechanism using transmitter accumulation–depletion in Section VIII.

## P. Transmitter Mobilization

The process of refining scales can be continued indefinitely. For example, if process $z_{ji}$ takes place within $N_{ji}$ but contributes transmitter for release from the $N_{ji}$ membrane, then transmitter will be transported to the membrane. Various models for this can be contemplated. The simplest again rely on linearity wherever possible. For example, let

$y_{ji}$ = the total amount of transmitter in $N_{ji}$

and

$w_{ji}$ = the total amount of transmitter in $N_{ji}$ at the membrane facing the synaptic cleft

The rate of transmitter release in this case involves $w_{ji}$, not $y_{ji}$, and is derived much as $H_{ji}$ was derived from $F_{ji}$. Thus we find, using linearity

wherever possible, that

$$\dot{y}_{ji} = \zeta_{ji}(\epsilon_{ji}z_{ji} - y_{ji}) - \eta_{ji}E_{ji}w_{ji} \tag{49}$$

and

$$\dot{w}_{ji} = \theta_{ji}(y_{ji} - w_{ji}) - \eta_{ji}E_{ji}w_{ji} - \kappa_{ji}[w_{ji} - \lambda_{ji}]^+ \tag{50}$$

Equation (50) can be understood by writing it as

$$\dot{w}_{ji} = U^*_{ji} + V^*_{ji} + W^*_{ji}$$

with

$$U^*_{ji} = \theta_{ji}(y_{ji} - w_{ji}) \tag{51}$$

$$V^*_{ji} = -\eta_{ji}E_{ji}w_{ji} \tag{52}$$

and

$$W^*_{ji} = -\kappa_{ji}[w_{ji} - \lambda_{ji}]^+ \tag{53}$$

Term (51) says that transmitter is mobilized at a rate proportional to the amount $(y_{ji} - w_{ji})$ of unmobilized transmitter. Term (52) gives the rate of releasing mobilized transmitter from $N_{ji}$. Term (53) says that mobilized transmitter can become spontaneously demobilized until only an amount $\lambda_{ji}$ of transmitter is still mobilized.

Equation (50) has some interesting properties. If we study its transient response, then the slowly varying $z_{ji}$ remains approximately constant. Suppose also that $\lambda_{ji} = 0$ and $\zeta_{ji} = \kappa_{ji}$. Then the equation can be explicitly integrated. Properties 1 to 3 of the previous section hold, and in addition the amount of mobilized transmitter is constant through time. The last property is not generally true if $\zeta_{ji} \neq \kappa_{ji}$.

In summary, this paper illustrates a procedure whereby the physiological equations themselves, and the network anatomy, can be successively refined to accommodate increasingly subtle psychological postulates. At each level of analysis, one finds phenomena that caution against arguing from local to global, or from linear to nonlinear, network properties.

## REFERENCES


Adams, J. A. (1967). "Human Memory." McGraw-Hill, New York.

Atkinson, R. C., and Estes, W. K. (1963). *In* "Handbook of Mathematical Psychology" (R. D. Luce, R. R. Bush, and E. Galanter, eds.), Vol. II, p. 121. Wiley, New York.

Atkinson, R. C., and Shiffrin, R. M. (1968). *In* "The Psychology of Learning and Motivation" (K. W. Spence and J. T. Spence, eds.), Vol. 2, p. 98. Academic Press, New York.

Boedtker, H. (1960). *J. Mol. Biol.* **2**, 171.

Brink, F. (1954). *Pharmacol. Rev.* **6**, 243, 276 and 284.

Cox, V. C., Kakolewski, J. W., and Valenstein, E. S. (1969). *J. Comp. Physiol. Psychol.* **68**, 337.

Dally, P. (1967). "Chemotherapy of Psychiatric Disorders." Plenum, New York.

del Castillo, J., and Engbaek, L. (1954). *J. Physiol. (London)* **124**, 370.





del Castillo, J., and Katz, B. (1954). *J. Physiol. (London)* **124**, 560.

Denny, M. R. (1971). *In* "Aversive Conditioning and Learning" (F. R. Brush, ed.), p. 000. Academic Press, New York.

de Robertis, E. D. P. (1964). "Histophysiology of Synapses and Neurosecretion." Macmillan, New York.

Dethier, V. G. (1968). "Physiology of Insect Senses." Methuen, London.

Dixon, M., and Webb, E. C., eds. (1958). "Enzymes," 1st ed. Academic Press, New York.

Eccles, J. C. (1964). "The Physiology of Synapses." Springer-Verlag, Berlin and New York.

Estes, W. K. (1969). *In* "Punishment and Aversive Behavior" (B. A. Campbell and R. M. Church, eds.), p. 57. Appleton, New York.

Friede, R. L. (1959). *Exp. Neurol.* **1**, 441.

Fruton, J. S., and Simmonds, S. (1958). "General Biochemistry." Wiley, New York.

Grastyán, E. (1968). *In* "Biological Foundations of Emotion" (E. Gellhorn, ed.), p. 000. Scott, Foresman, Glenview, Illinois.

Grossberg, S. (1968a). *J. Math. Anal. Appl.* **21**, 643.

Grossberg, S. (1968b). *J. Math. Anal. Appl.* **22**, 490.

Grossberg, S. (1969a). *J. Math. Psychol.* **6**, 209.

Grossberg, S. (1969b). *J. Math. Mech.* **19**, 53.

Grossberg, S. (1969c). *J. Differential Equations* **5**, 531.

Grossberg, S. (1969d). *J. Statis. Phys.* **1**, 319.

Grossberg, S. (1969e). *Math. Biosci.* **4**, 201.

Grossberg, S. (1969f). *J. Theor. Biol.* **22**, 325.

Grossberg, S. (1970a). *J. Theor. Biol.* **27**, 291.

Grossberg, S. (1970b). *Stud. Appl. Math.* **49**, 135.

Grossberg, S. (1971a). *J. Cybernet.* **1**, 28.

Grossberg, S. (1971b). *J. Theor. Biol.* **33**, 225.

Grossberg, S. (1971c). *Proc. Nat. Acad. Sci. U.S.* **68**, 828.

Grossberg, S. (1972a). *Kybernetik* **10**, 49.

Grossberg, S. (1972b). *In* "Delay and Functional Differential Equations and Their Applications" (K. Schmitt, ed.), p. 121.

Grossberg, S. (1972c). *Math. Biosci.* **15**, 39.

Grossberg, S. (1972d). *Math. Biosci.* **15**, 253.

Grossberg, S. (1973). *Studies Appl. Math.* **52**, 213.

Grossberg, S. (1975). *Int. Rev. Neurobiol.* **18**. (To be published).

Grossberg, S., and Pepe, J. (1971). *J. Statist. Phys.* **3**, 95.

Hamberger, A., and Hydén, H. (1963). *J. Cell Biol.* **16**, 521.

Harvey, A. M., and MacIntosh, F. C. (1940). *J. Physiol. (London)* **97**, 408.

Hebb, C. O., and Krnjević, K. (1962). *In* "Neurochemistry" (K. A. C. Elliott, I. H. Page, and J. H. Quastel, eds.), p. 452. Thomas, Springfield, Illinois.

Hebb, D. O. (1955). *Physiol. Rev.* **62**, 243.

Hodgkin, A. L., and Keynes, R. D. (1957). *J. Physiol. (London)* **138**, 253.

Hubbard, J. I. (1961). *J. Physiol. (London)* **159**, 507.

Hubel, D., and Wiesel, T. N. (1968). *In* "Physiological and Biochemical Aspects of Nervous Integration" (F. O. Carlson, ed.), p. 153. Prentice-Hall, Englewood Cliffs, New Jersey.

Hutter, O. F., and Kostial, K. (1955). *J. Physiol. London* **129**, 159.

Hydén, H. (1962). *In* "Neurochemistry" (K. A. C. Elliott, I. H. Page, and J. H. Quastel, eds.), p. 331. Thomas, Springfield, Illinois.

Jenkinson, D. H. (1957). *J. Physiol. (London)* **138**, 434.
Jensen, A. R. (1962). *J. Psychol.* **53**, 127.
Julesz, B. (1964). *Science* **145**, 356.
Katz, B. (1966). "Nerve, Muscle and Synapse." McGraw-Hill, New York.
Kennedy, D. (1968). *In* "Physiological and Biochemical Aspects of Nervous Integration" (F. O. Carlson, ed.), p. 285. Prentice-Hall, Englewood Cliffs, New Jersey.
Koenig, H. (1964). *In* "Morphological and Biochemical Correlates of Neural Activity" (M. M. Cohen and R. S. Snider, eds.), p. 39. Harper, New York.
Kornetaky, C., and Eliasson, M. (1969). *Science* **165**, 1273.
Krnjević, J. (1965). *In* "Pharmacology of Cholinergic and Adrenergic Transmission" (G. B. Koelle, W. W. Douglass, and A. Carlson, eds.), p. 21. Macmillan, New York.
Lashley, K. S. (1951). *In* "Cerebral Mechanisms in Behavior: The Hixon Symposium" (L. P. Jeffress, ed.), p. 112. Wiley, New York.
Lehninger, A. L. (1965). "The Mitochondrion." Benjamin, New York.
Lehrman, D. S. (1965). *In* "Sex and Behavior" (F. A. Beach, ed.), p. 355. Wiley, New York.
Liley, A. W. (1956). *J. Physiol. (London)* **134**, 427.
Livingston, R. B. (1967). *In* "The Neurosciences" (F. O. Schmitt *et al.*, eds.), Vol. 2, p. 91. MIT Press, Cambridge, Massachusetts.
Logan, F. A. (1969). *In* "Punishment and Aversive Behavior" (B. A. Campbell and R. M. Church, eds.), p. 43. Appleton, New York.
Maher, B. A. (1968). *Psychol. Today* **2**, 30.
Masterson, F. A., (1970). *J. Comp. Physiol. Psychol.* **72**, 471.
Miller, G. A. (1956). *Psychol. Rev.* **63**, 81.
Miller, N. E. (1963). *In* "Nebraska Symposium on Motivation" (M. R. Jones, ed.), p. 65. Univ. of Nebraska Press, Lincoln.
Ochs, S., and Burger, E. (1958). *Amer. J. Physiol.* **194**, 499.
Olds, J. (1955). *In* "Nebraska Symposium on Motivation" (M. R. Jones, ed.) Univ. of Nebraska Press, Lincoln.
Orkand, R. K., Nicholls, J. G., and Kuffler, S. W. (1966). *J. Neurophysiol.* **29**, 788.
Osgood, C. E. (1953). "Method and Theory in Experimental Psychology." Oxford Univ. Press, London and New York.
Penfield, W. (1958). "The Excitable Cortex in Conscious Man." Liverpool Univ. Press, Liverpool.
Phillips, M. I., and Bradley, P. B. (1970). *Science* **168**, 1122.
Quastel, J. H. (1962). *In* "Neurochemistry" (K. A. C. Elliott, I. H. Page, and J. H. Quastel, eds.), p. 226. Thomas, Springfield, Illinois.
Ratliff, F. (1965). "Mach Bands: Quantitative Studies of Neural Networks in the Retina." Holden-Day, San Francisco, California.
Reynierse, J. H., and Rizley, R. C. (1970). *J. Comp. Physiol. Psychol.* **72**, 223.
Ruch, T. C., Patton, H. D., Woodbury, J. W., and Towe, A. L. (1961). "Neurophysiology." Saunders, Philadelphia, Pennsylvania.
Spirin, A. S. (1964). "Macromolecular Structure of Ribonucleic Acids," Part II, Chapter 4. Van Nostrand-Reinhold, Princeton, New Jersey.
Stein, P. S. G. (1971). *J: Neurophys.* **34**, 310.
Sumner, J. B., and Somers, G. F. (1953). "Chemistry and Methods of Enzymes," 3rd rev. ed., p. 351. Academic Press, New York.
Thomas, G. J., Hostetter, G., and Barker, D. J. (1968). *Progr. Physiol. Psychol.* **2**, 265.
Thompson, R. F. (1967). "Foundations of Physiological Psychology." Harper, New York.



Ussing, H. H. (1960). *In* "Handbuch der Experimentellen Pharmakologie," Vol. 6, Part B, p. 1. Springer-Verlag, Berlin and New York.
Valenstein, E. S., Cox, V. C., and Kakolewski, J. W. (1970). *Psychol. Rev.* **77**, 16.
Waelsch, H., and Lajtha, A. (1960). *In* "Neurochemistry of Nucleotides and Amino Acids" (R. O. Brady and D. B. Tower, eds.), p. 205. Wiley, New York.
Watson, J. D. (1965). "Molecular Biology of The Gene." Benjamin, New York.
Weiss, P., and Hiscoe, H. B. (1946). *J. Exp. Zool.* **107**, 315.
Willows, A. O. D. (1968). *In* "Physiological and Biochemical Aspects of Nervous Integration" (F. O. Carlson, ed.), p. 217. Prentice-Hall, Englewood Cliffs, New Jersey.
Wyatt, H. W. (1964). *J. Theor. Biol.* **6**, 441.