On Learning, Information, Lateral Inhibition, and Transmitters

STEPHEN GROSSBERG
Department of Mathematics, Massachusetts Institute of Technology, Cambridge Massachusetts

Communicated by Richard Bellman

ABSTRACT

This article continues the derivation of a learning theory with neurophysiological implications. Equations are derived that contain formal analogs of physiological interactions between membrane potentials, spiking frequencies, transmitter production and release, and various trophic effects. These equations reduce to the Hartline–Ratliff equation for lateral inhibition in a special case. The need for inhibition in making possible perfect learning is suggested. A formal connection between contour enhancement due to lateral inhibition, the Ward–Hovland or "reminiscence" phenomenon in learning (i.e., spontaneous improvement of memory), bowing in serial verbal learning, and the information functional is described. A unified explanation of spatiotemporal masking and of the decrease of reaction time with increased learning or increased stimulus energy is given. A formal analog of the inward flow of Na^+ and the outward flow of K^+ through the membrane in response to excitatory transmitter and of the outward flow of K^+ in response to inhibitory transmitter is derived.

1. INTRODUCTION

A. Some "Neural" Properties

This article continues the program, introduced in [1], of deriving a learning theory with physiological implications from psychological postulates. Psychological postulates are herein reviewed and extended that give rise to equations that resemble such familiar neural phenomena as lateral inhibition [2], transmitter production and release [3], the production of spikes at suprathreshold membrane potential values [4],

the inward flow of Na^+ and outward flow of K^+ across the cell membrane in response to an excitatory transmitter [3], and the outward flow of K^+ in response to an inhibitory transmitter [3]. The equations reduce in a special case to the Hartline–Ratliff equation for lateral inhibition in the *Limulus* retina [5], and provide some theoretical information about the parameters that enter into determining the empirical coefficients of that equation. The equations also contain predictions about phenomena that have not as yet been experimentally determined; for example, a binding of Na^+, K^+, Ca^{++}, and Mg^{++} within the end bulb in complexes of varying strength to produce transmitter production and release rates that are sensitive to prior presynaptic and postsynaptic levels of membrane potential. Some of these facts are only sketched herein, since the article is primarily concerned with heuristic *gedanken* experiments that give rise to these equations. Later papers will derive the equations on a postulational basis, extend the formalism, and analyze the neural results in greater detail.

B. Lateral Inhibition and Perfect Learning

Lateral inhibition plays a useful role in making possible *perfect*, or *deterministic*, learning in the systems that we derive herein. In [1] are derived learning machines in which no manifestly inhibitory processes occur, and yet these machines can learn, remember, and recall in specific experimental situations [6–11]. Nonetheless, because some of the postulates used in this derivation are merely approximate, a severe limitation is placed on the behavior sequences that these machines can perfectly learn. The alphabet $\mathcal{A} = \text{ABC}, \ldots, \text{Z}$ can, in principle, be perfectly learned by some of the machines in [1], but the alternating sequence $\mathcal{B} = \text{ABACABAC}$ can be learned only statistically (i.e., some machines can learn that B and C each follow A 50% of the time in $\mathcal{B}$). Lateral inhibition arises in this article as one of the mechanisms used to construct machines that learn $\mathcal{B}$ deterministically.

Statistical learning of $\mathcal{B}$ is simpler than deterministic learning for an obvious reason. Statistical learning merely requires that a machine know the frequency with which B and C follow A. Deterministic learning requires that the machine know the subsequences of length greater than one that determine the next letter uniquely. For example, the subsequence BA uniquely determines C, and the subsequence CA uniquely determines B. Deterministic learning of $\mathcal{B}$ thus requires that the machine be able to

learn the *context* of letters that precede a given letter, not only the frequency with which one letter follows the next.

Inhibition is thus a suggested tool for solving the problem of deterministically learning behavior sequences in which nontrivial context effects occur. This problem arises in any behavioral language that has an alphabet (e.g., letters) that is smaller than its vocabulary (e.g., words). Only the simplest behavioral languages are not of this type, if only because the number of distinct behavioral signs that an individual can produce is so small. The universal appearance of lateral inhibition in neural systems therefore emerges as a counterpart of the universal problem of deterministically learning nontrivial contexts.

C. Lateral Inhibition and Information

We will show that lateral inhibition and the mathematical concept of information are closely related. It is well known that the information function, or entropy, of mathematical information theory can be uniquely derived from postulates that are a plausible formal realization of our intuitive concept of information [12]. The strong intuitive impression also exists that the nervous systems of animals are excellent information-processing systems. The basic question therefore arises: how does the information function "sit" in an animal's nervous system? We find that the transformation from inputs to outputs executed by our equations behaves qualitatively in special cases *as if* the mathematical information contained in the inputs were being computed by these equations. Lateral inhibition is crucial in producing these qualitative effects.

This article begins by pointing out some limitations of the systems in [1] and sketches a simple plausible way out of these limitations. We then label the mathematical variables of the new system in a natural neural way. The equations governing these variables thereby give rise to some neural expectations, which are stated along with references containing confirmatory neural data.

2. EXCITATORY EMBEDDING FIELDS

We briefly review some conclusions from [1] about our learning machines $\mathcal{M}$ as a point of departure for the present account.

A. Recognition

1. A "simple" behavioral symbol r_i (such as the letter A as used in speech) is represented in $\mathcal{M}$ by a single abstract state v_i.

2. The presentation of r_i to $\mathcal{M}$ at time $t = t_0$ is represented in $\mathcal{M}$ as an input pulse to v_i with onset time $t = t_0$. More complicated behavioral symbols are represented by correspondingly more complicated input patterns.

3. A nonnegative state function $x_i(t)$ sits at v_i and fluctuates in response to inputs. When all inputs are zero, $x_i(t)$ decays to zero. Increasing the input $I_i(t)$ received by v_i increases $x_i(t)$.

These facts are summarized in the equations

$$\dot{x}_i(t) = -\alpha x_i(t) + I_i(t), \qquad x_i(0) \geqslant 0, \tag{1}$$

for each symbol r_i, where $i = 1, 2, \ldots, n$. The input $I_i(t)$ is given by

$$I_i(t) = \sum_{m=1}^{N_i} J_i(t - t_{im})$$

where $J_i(t)$ is an input pulse with onset time $t = 0$, and t_{im} is the mth onset time of r_i.

In the special case that $x_i(t)$ has a fixed upper bound M_i, (1) is replaced by

$$\dot{x}_i(t) = -\alpha x_i(t) + (M_i - x_i(t))I_i(t), \qquad 0 \leqslant x_i(0) \leqslant M_i. \tag{1'}$$

B. Learning

1. After a list $r_i r_j$ of symbols has been learned by $\mathcal{M}$, a presentation of r_i to $\mathcal{M}$ creates the reply r_j a short time, say, τ_{ij} time units, later. That is, an *input* to v_i creates an *output* from v_j τ_{ij} time units later. To accomplish this, a signal will be sent from v_i to v_j at a finite velocity over a pathway e_{ij}; e_{ij} is a *directed* pathway because the lists $r_i r_j$ and $r_j r_i$ are not the same. The pathway e_{ij} is therefore drawn as an arrow from v_i to v_j.

2. Before learning occurs, $\mathcal{M}$ will be able to learn several lists $r_i r_{j_1}, r_i r_{j_2}, \ldots, r_i r_{j_k}$, or else r_i would already be $\mathcal{M}$'s only reply to r_i. Therefore, v_i can send signals to all points v_j that represent possible replies r_j to r_i.

By Section 2, B, 1 and 2, a process $z_{ij}(t)$ will exist that controls the size of signals from v_i to v_j at time t; $z_{ij}(t)$ measures the frequency with which r_i and r_j have been presented consecutively to $\mathcal{M}$ in the past, and cuts off the signal from v_i to v_j if this frequency is small.

3. Since $z_{ij}(t)$ measures the frequency of consecutive r_i and r_j presentations, it takes place in $\mathcal{M}$ at a position where past $x_i(\xi)$ signals ($\xi < t$) and present $x_j(t)$ signals coexist. Only one such place exists in $\mathcal{M}$, namely, the *arrowhead* N_{ij} of e_{ij} at which the signal from v_i to N_{ij} is contiguous with $x_j(t)$ in v_j.

Section 2, A and B gives rise to the equations (system (*))

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta \sum_{m=1}^{n} x_m(t - \tau_{mi}) p_{mi} z_{mi}(t) + I_i(t) \tag{2}$$

and

$$\dot{z}_{jk}(t) = -u z_{jk}(t) + \beta p_{jk} x_j(t - \tau_{jk}) x_k(t) \tag{3}$$

where $i, j, k = 1, 2, \ldots, n$. The rate parameters α, β, u are positive, and the time lags τ_{jk} are positive. The nonnegative coefficients p_{jk} determine the relative strength of signals from v_j to N_{jk}. The initial data of (2) and (3) are nonnegative and continuous.

In the special case that $x_i(t)$ has a finite maximum M_i and $z_{jk}(t)$ has a finite maximum M_{jk}, Eqs. (2) and (3) are replaced by

$$\dot{x}_i(t) = -\alpha x_i(t) + (M_i - x_i(t))\left[\beta \sum_{m=1}^{n} x_m(t - \tau_{mi}) p_{mi} z_{mi}(t) + I_i(t)\right] \tag{2'}$$

and

$$\dot{z}_{jk}(t) = -u z_{jk}(t) + \beta p_{jk}(M_{jk} - z_{jk}(t)) x_k(t - \tau_{jk}) x_k(t), \tag{3'}$$

with initial data chosen to satisfy $0 \leqslant x_i \leqslant M_i$ and $0 \leqslant z_{jk} \leqslant M_{jk}$. Henceforth we concentrate on the system (*).

Although (*) is a mathematically well-defined system, its learning is hampered by the following fact. Suppose that $p_{ij} > 0$ and $p_{ik} > 0$: that is, $\mathcal{M}$ can in principle learn both $r_i r_j$ and $r_i r_k$. Suppose that $r_i r_j$ is presented to $\mathcal{M}$. Then signals are sent by v_i to both v_j and v_k. $z_{ij}(t)$ grows faster than $z_{ik}(t)$ does, because v_j also receives a positive input whenever r_j occurs. But the signal from v_i to v_k excites $x_k(t)$, which in turn helps to keep $z_{ik}(t)$ from decaying, as (3) shows. Although some learning of $r_i r_j$ occurs, the process is not very efficient in all cases.

In [1] we go on to seek a modification of (*) in which the inequality

$$p_{ij} z_{ij}(t) \gg \max\{p_{ik} z_{ik}(t) : k \neq j\}$$

will suffice to cut off the signal from v_i to v_k in a more efficient manner. This is accomplished in the case $\tau_{ij} = \tau$ by replacing $p_{mi} z_{mi}(t)$ in (2) by the normalized function

$$y_{mi}(t) = p_{mi} z_{mi}(t)\left(\sum_{m=1}^{n} p_{mk} z_{mk}(t)\right)^{-1}.$$

That is, we replace (*) by the system (**)

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta \sum_{m=1}^{n} x_m(t - \tau) y_{mi}(t) + I_i(t), \tag{4}$$

$$y_{jk}(t) = p_{jk} z_{jk}(t)\left(\sum_{m=1}^{n} p_{jm} z_{jm}(t)\right)^{-1}, \tag{5}$$

and

$$\dot{z}_{jk}(t) = -\alpha z_{jk}(t) + \beta p_{jk} x_j(t-\tau) x_k(t). \tag{6}$$

Systems (*) and (**) are both examples of *embedding fields*. These fields are called *excitatory* because their variables, including the interaction terms $\beta x_m(t-\tau_{mi}) p_{mi} z_{mi}(t)$ of (*) and $\beta x_m(t-\tau) y_{mi}(t)$ of (**) are always nonnegative.

We now itemize some limitations of excitatory embedding fields.

3. NONLOCAL ASSOCIATIONS IN A LOCAL FLOW

Consider the interaction term $\beta x_m(t-\tau) y_{mi}(t)$ from v_m to v_i in (**). This interaction has the following interpretation. Vertex v_m sends out a signal $\beta x_m(t-\tau)$ along e_{mi} at time $t-\tau$. This signal travels along e_{mi} until it reaches the arrowhead N_{mi} of e_{mi} at time t. The process $y_{mi}(t)$ in N_{mi} is thereupon activated and the quantity $\beta x_m(t-\tau) y_{mi}(t)$ is instantaneously transmitted from N_{mi} to v_i. Process $y_{mi}(t)$, in turn, is constructed from the ratio of $p_{mi} z_{mi}(t)$ and $\sum_{k=1}^{n} p_{mk} z_{mk}(t)$. Each $z_{mk}(t)$ cross-correlates the pulse $\beta p_{mk} x_m(t-\tau)$ received from v_m by the arrowhead N_{mk} of e_{mk} at time t with the value $x_k(t)$ of v_k contiguous to the arrowhead at this time. The puzzle thereupon arises: for any $k \neq i$, how does the cross-correlation $z_{mk}(t)$, which sits at the arrowhead N_{mk} at time t, make itself instantaneously felt at the arrowhead N_{mi} in $y_{mi}(t)$?

This difficulty cannot be overcome without changing (**), since no provisions are therein made for a transport of any $z_{mk}(t)$ quantity from one arrowhead to another. This change must not, however, destroy the improvements in learning achieved by replacing $p_{mi} z_{mi}(t)$ in (2) by $y_{mi}(t)$ in (4). The basic improvement is that an increase in $z_{mk}(t)$ causes not only an increase in $y_{mk}(t)$, but also a *decrease* in $y_{mi}(t)$, for all $i \neq k$, and conversely. That is, growth of the "associational strength" $y_{mi}(t)$ is *inhibited* by growth of the associational strength $y_{mk}(t)$, $i \neq k$, and conversely. The replacement of $p_{mi} z_{mi}(t)$ by $y_{mi}(t)$ is thus a way of introducing inhibition into an embedding field without sacrificing the nonnegativity of all its variables.

A heavy conceptual price is paid for this formal improvement; namely, the introduction of a "virtual" process that instantaneously leaps from arrowhead to arrowhead without any apparent geometrical mechanism

underlying it. Since this price is conceptually intolerable, we must seek another way to realize the mutual inhibition of associational strengths.

We will find that essentially only one method of doing this is available to us. This method will sacrifice the nonnegativity of the $x_i(t)$ values to eliminate the virtual inhibitory process. As shown in [1], nonnegativity of the $x_i(t)$'s means that all states of $\mathcal{M}$ are observable to a psychological experimenter studying $\mathcal{M}$. Thus, observability must be sacrificed to create a dynamics of learning that corresponds in a sensible way to the geometry on which it plays.

4. LATERAL INHIBITION: LOCALITY VERSUS OBSERVABILITY

A way of preserving inhibition between associations $y_{ij}(t)$ and $y_{ik}(t)$ without requiring a virtual process is easily seen. It is clear that *some* process must carry this inhibition between N_{ij} and N_{ik}. We know only one process that carries information around the network. This is the process whereby the *states* v_j and v_k send signals to one another. In other words, *inhibitory signals* must pass between the states v_j and v_k. Only the process $x_j(t)$ can create a signal from v_j to v_k, and only the process $x_k(t)$ can create a signal from v_k to v_j. The signals with which we are familiar in (**) are *excitatory* in the sense that an increase in $x_j(t)$ causes an increase in $x_k(t)$ after a suitable time lag elapses. An *inhibitory* signal from v_j to v_k has the property that an increase in $x_j(t)$ causes a *decrease* in $x_k(t)$ after a suitable time lag elapses.

Can such an inhibitory signal from v_j to v_k alter $z_{ik}(t)$ in much the same way that the transformation from $z_{ik}(t)$ to $y_{ik}(t)$ does? The answer is manifestly yes, because as (6) shows, a decrease (or increase) in $x_k(t)$ causes a corresponding decrease (or increase) in $z_{ik}(t)$. Just as the virtual inhibition from $y_{ij}(t)$ to $y_{ik}(t)$ occurs instantaneously, the inhibitory signals from v_j to v_k must occur "rapidly" where by this we clearly mean: (1) more rapidly than excitatory signals, and (2) rapidly compared to the rate with which the correlations $z_{ij}(t)$ and $z_{ik}(t)$ change in response to changes in $x_j(t)$ and $x_k(t)$, respectively.

We conclude that rapid inhibitory signals from v_j to v_k can inhibit $z_{ik}(t)$ much as the transformation from $z_{ik}(t)$ to $y_{ik}(t)$ does. The great advantage of the former method of inhibiting $z_{ik}(t)$ is that all interactions can then be described by *local* processes; that is, by processes that move along the geometry of $\mathcal{M}$ in a sensible way. The loss of observability is a small price to pay for this advantage.

5. INHIBITION IN THE MIDDLE OF A LONG LIST

In a previous paper [8], the bowed curve of serial verbal learning was explained, using (**). Bowing occurs at a fixed item r_i in the middle of a long list because the associations $y_{jk}(t)$ remain quite uniformly distributed in k for a relatively long amount of time.

In terms of inhibition between states, this means that inhibition among states is maximal when items in the middle of the list occur. This fact corroborates the classical work of Hull and his associates ([13], page 516), who also postulated an "accumulation of inhibition" near a list's middle. Of course, the actual mechanisms that Hull envisaged are not the same as our own, but it is altogether remarkable that he reached this qualitative conclusion some thirty odd years ago.

The introduction of inhibition between states to replace virtual inhibition between correlations carries with it a substantial conceptual consequence. Having identified Hull's "accumulating inhibition" as inhibition between states, we will, in Sections 11–13, identify inhibition between states as lateral inhibition of the neural variety, and will be able to formally derive the Hartline–Ratliff equation. A unifying conceptual bridge is hereby constructed from the seemingly diverse phenomena of serial verbal learning in humans to lateral inhibition within the *Limulus* retina.

6. INHIBITION AND SELF-IMPROVING MEMORY

Using the functions $y_{ij}(t)$ instead of $z_{ij}(t)$ in (**) has an important effect on the way in which $\mathcal{M}$ remembers. The following fact has been rigorously proved [7, 9, 10]. If $\mathcal{M}$ has learned a given task to a "moderate" degree of accuracy, then $\mathcal{M}$'s memory of the task improves spontaneously without further practice. Since we have replaced the $y_{ij}(t)$'s by inhibitory signals, we should now be able to say that inhibition between states helps to create self-improving memories. This is indeed the case.

Speaking roughly, the mechanism is as follows. First, a rapid inhibition between states eliminates behaviorally insignificant (e.g., uniformly distributed) background noise. The more slowly varying associations $z_{jk}(t)$ then learn only from the behaviorally significant (e.g., dominant) states. The data on which $\mathcal{M}$ constructs its memory are thus "crisper" than the data that $\mathcal{M}$ receives. Once a significant $z_{jk}(t)$ association is

formed, it tends to create a distribution of state values that are compatible with its existence by funneling large excitatory signals from v_j to v_k. Then inhibition between states "crispens" these state values and thereby causes a further increase in $z_{jk}(t)$.

The "crispening" of associations $z_{jk}(t)$ is a new idea concerning the way in which $\mathcal{M}$ learns. It is, however, intimately connected with the "perceptual crispening" or "contour enhancement" due to lateral inhibition that has been a subject of experimental interest since the time of Mach [5]. Perceptual crispening has been noticed by experimentalists for some time because it leads to a crispening of neural signals and thus of behavior itself. Crispening of associations has remained unnoticed both because it is coupled so intimately to perceptual crispening effects, and because it describes a change within the end bulbs, and therefore creates no directly observable signals.

7. OBSERVABLE PROBABILITY MODELS VERSUS UNOBSERVABLE INHIBITION BETWEEN STATES

At the risk of belaboring the now obvious, we emphasize an important conceptual consequence of replacing virtual inhibition between correlations by a real inhibition between states. We perhaps all share the strong intuitive belief that we can legitimately hope to describe the observable behavior of many organisms by using some kind of probability model, and it seems perfectly natural to discuss the transition probability that the organism will move from one behavioral state to the next at any given time t. We are equally aware that much of our behavior seems to be quite deterministic. For example, the very act of writing a mathematical paper has approximately probability zero in any reasonably simple probabilistic model for producing letters and numbers. How can the seemingly contradictory intuitive impressions of merely probabilistic versus deterministic behavior be reconciled in a philosophically satisfactory way?

The simple model (**) suggests a way: (**) is manifestly a deterministic model, in the sense that it describes a prediction theory for individual inputs and outputs. System (**) also contains transition probabilities of moving from a given behavioral state v_i to another state v_j at time t, namely, the associational strengths $y_{ij}(t)$. In particular, when $y_{ij}(t) \cong 1$, then the probabilistic transition from v_i to v_j at time t will look as deterministic as we please. In this sense, (**) is both deterministic and probabilistic at every instant of time.

The properties that define a probability distribution, namely,

$$\text{(a) } y_{1i}(t) \geqslant 0 \quad \text{and} \quad \text{(b) } \sum_{j=1}^{n} y_{1j}(t) = 1,$$

are, within (**), consequences of two facts: (a′) observability, and (b′) the need in an observable model for virtual inhibition between correlations to guarantee efficient learning. Observability, or (a′), must, however, be sacrificed to preserve the more basic property of locality. Virtual inhibition between correlations, or (b′), must then be replaced by real inhibition between states. The "impression" that the observable behavior of system (**) is probabilistic is hereby formally translated into a wholly deterministic, but not wholly observable, process in which both excitatory and inhibitory factors are admitted.

Since the basic ideas, such as locality, that led to these conclusions are so very simple and general, it seems quite possible that analogs therefore exist in many other systems in which evolutionary (or learning) trends appear. That is, the seemingly probabilistic dynamics of these processes might well be the observable epiphenomena of a not entirely observable interaction between excitatory and inhibitory factors.

8. ENTROPY AND SUFFICIENT REASON: NONLOCAL OUTPUTS OF LOCAL STATES

One of the conceptual deficiencies of model (**) is that it does not abide by the principle of sufficient reason. By this we mean that presentation of a symbol r_j to a machine $\mathcal{M}$ can produce large outputs from several states v_i even when no symbol r_i is a preferred response to r_j.

For example, consider an outstar [9] with source v_1 and border $B = \{v_i: i = 2, 3, \ldots, n\}$. That is, let $p_{1i} = 1/(n-1)$, $i = 2, 3, \ldots, n$, and let all other $p_{ij} = 0$. Then (**) becomes

$$\dot{x}_1(t) = -\alpha x_1(t) + I_1(t),$$

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta x_1(t - \tau) y_{1i}(t) + I_i(t),$$

$$y_{1i}(t) = z_{1i}(t)\left[\sum_{m=2}^{n} z_{1m}(t)\right]^{-1},$$

and

$$\dot{z}_{1i}(t) = -u z_{1i}(t) + \beta x_1(t - \tau) x_i(t), \qquad i = 2, 3, \ldots, n.$$

Suppose that no list $r_1 r_i$, $i = 2, 3, \ldots, n$, is preferred to any other in $\mathcal{M}$ before time $t - \tau$. That is, all $x_i(\xi)$ are equal and all $y_{1i}(\xi) = 1/(n-1)$, where $\xi \leqslant t - \tau$ and $i = 2, 3, \ldots, n$. Let r_1 be presented to

$\mathcal{M}$ at time $\xi = t - \tau$. Then $x_1(\xi)$ becomes large shortly after time $\xi = t - \tau$. Consequently, the signal $\beta x_1(\xi) y_{1i}(\xi + \tau)$ from v_1 to v_i also increases. Since $y_{1i}(\xi + \tau) = 1/(n-1)$, each signal is the same. Moreover, the increase in this common signal is substantial, and in any case depends only on the number $n - 1$ of points in B. The outputs $x_i(\xi)$ from all v_i therefore increase substantially shortly after time $\xi = t$. These large outputs occur in spite of the fact that all $x_i(\xi)$ remain equal and all $y_{1i}(\xi) = 1/(n-1)$ throughout this experiment on $\mathcal{M}$. That is, a large output from v_i can occur in response to the presentation of r_1 even though r_i is not a preferred response alternative to r_1.

This difficulty is formally overcome in [9] by modifying the output from v_i. Instead of using $x_i(t)$ itself as the output, we use

$$O_i^{(1)}(t) = \max\left\{0, x_i(t)\left[1 - \frac{H(X_2(t), X_3(t), \ldots, X_n(t))}{\ln_2(n-1)}\right] - \Gamma_i\right\},$$

$i = 2, 3, \ldots, n$, where Γ_i is some positive response threshold,

$$X_i(t) = x_i(t)\left(\sum_{m=2}^{n} x_m(t)\right)^{-1},$$

and

$$H(q_1, q_2, \ldots, q_{n-1}) = -\sum_{m=1}^{n-1} q_m \ln_2 q_m$$

for any probability distribution $\{q_i: i = 1, 2, \ldots, n-1\}$; H is the familiar information function, or entropy, of mathematical information theory [12].

The modified output $O_i^{(1)}(t)$ realizes the principle of sufficient reason in $\mathcal{M}$ in the following way. Suppose that no list $r_1 r_j$ is preferred by $\mathcal{M}$ at time t; in particular, all $x_j(t)$ are equal. Then $O_i^{(1)}(t) = 0$ for all $i = 2, 3, \ldots, n$ no matter how large the common $x_j(t)$ value is. That is, in order that a modified output $O_i^{(1)}(t)$ from v_i occur, it must represent a preference by $\mathcal{M}$ for the list $r_1 r_i$.

Consider now the case in which a definite preference for $r_1 r_i$ does exist within $\mathcal{M}$ at time t. Suppose in particular that $x_i(t)$ far exceeds all $x_j(t)$ values in size, $j \neq 1, i$. Then all $O_i^{(1)}(t)$ reduce essentially to the old outputs $x_j(t)$, with the improvement that background noise is eliminated by the thresholds Γ_j. In mathematical terms this means:

$$x_i(t) \gg x_j(t) \cong 0, \qquad \text{for all } j \neq 1, i,$$

implies that

$$O_j^{(\Gamma)}(t) \cong \max\{0, x_j(t) - \Gamma_j\}.$$

Since $x_j(t) \cong 0$ and $\Gamma_j > 0$ for all $j \neq 1, i$,

$$O_j^{(\Gamma)}(t) \cong 0, \qquad j \neq i,$$

whereas

$$O_i^{(\Gamma)}(t) \cong x_i(t) - \Gamma_i,$$

which differs from $x_i(t)$ only by an inessential scaling factor Γ_i.

Although the modified outputs $O_i^{(\Gamma)}(t)$ eliminate an unpleasant conceptual difficulty in a formal way, they create yet another conceptual difficulty. In order to evaluate $O_i^{(\Gamma)}(t)$, $\mathcal{M}$ must instantaneously bring together *all* $x_j(t)$ values from B to evaluate the probabilities $X_j(t)$, must then compute the entropy of these probabilities, and must then compute $O_i^{(\Gamma)}(t)$ itself at each state v_i. Yet within $\mathcal{M}$ itself, there exist no geometrical pathways, let alone dynamical reasons, by which these transformations can take place. Again locality has been violated. Model (**) must therefore be modified to eliminate this conceptual difficulty without losing the formal advantages of $O_i^{(\Gamma)}(t)$.

The transformation from $x_i(t)$ to $O_i^{(\Gamma)}(t)$ replaces equal $x_i(t)$'s by zeros, but allows the $O_i^{(\Gamma)}(t)$ value corresponding to a unique large $x_i(t)$ value to remain uninfluenced by the other $x_j(t)$ values except for a threshold shift. Thus the transformation from $x_i(t)$ to $O_i^{(\Gamma)}(t)$ acts as if the $x_i(t)$'s inhibit each other out of existence when they are equal, but a unique large $x_i(t)$ value inhibits all other $x_j(t)$ values to zero without suffering reciprocal inhibition.

We therefore conclude that both the transformation

$$z_{ij}(t) \to y_{ij}(t)$$

between associations and the transformation

$$x_i(t) \to O_i^{(\Gamma)}(t)$$

between outputs covertly describe an inhibition between states.

The remarks above suggest that the entropy H "sits" in $\mathcal{M}$ as an inhibitory mechanism that accentuates dominant $x_i(t)$ values and annihilates background noise to prepare these data for subsequent learning by the associations $z_{jk}(t)$. This role for H is certainly compatible with the intuitive idea that H plays some part in the information processing of a learning machine.

9. WHAT IS A "SIMPLE" ACT: NONLOCAL PATTERNS WITH LOCAL CONTROLS

Every "simple" behavioral symbol r_i has been placed in correspondence with a single abstract state v_i in $\mathcal{M}$. Learning a sequence of simple symbols $r_1 r_2 \cdots r_L$ can be viewed as the formation of a new composite symbol that itself gradually becomes simple as our experience with it grows. Should not then the new symbol $r_1 r_2 \cdots r_L$ eventually correspond to a new state? If this is so, then are not the states v_i to which the original simple symbols r_i correspond also, in some sense, composite states?

Questions such as these illustrate the approximate status of the postulate that a simple symbol r_i corresponds to a single abstract state v_i in $\mathcal{M}$. Even the production of such seemingly simple symbols as A requires the integration of very complicated muscular motions that cannot adequately be described by a single state, and are carried through the air by sound waves of great complexity. We must therefore expect to eventually find that even simple behavioral symbols are represented by aggregates of many states interacting together. In the next section we discuss an example that provides some formal reasons why this will be so. This one–many correspondence between symbols and states will gradually lose its abstract character as we derive better neural equations for the states.

Given such a one–many correspondence between symbols and states, the question why simple behavioral symbols "seem" to be simple becomes an urgent one. A partial answer is fortunately already provided by studying the way in which (**) learns the alphabet ABC $\cdots$ Z.

Consider a machine $\mathcal{M}$ obeying (**) that can learn the alphabet. Once the alphabet is learned, a *chain* of associations

$$y_{AB} \cong y_{BC} \cong y_{CD} \cong \cdots \cong y_{XY} \cong y_{YZ} \cong 1$$

is formed. The alphabet seems simple to us once we know it because we can simply "rattle it off" once we choose to do so. Can $\mathcal{M}$ "rattle off the alphabet?" The answer is manifestly yes.

Simply present A to $\mathcal{M}$ after $\mathcal{M}$ has learned the alphabet. Then $x_A(t)$ grows and sends a signal *only* to v_B. A large output $x_B(t)$ (or $O_B^{(\Gamma)}(t)$!) is created at v_B τ time units later. Suppose that this output rapidly creates a *feedback input* to v_B through the medium surrounding $\mathcal{M}$ (much as we "hear ourselves talk"). Then $x_B(t)$ grows further and a signal is sent *only* to v_C. A large output $x_C(t)$ occurs from v_C τ time units later, and hereby creates a rapid feedback input to v_C. The process of sending signals one

step at a time and reinforcing these signals by rapid feedback inputs through the medium surrounding $\mathcal{M}$ continues in this way until all the outputs $x_A(t), x_B(t), \ldots, x_Z(t)$ have become large, one at a time, every τ time units; that is, until the entire alphabet has been said by $\mathcal{M}$.

The alphabet thus seems simple to $\mathcal{M}$, even though it is a composite symbol, because the associations $y_{AB}, y_{BC}, \ldots, y_{YZ}$ form a "one-dimensional" chain. Once the "local control" v_A is activated by an input, the entire chain is activated step by step.

This analysis requires some kind of feedback input to keep the signals in successive links of the chain from dying out. It is in fact well known [14] that performing long and complicated behavioral sequences is hampered if sensory or neural feedback created by performance of the links in the sequence is eliminated.

The following section, which contains a somewhat technical example, can be skipped on a first reading. In it are pointed out some formal reasons why a one–many correspondence between symbols and states is needed if we wish aggregates of simple symbols to become simple when we know them well enough.

10. HIGHER-ORDER ASSOCIATIONS

By taking seriously the idea that composite symbols gradually become simple as our experience with them grows, we can gain further insight into some of the properties that an improved version of (**) might have. Let a machine $\mathcal{M}$ be given, with points v_i, $i = 1, 2, \ldots, n$, and edges e_{jk} leading from each point v_j to every *other* point v_k with equal weight $p_{jk} = 1/(n-1)$. $\mathcal{M}$ is thus a *complete graph without loops* [8, 10, 11]. Consider the problem of teaching $\mathcal{M}$ the alphabet $r_1 r_2 \cdots r_n$. We wish to say that every subsequence $r_i r_{i+1} r_{i+2} \cdots r_{i+k}$ of the alphabet begins to act like a simple state as the alphabet (and all its subsequences) is learned. What is a simple way of changing the equations of an excitatory embedding field to accomplish this aim? The following discussion provides an answer to this question.

Consider the alphabet ABCD $\cdots$ Z for definiteness. *Before* the alphabet is learned, none of the letters is linked to any other by preferential associations. The prediction of (say) the letter D by the machine will thus be determined by the previous letters A, B, or C *acting separately*. We express this fact symbolically by

$$(A)(B)(C) \rightarrow D.$$

After the alphabet is learned, each subset of it has some status as a simple symbol, and thus *all* the subsets constructed from A, B, and C can contribute to a prediction of D, but to differing degrees. Thus we write

$$\left.\begin{array}{l}(A)(B)(C)\\(ABC)\\(BC)\\(AB)\\\cdot\\\cdot\\\cdot\end{array}\right\} \rightarrow D.$$

How much weight does each of these subsets carry in the prediction of D? The answer manifestly depends on the amount of time that elapses between presentation of successive letters to the machine [1]. If this time is large, then only C immediately precedes presentations of D, and so only (C) can influence the prediction of D. If A has been said long ago, but B and C are presented rapidly before D is, then only (B), (C), and (BC) can influence the prediction of D. We must therefore seek a way of mathematically expressing the fact that only sets of letters that were presented contiguously in time to a given letter substantially influence the prediction of that letter.

There is a simple way to say at time t that B and C have recently been said. It is: the product $x_B(t)x_C(t)$ is large. Indeed, if any set

$$I(J) = \{r_k : k \in J \equiv \{j_1, j_2, \ldots, j_m\}\}$$

of distinct symbols has recently been said, then the product

$$\prod_{k \in J} x_k(t)$$

is large. We need now only find a way to say that if $I(J)$ and r_j have been consecutively presented very often in the past, then presenting $I(J)$ alone in the future generates an output from v_j. Model (**) solves this problem for the special case that

$$I(J) = \{r_i\}.$$

Since the interaction corresponding to the transition $r_i \rightarrow r_j$ is

$$\beta x_i(t-\tau) y_{ij}(t),$$

we suppose by analogy that the interaction of the transition $I(J) \to r_j$ is given by

$$\dot{z}_{Jj}(t) = -az_{Jj}(t) + \beta p_{Jj} \prod_{k \in J} x_k(t-\tau)x_j(t).$$

It remains only to define $y_{Jj}(t)$ in terms of the correlations $z_{Jm}(t)$. Formally this can be done by analogy with (5) as

$$y_{Jj}(t) = p_{Jj}z_{Jj}(t)\left(\sum_{m=1}^{n} p_{Jm}z_{Jm}(t)\right)^{-1}.$$

We have hereby derived the following equations (system (***)) for an excitatory embedding field *with higher-order correlations*.

$$\dot{x}_i(t) = -\alpha x_i(t) + \beta \sum_{J \subset \mathcal{A}} \prod_{k \in J} x_k(t-\tau)y_{Ji}(t) + I_i(t).$$

$$y_{Ji}(t) = p_{Ji}z_{Ji}(t)\left(\sum_{m=1}^{n} p_{Jm}z_{Jm}(t)\right)^{-1},$$

and

$$\dot{z}_{Ji}(t) = -az_{Ji}(t) + \beta p_{Ji} \prod_{k \in J} x_k(t-\tau)x_i(t),$$

$i = 1, 2, \ldots, n$, where $\mathcal{A} = \{1, 2, \ldots, n\}$.

This system has the following desirable property. If a set $I(J)$ is often paired with r_j, then z_{Jj} and hence y_{Jj} will grow. A future occurrence of $I(J)$ will therefore guarantee that the interaction $\prod_{k \in J} x_k(t-\tau)y_{Jj}(t)$ becomes large, and thus that the output $x_j(t)$ is large. In this sense, the set $I(J)$ eventually acts like a simple symbol within $\mathcal{M}$. If $I(J)$ is often paired with r_j, then every subset $\tilde{I}$ of $I(J)$ will also be paired with r_j. The relative contribution of each subset $\tilde{I}$ to the output v_j depends on the relative timing whereby symbols in $\tilde{I}$ and r_j are presented to $\mathcal{M}$. Consider for example the case in which

$$p_{Ji} = p^{|J|}, \qquad |J| > 1,$$

where $|J|$ is the number of indices in J and p is chosen so small that

$$px_m(t) \leqslant 1$$

for all $t \geqslant 0$. Start $\mathcal{M}$ in a state of "rest" and "maximal ignorance" [1], and let

$$I(i, j) = \{j - i, j - i + 1, j - i + 2, \ldots, j - 1\}.$$

Then the correlations

$$z_{I(i,j)j}(t)$$

will be monotone decreasing in i, corresponding to the idea that symbols r_{j-i}, which occur long before r_j does, cannot easily become associated with r_j.

Although the system (***) embodies some useful formal properties, it is filled with conceptual difficulties. These difficulties are, however, quite informative. We now discuss some of them.

A. Reducing Higher-Order Associations to Simple Associations

Where and how are the products $\prod_{k \in J} x_k(t)$ computed? Figure 1 partially answers this question: it introduces a *compound edge* e_{Ji} for

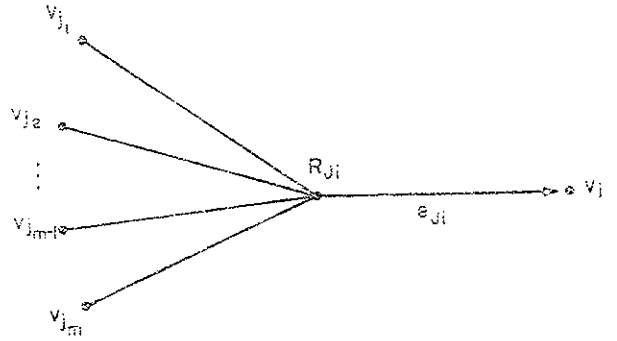

FIG. 1

every set of indices J, in addition to the simple edges e_{ji} from v_j to v_i. The signals $x_k(t)$, $k \in J$, travel from v_k along e_{ji} until they reach R_{Ji}, where the product $p_{Ji}\prod_{k \in J} x_k(t)$ travels to the compound arrowhead N_{Ji}, where $z_{Ji}(t+\tau)$ correlates it with $x_i(t+\tau)$.

The process whereby this product is computed and then correlated within z_{Ji} is entirely mysterious in Fig. 1. We will fortunately find, in Section 19, that the process mentioned in Sections 3 and 8 can (a) compute products of states, and can thereby (b) reduce higher-order associations $z_{Ji}(t)$ to combinations of first-order associations $z_{jk}(t)$. This reduction property provides an alternative tool for "simplifying" learned behavior without creating formidable technical difficulties, as the following paragraphs show.

B. Overlapping Nonlocal Patterns Reduce Background Noise

There is too much background noise in (***). Every time m points v_j, $j \in I(J)$, are rapidly excited, $n(2^m - 1) - m$ edges of the form e_{Ki} are activated, over subsets $K \subset J$ and $i = 1, 2, \ldots, n$. The signals

received by the n points v_i are therefore enormous and often obscure correlations that have behavioral significance. This difficulty arises because we have fixed the number of states v_i in a one-to-one correspondence with the simple symbols r_i, and have then been forced to describe the action of all composite symbols using a hierarchy of higher-order associations between the states. The result is a graph with n points and $n(2^n - 2)$ edges in which each point is victimized by furious input signals.

The way out is clear: divide the burden of representing composite symbols more equally between points and edges. Merely adding points will not help unless these points represent symbols in the proper manner. For example, suppose that every simple symbol r_i corresponds to k_i points $V_i = \{v_{i1}, v_{i2}, \ldots, v_{ik_i}\}$, where $\sum_{i=1}^{n} k_i = n(2^n - 2)$, and no point in V_i is the same as any point in V_j, $j \neq i$. The number of points now equals the number of edges, but the input signals reaching a given point can be just as powerful as they were in the previous example.

We will, instead, allow some of the points in V_i to be the same as some of the points in V_j. That is, each simple symbol r_i can correspond to *many* points v_i, and each v_i can participate in the representation of *many* symbols r_i if we wish well-learned composite symbols to act in a simple way without creating uncontrollable background noise. It is, of course, well known from electrode studies of individual neurons that a single neuron can fire very actively in response to many seemingly diverse inputs [15]. This fact should not, as is often the case, be viewed as an example of "chaos" within the brain. Quite the contrary, it helps to guarantee efficient learning.

11. UNBOUNDED SIGNED EMBEDDING FIELDS

Section 4 provides a directive for eliminating deficiencies of excitatory embedding fields by introducing fields with both excitatory and inhibitory elements, which we choose to call *signed embedding fields*. We now translate this directive into the simplest set of equations that can replace (**), and then proceed to reap some happy neural consequences.

A. Signal Thresholds

In (**), the output from the state v_i at time t is proportional to $x_i(t)$ Section 8 replaces this output by $O_i^{(\Gamma)}(t)$, but $O_i^{(\Gamma)}(t)$ is conceptually unappealing because it is "nonlocal" relative to the geometry of the flow.

Output $O_i^{(\Gamma)}(t)$ has a "local" form, however, in exactly one case, namely, if *only* the state v_i has a large $x_i(t)$ value. Then

$$O_i^{(\Gamma)}(t) \cong \max(x_i(t) - \Gamma_i, 0), \tag{7}$$

which depends only on $x_i(t)$ and a known positive threshold Γ_i. The principle of sufficient reason can now be invoked to show that $x_i(t)$ must be replaced by (7) in *all* output expressions. To see this, let a "local"(or "near-sighted") observer $\mathcal{O}$ inhabit the state v_i; $\mathcal{O}$ cannot determine whether or not exactly one state has a large value at time t, if only because signals from state to state are not transmitted instantaneously. Nor can $\mathcal{O}$ distinguish the destination of one output signal emitted by v_i from that of any other. In this sense, sufficient reason requires that the output function from v_i be independent of the distribution of values at other states. Since locality requires that this function be (7) for one distribution of values, (7) is the output for *all* distributions of values. In particular,the excitatory signal received by the arrowhead N_{ij} from v_i at time t is, instead of $\beta_i x_i(t - \tau_{ij}) p_{ij}$,

$$\beta_i \max[x_i(t - \tau_{ij}) - \Gamma_{ij}, 0] p_{ij}, \tag{8}$$

where we have indexed the threshold Γ_{ij} by j as well as i to admit the possibility that signals to different arrowheads have different thresholds. Similarly, β_i can depend on the state v_i.

In order to unambiguously designate that (8) represents an excitatory signal, we label the parameters in (8) with superscript +'s; (8) becomes

$$\beta_i^+ \max[x_i(t - \tau_{ij}^+) - \Gamma_{ij}^+, 0] p_{ij}^+. \tag{9}$$

To shorten the writing of (9), we introduce the notation

$$[\omega]^+ = \max(\omega, 0). \tag{10}$$

Then (9) becomes

$$\beta_i^+ [x_i^+(t - \tau_{ij}) - \Gamma_{ij}^+]^+ p_{ij}^+. \tag{11}$$

REMARK. Given that $x_i(t)$ is always replaced by the local form $\max(x_i(t) - \Gamma_i, 0)$ of $O_i^{(\Gamma)}(t)$, how can we possibly recapture the *nonlocal* properties of $O_i^{(\Gamma)}(t)$ when more than one state has a large value? We will find that these nonlocal properties are consequences of the *interactions*, both excitatory and inhibitory, between the various states.

The nonlocal association $y_{ij}(t)$ of (**) will now be replaced by the local association $z_{ij}(t)$ of (*). The excitatory signal received by v_j from v_i at time t is therefore, by (11),

$$\beta_i^+ [x_i(t - \tau_{ij}^+) - \Gamma_{ij}^+]^+ p_{ij}^+ z_{ij}(t). \tag{12}$$

The total excitatory signal received by v_i from all states v_m at time t is

$$J_i^+(t) = \sum_{m=1}^{n} \beta_{mi}^+ [x_m(t - \tau_{mi}^+) - \Gamma_{mi}^+]^+ p_{mi}^+ z_{mi}(t). \tag{13}$$

The inhibitory signal received by v_j from v_i at time t is defined by analogy with (12). Two considerations determine this signal: (1) an increase in x_i causes a *decrease* in x_j τ_{ij}^- time units later; and (2) no reason

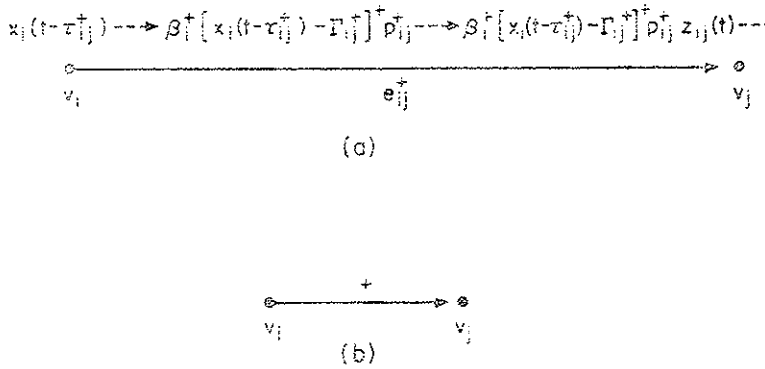

FIG. 2

yet exists for supposing that an inhibitory association exists. (A reason will be provided in Section 21.) The inhibitory signal received by v_j from v_i at time t is therefore

$$-\beta_i^- [x_i(t - \tau_{ij}^-) - \Gamma_{ij}^-]^+ p_{ij}^-, \tag{14}$$

where, for example, τ_{ij}^- is a positive time lag, and p_{ij}^- is a nonnegative inhibitory path weight. The minus sign preceding β_i^- guarantees that an increase in $x_i(t - \tau_{ij}^-)$ can only cause a decrease in $x_j(t)$. The total inhibitory signal received by v_i from all states v_m at time t is

$$-J_i(t) = -\sum_{m=1}^{n} \beta_{mi}^- [x_m(t - \tau_{mi}^-) - \Gamma_{mi}^-]^+ p_{mi}^-. \tag{15}$$

Embedding fields in which the expressions (12) and (14) occur are called signed fields because (12) and (14) can be pictured as flows over a *signed graph;* (12) can be pictured as in Fig. 2a, or generically as in Fig. 2b, and (14) can be pictured as in Fig. 3a, or generically as in Fig. 3b. The plus and minus signs stand for excitatory and inhibitory, respectively.

B. Signed Equilibrium and Decay

The preceding subsection merely localizes various expressions and admits inhibitory signals. The introduction of inhibitory signals requires

that we rescale the equilibrium values of $x_i(t)$ and $z_{jk}(t)$. This we now do.

Consider the decay term $-\alpha x_i(t)$ of (2). This term implies that $x_i(t)$ decay to zero if all inputs to v_i are identically zero; that is, zero is the *equilibrium value* of $x_i(t)$. A zero equilibrium value is needed in (2) because all variables are nonnegative. Thus an increase in the excitatory signals to v_i makes $x_i(t)$ more positive, and an elimination of all inputs permits $x_i(t)$ to decay to its zero equilibrium value.

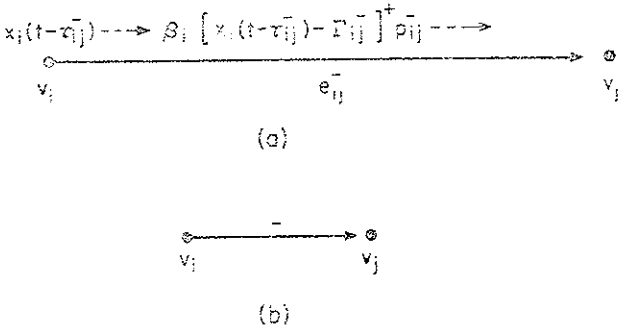

FIG. 3

In a signed field, both excitatory and inhibitory inputs are admissible. No longer need the variables $x_i(t)$ and $z_{jk}(t)$ be nonnegative. The zero equilibrium value thus loses its privileged place, and can be replaced by any equilibrium value P_i that $x_i(t)$ approaches, from either above or below, when all inputs are zero. Thus we generalize the decay term $-\alpha x_i(t)$ to

$$\alpha_i^+ [P_i - x_i(t)]^+ - \alpha_i^- [x_i(t) - P_i]^+, \tag{16}$$

which describes an exponential decay to equilibrium in which the decay rates from above and below P_i need not be equal. In the special case $\alpha_i^+ = \alpha_i^- \equiv \alpha$, (16) reduces to

$$\alpha(P_i - x_i(t)).$$

If, moreover, $P_i = 0$, then we recapture the original decay term $-\alpha x_i(t)$.

The condition that the $\Gamma_{mi}^{\pm}$ be positive thresholds is now generalized to

$$\Gamma_{mi}^{\pm} > P_m. \tag{17}$$

Having abandoned a zero equilibrium value for $x_i(t)$, there no longer exists any reason to require a zero equilibrium value for $z_{jk}(t)$. We need

only measure whether $z_{jk}(t)$ is above or below various prescribed equilibrium and threshold values. Once this is realized, Eq. (3)

$$\dot{z}_{jk}(t) = -\alpha z_{jk}(t) + \beta p_{jk} x_j(t - \tau_{jk}) x_k(t)$$

of (*) is easily adapted to the signed case.

If we let Q_{jk} be the new equilibrium value of $z_{jk}(t)$, the decay term $-\alpha z_{jk}(t)$ of (3) is generalized to

$$\alpha_{jk}^+[Q_{jk} - z_{jk}(t)]^+ - \alpha_{jk}^-[z_{jk}(t) - Q_{jk}] . \tag{18}$$

The term $\beta p_{jk} x_j(t - \tau_{jk})$ in (3) designates the excitatory signal received by N_{jk} from v_j at time t, which is replaced by (11) in the signed case. The term $x_k(t)$ in (3) need not be compared to its zero equilibrium in the signed case. It suffices to measure whether or not $x_k(t)$ exceeds some threshold value Λ_k^+. The term $x_k(t)$ in (3) can therefore be replaced by

$$[x_k(t) - \Lambda_k^+]^+ \tag{19}$$

in the signed case. In all, (3) is replaced by

$$\begin{aligned} \dot{z}_{jk}(t) = {} & \alpha_{jk}^+[Q_{jk} - z_{jk}(t)]^+ - \alpha_{jk}^-[z_{jk}(t) - Q_{jk}]^+ \\ & + \beta_j' \gamma_{jk} p_{jk}[x_j(t - \tau_{jk}) - \Gamma_j^+]^+ [x_k(t) - \Lambda_k^+] . \end{aligned} \tag{20}$$

The numerical factor γ_{jk} in (20) is merely a scaling factor. The only immediately obvious constraint to be placed on Λ_k^+ is

$$\Lambda_k^+ \geqslant P_k , \tag{21}$$

which means heuristically that $z_{jk}(t)$ does not grow if $x_k(t)$ is "too small." The choice $\alpha_{jk}^+ = \alpha_{jk}^- = 0$ of parameters in (20) is not forbidden.

It remains only to rescale $z_{ij}(t)$ in (12); $z_{ij}(t)$ is replaced by $[z_{ij}(t) - \Omega_{ij}^+]^+$, where Ω_{jk}^+ satisfies

$$\Omega_{jk}^+ \geqslant Q_{jk} . \tag{22}$$

Then (12) becomes

$$\beta_i'[x_i(t - \tau_{ij}^+) - \Gamma_i^+]^+ p_{ij}^+ [z_{ij}(t) - \Omega_{ij}^+]^+ , \tag{23}$$

and (13) becomes

$$J_i^+(t) = \sum_{m=1}^{n} \beta_m'[x_m(t - \tau_{mi}^+) - \Gamma_{mi}^+]^+ p_{mi}^+ [z_{mi}(t) - \Omega_{mi}^+]^+ . \tag{24}$$

The equation for $x_i(t)$ is therefore

$$\begin{aligned} \dot{x}_i(t) = {} & \alpha_i^+[P_i - x_i(t)]^+ - \alpha_i^-[x_i(t) - P_i]^+ \\ & + J_i^+(t) - J_i^-(t) + I_i^+(t) - I_i^-(t) \end{aligned} \tag{25}$$

where $I_i^+(t)$ and $I_i^-(t)$ are known excitatory and inhibitory inputs, respectively; $J_i^+(t)$ is defined in (24); and $J_i^-(t)$ is defined in (15). Equations (20)

and (25) arise from (*) merely by (a) admitting inhibitory signals, and (b) introducing the more general equilibria and thresholds required by (a).

12. A NEURAL INTERPRETATION

Reference [1] contains a qualitative neural interpretation of excitatory embedding fields. We now extend this interpretation to signed fields.

A. Cell Body, Axon, End Bulb, and Synapse

Inspection of Figs. 2 and 3 readily suggests a neural interpretation of these networks. Each point v_i corresponds to a nerve cell body, or to a cluster of mutually "indistinguishable" nerve cell bodies. Each edge e_{ij} corresponds to the axons, or clusters of axons, leading from v_i to v_j. Each arrowhead N_{ij} corresponds to the end bulbs associated with the axons of e_{ij}. And the spaces between the "end bulbs" N_{ij} and the "postsynaptic cell bodies" v_j are synapses.

B. Membrane Potential, Spiking Frequency, and Threshold

The process $x_i(t)$ sits in the cell bodies v_i. It is therefore plausible to identify it with the *average membrane potential* of these cells; $x_i(t)$ gives rise to an excitatory signal along the axons e_{ij} that equals

$$\beta_i'[x_i(t) - \Gamma_i^+]^+ p_{ij}^+ . \tag{26}$$

This signal cannot represent an individual neural event, because $x_i(t)$ is itself an average. It can represent the *frequency* of individual neural signals in e_{ij} throughout a unit time interval, if such signals exist. Indeed neural signals exist, and an individual neural signal is called a *spike*. Expression (26) is thus the *average spiking frequency* created by the cells v_i in the axons e_{ij} at time t. This signal equals zero whenever $x_i(t) \leqslant \Gamma_i^+$. Thus Γ_i^+ is the *spiking threshold*, if such exists, which it does, and P_i is the *equilibrium potential* of $x_i(t)$, if such exists, which it does. Given these identifications, (26) claims that the spiking frequency equals zero until the membrane potential rises above a spiking threshold that strictly exceeds the equilibrium potential. For suprathreshold values, average spiking frequency should be linearly related to membrane potential. (In the bounded case to be discussed in Section 14, "saturation" occurs at large values of membrane potential, and thus also in the spiking frequency.)

These claims are a natural consequence of an obvious neural labeling of our psychologically derived mathematical variables. It is therefore gratifying that some neural data behave like the claims. In particular, the

existence of a process $x_i(t)$ fluctuating at v_i is well known [2, 3, 4]. That this process gives rise to signals in e_{ij} is also well known [4]. That a threshold exists above which these signals vary linearly in $x_i(t)$ has also been reported [5, 16, 17, 18]. Our equations are also compatible with the so-called slow potential view, since the fluctuations in the average membrane potential v_i determine changes in spiking frequency along e_{ij} that activate the potentials of the recipient cells v_j [19, 20, 21], as the next paragraphs show.

C. Transmitter Production and Release

The excitatory signal (26) reaches the end bulbs N_{ij} at time $t + \tau_{ij}^+$ and thereupon activates the process $z_{ij}(t + \tau_{ij}^+)$. A quantity

$$p_{ij}^+[x_i(t) - \Gamma_{ij}^+]^+ p_{ij}^-[z_{ij}(t + \tau_{ij}^+) - \Omega_{ij}]^+ \tag{27}$$

is then released into the "synaptic clefts" between N_{ij} and the postsynaptic cell bodies v_j, and thereupon causes a proportional change in the rate of change, namely, $\dot{x}_j(t + \tau_{ij}^+)$, of the "average postsynaptic potential." A transmission from end bulbs over the synapses to postsynaptic cells is experimentally well established [3, 22, 23]. We therefore identify (27) with the amount of *excitatory transmitter* released into the synaptic clefts at time $t + \tau_{ij}^+$. The process $z_{ij}(t + \tau_{ij}^+)$ taking place in the end bulbs N_{ij} corresponds to the *transmitter production and storage* process in these end bulbs at time $t + \tau_{ij}^+$. The transmission law (27) then says that the average amount of excitatory transmitter released from N_{ij} into the synaptic clefts increases if either the presynaptic spiking frequency or the amount of available transmitter in N_{ij} increases.

D. Presynaptic and Postsynaptic Control of Transmitter Production

Equation (20) can now be read as an equation describing the average rate of transmitter production in N_{jk}. The most important term in (20) is

$$\beta_j^+ \gamma_{jk}^+ p_{jk}^+[x_j(t - \tau_{jk}^+) - \Gamma_{jk}^+]^+[x_k(t) - \Lambda_k^-]^+. \tag{28}$$

As this expression increases, the rate of excitatory transmitter production also increases; (28) can be increased either by increasing the spiking frequency (11)

$$\beta_j^+ p_{jk}^-[x_j(t - \tau_{jk}^+) - \Gamma_{jk}^-]^+$$

of e_{jk}, or by increasing the postsynaptic potential $x_k(t)$ of v_k, and thereby increasing (19)

$$[x_k(t) - \Lambda_k^-]^+.$$

Both (11) and (19) must be positive in order for $z_{jk}(t)$ to grow. This property leads to the prediction that excitatory transmitter production in N_{jk} depends on both the presynaptic spiking frequency in e_{jk} and the postsynaptic potential of v_k. In particular, if the value of $x_k(t)$ is depressed below its equilibrium potential P_k by an inhibitory input to v_k, then since $\Lambda_k^- \geqslant P_k$, $z_{jk}(t)$ cannot grow. This is, moreover, just the kind of inhibition that the transformation $z_{jk}(t) \to y_{jk}(t)$ in (**) describes.

E. Exponential Decay of Membrane Potential

Expression (16) states that the average membrane potential decays at an exponential rate to its equilibrium potential in the absence of incoming transmissions from other cells. This fact has been experimentally reported [2].

13. THE HARTLINE–RATLIFF EQUATION

In a series of distinguished papers (e.g., [5, 24–26]), Hartline, Ratliff, and their colleagues have developed an empirical equation concerning the inhibitory interactions within the *Limulus* retina. This equation has the following form.

$$r_i = e_i - \sum_{j=1}^{n} K_{ij}[r_j - r_{ij}^0]^+ \tag{29}$$

where $i = 1, 2, \ldots, n$; e_i is the frequency of impulses within the ith ommatidium axon under a fixed light source in the absence of inhibition from nearby cells. When inhibitory contributions are not negligible, e_i is reduced to r_i, which is the net frequency of discharge of impulses in the ith axon due to the combined effects of the light source and the inhibition from nearby cells. The K_{ij} are "inhibitory coefficients," and the r_{ij}^0 are "threshold frequencies." We now show that the Hartline–Ratliff equation can be formally derived from (25). We carry out the derivation in the simplest possible way to clearly delineate the main ideas. In particular, various geometrical complexities of the *Limulus* retinal network that can, in principle, be discussed using (25) will be ignored.

All interactions in (29) are inhibitory. Thus we set all $p_{ki}^+ = 0$, ignoring the possibility that excitatory links activate the inhibition indirectly. Only the light sources perturb the retina, and so all $I_i^-(t) \equiv 0$. Equation (25) becomes

$$\dot{x}_i(t) = \alpha_i^-[P_i - x_i(t)]^- - \alpha_i^-[x_i(t) - P_i]^+ - J_i^-(t) + I_i^+(t). \tag{30}$$

The light source is stationary in time. That is,

$$I_i^-(t) = I_i^- = \text{constant}.$$

Equation (29) describes the retina's steady-state response to this source. In other words,

$$\dot{x}_i(t) = 0 \quad \text{and} \quad x_i(t) = x_i = \text{constant},$$

for all i. Two cases now arise.

Case 1. Only one ommatidium, say v_i, receives light. Therefore v_i receives no inhibition from other v_j, and since x_i is excited above its equilibrium value by the light,

$$\alpha_i^-[P_i - x_i(t)]^- = 0 \quad \text{and} \quad \alpha_i^-[x_i(t) - P_i]^- = \alpha_i^-(x_i(t) - P_i).$$

Equation (30) becomes

$$0 = -\alpha_i^-(x_i - P_i) + I_i^-,$$

or

$$x_i = P_i + \frac{1}{\alpha_i^-} I_i^-. \tag{31}$$

The spiking frequency along the ommatidium axon is, as in (26), of the form

$$e_i = \mu_i^+[x_i - \Gamma_i^+]^-, \tag{32}$$

where μ_i^+ plays the role of a composite coefficient $\beta_i^+ p_{ij}^-$, and $\Gamma_i^+ > P_i$. By (31),

$$e_i = \mu_i^+\left[P_i - \Gamma_i^+ + \frac{1}{\alpha_i^-} I_i^-\right]^+,$$

and choosing the light I_i^- sufficiently intense to make e_i suprathreshold,

$$e_i = \mu_i^-\left(P_i - \Gamma_i^+ + \frac{1}{\alpha_i^-} I_i^-\right). \tag{33}$$

Case 2. Let several ommatidia receive intense light. Then not all $J_i^- = 0$. Since all $x_m(t)$ are constant, (15) becomes

$$J_i^- = \sum_{m=1}^{n} \beta_m^-[x_m - \Gamma_{mi}^-]^+ p_{mi}^-,$$

and thus for any index i such that $x_i \geqslant P_i$, (25) becomes

$$0 = -\alpha_i^-(x_i - P_i) + I_i^+ - \sum_{m=1}^{n} \beta_m^-[x_m - \Gamma_{mi}^-]^+ p_{mi}^-.$$

Rearranging terms yields

$$x_i = P_i + \frac{1}{\alpha_i^-} I_i^+ - \frac{1}{\alpha_i^-} \sum_{m=1}^{n} \beta_m^-[x_m - \Gamma_{mi}^-]^+ p_{mi}^-. \tag{34}$$

Denote the spiking frequency along the ith ommatidium by r_i in this case. Then

$$r_i = \mu_i^+[x_i - \Gamma_i^+]^+, \tag{35}$$

and so by (34),

$$r_i = \mu_i^-\left[P_i - \Gamma_i^+ + \frac{1}{\alpha_i^-} I_i^+ - \frac{1}{\alpha_i^-} \sum_{m=1}^{n} \beta_m^-[x_m - \Gamma_{mi}^-]^+ p_{mi}^-\right]^+.$$

Consider a value of i for which r_i is suprathreshold (and in particular $x_i \geqslant P_i$). Then

$$r_i = \mu_i^-\left(P_i - \Gamma_i^+ + \frac{1}{\alpha_i^-} I_i^- - \frac{1}{\alpha_i^-} \sum_{m=1}^{n} \beta_m^-[x_m - \Gamma_{mi}^-]^+ p_{mi}^-\right),$$

which by (33) is

$$r_i = e_i - \sum_{m=1}^{n} [x_m - \Gamma_{mi}^-]^+ \frac{\mu_i^- \beta_m^- p_{mi}^-}{\alpha_i^-}. \tag{36}$$

Now suppose for simplicity that *all* r_i that receive light are suprathreshold. We can without loss of generality suppose for our formal convenience that all r_i are positive. Then by (35),

$$x_i = \Gamma_i^+ + \frac{1}{\mu_i^+} r_i,$$

which when substituted into (36) yields

$$r_i = e_i - \sum_{m=1}^{n} \left[\frac{1}{\mu_m^+} r_m + \Gamma_m^+ - \Gamma_{mi}^-\right]^+ \frac{\mu_i^- \beta_m^- p_{mi}^-}{\alpha_i^-},$$

which is the same as

$$r_i = e_i - \sum_{m=1}^{n} \left[r_m - \mu_m^+(\Gamma_{mi}^- - \Gamma_m^+)\right]^+ \frac{\mu_i^- \beta_m^- p_{mi}^-}{\mu_m^+ \alpha_i^-}; \tag{37}$$

(37) agrees formally with the Hartline–Ratliff equation (29) if we make the identifications

$$K_{ij} = \frac{\mu_i^+ \beta_j^-}{\mu_j^- \alpha_i^-} p_{ji}^- \tag{38}$$

and

$$r_{ij}^0 = \mu_j^-(\Gamma_{ji}^- - \Gamma_j^+). \tag{39}$$

Although the geometry of the *Limulus* retina is presumably far more complicated than our assumptions on the coefficients p_{ij}^+ and p_{ij}^-, the fact that (25) can be derived from simple psychological ideas and has a formal structure in (37) that agrees with the Hartline–Ratliff equation—which itself resulted from much arduous and ingenious physiological experimentation—acts as a helpful and gratifying check on the psychological approach used to derive (25).

Let us now suppose that our assumptions on I_{ij}^{+} and p_{ij}^{+} are not so unreasonable as to entirely invalidate the formulas (38) and (39): (38) then shows that the inhibitory coefficients K_{ij} are composites of the rate parameters μ_i^+, μ_j^-, β_j^-, and α_i^-, and of the inhibitory path weight p_{ji}^-. If the shape and size of the retinal cells (say, at the retina's border, or due to its elliptical symmetry) vary systematically with retinal position, then the rate parameters need not be independent of i and j. In this case, measurements of K_{ij} might well give a systematically distorted view of the path weights p_{ji}^-.

A curious fact emerges if we interpret r_{ij}^0 as a threshold spiking frequency, as (29) bids. Then $r_{ij}^0 \geqslant 0$ and (39) implies

$$\Gamma_{ji}^- \geqslant \Gamma_j^+. \tag{40}$$

This inequality is plausible, for example, if the ommatidium axon must start firing before inhibitory interactions can set in.

14. BOUNDED SIGNED EMBEDDING FIELDS

The system of (20) and (25) has the property that the values $x_i(t)$ can become arbitrarily large if the inputs $I_i^-(t)$ are chosen sufficiently large. It is in some ways physically more plausible to assume that $x_i(t)$ has fixed bounds that cannot be exceeded under any circumstances, much as we find in (2′) and (3′). We therefore derive below the natural adaptations of (2′) and (3′) to the signed case.

Suppose that $x_i(t)$ has a fixed maximum M_i and a fixed minimum m_i. The excitatory term due to excitatory signals $J_i^+(t)$ from states v_m and to other excitatory sources $I_i^+(t)$ can be written down by direct analogy with (2′) in the form

$$\alpha_i^+(M_i - x_i(t))(J_i^+(t) + I_i^+(t)), \tag{41}$$

with initial data that satisfy $x_i \leqslant M_i$. The inhibitory term due to inhibitory signals $J_i^-(t)$ from states v_m and to other inhibitory sources $I_i^-(t)$ can be written down by analogy with (41) in the form

$$-\alpha_i^-(x_i(t) - m_i)(J_i^-(t) + I_i^-(t)), \tag{42}$$

with initial data subject to the constraint $x_i \geqslant m_i$. It remains only to determine the decay term for $x_i(t)$. One possibility is to simply use the decay term of (16). Another possibility arises by noting in (41) and (42) that $x_i(t)$ is compared with its extrema M_i and m_i, rather than with its

equilibrium value P_i. We therefore single out for particular attention the following equation for $x_i(t)$.

$$\begin{aligned}\dot{x}_i(t) = {} & \alpha_i^+(M_i - x_i(t))(\gamma_i^+ + J_i^+(t) + I_i^+(t)) \\ & -\alpha_i^-(x_i(t) - m_i)(\gamma_i^- + J_i^-(t) + I_i^-(t)),\end{aligned} \tag{43}$$

with decay term

$$\alpha_i^+\gamma_i^+(M_i - x_i(t)) - \alpha_i^-\gamma_i^-(x_i(t) - m_i), \tag{44}$$

and initial data subject to the constraints $m_i \leqslant x_i \leqslant M_i$. The inequalities $m_i \leqslant x_i \leqslant M_i$ are then satisfied for all time, since $x_i(t) = m_i$ implies $\dot{x}_i(t) \geqslant 0$, and $x_i(t) = M_i$ implies $\dot{x}_i(t) \leqslant 0$.

The equilibrium P_i of $x_i(t)$ in (43) is defined by the value of x_i that satisfies (43) when $\dot{x}_i(t)$ and all inputs $J_i^+(t)$, $I_i^+(t)$, $J_i^-(t)$, and $I_i^-(t)$ are zero. Under these circumstances, (43) becomes

$$0 = \alpha_i^+\gamma_i^+(M_i - P_i) - \alpha_i^-\gamma_i^-(P_i - m_i),$$

or

$$P_i = \frac{\alpha_i^+\gamma_i^+ M_i + \alpha_i^-\gamma_i^- m_i}{\alpha_i^+\gamma_i^+ + \alpha_i^-\gamma_i^-}. \tag{45}$$

A similar argument holds if $z_{jk}(t)$ has the fixed maximum M_{jk}. Then the term (28)

$$\beta_j^+\gamma_{jk}^+ p_{jk}^+[x_j(t - \tau_{jk}^+) - \Gamma_{jk}^+]^+[x_k(t) - \Lambda_k^+]^+$$

in (20) is premultiplied by $(M_{jk} - z_{jk}(t))$ and the initial data of z_{jk} are constrained by the inequality $z_{jk} \leqslant M_{jk}$. If $z_{jk}(t)$ also has a fixed minimum m_{jk}, then the initial data satisfy $z_{jk} \geqslant m_{jk}$, and the decay term of (20) either remains unchanged or is replaced by an expression of the form

$$u_{jk}^+(M_{jk} - z_{jk}(t)) - u_{jk}^-(z_{jk}(t) - m_{jk})$$

by analogy with (44). In the latter case,

$$\begin{aligned}\dot{z}_{jk}(t) = {} & -u_{jk}^-(z_{jk}(t) - m_{jk}) + (M_{jk} - z_{jk}(t)) \\ & \times (u_{jk}^+ + \beta_j^+\gamma_{jk}^+ p_{jk}^+[x_j(t - \tau_{jk}^+) - \Gamma_{jk}^+]^+[x_k(t) - \Lambda_k^+]^+).\end{aligned} \tag{46}$$

The possibility that $u_{jk}^+ = u_{jk}^- = 0$ is not ruled out.

15. INFORMATION AND LATERAL INHIBITION

This section introduces a signed embedding field $\mathcal{M}$ that illustrates the connection between information and lateral inhibition that was conjectured in Section 3. For simplicity we consider a signed field that is a direct analog of an outstar [9]. The smallest (excitatory) outstar is pictured in

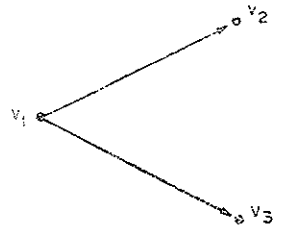

FIG. 4

Fig. 4. Section 3 shows that the transformation $z_{1i} \to y_{1i}$ in an excitatory outstar can be replaced by mutual inhibitory signals between the states v_2 and v_3 in the outstar's border. We therefore introduce the *signed outstar* $\mathcal{M}$ of Fig. 5. By analogy with the case of excitatory outstars, we call v_1 the *excitatory source*, $V^+ = \{v_2, v_3\}$ the *excitatory border*, and $V^- = \{v_4, v_5\}$ the *inhibitory border* of $\mathcal{M}$. The inhibitory signal from v_2 to v_3 is created in the following way. v_2 sends an excitatory signal to v_4, and v_4 thereupon sends an inhibitory signal to v_3. Similarly, v_3 inhibits v_2 by sending an excitatory signal to v_5, which thereupon sends an inhibitory signal to v_2. Just as in an excitatory outstar, only the set of points $V_O = \{v_1, v_2, v_3\}$ receives inputs from an experimentalist E, and only these points send outputs back to E: that is,

$$I_i^+(t) = I_i^-(t) \equiv 0, \qquad i = 4, 5.$$

Set V_O is therefore called the set of *observable states of* $\mathcal{M}$. The set $V_C = \{v_4, v_5\}$ merely mediates inhibitory signals between the observable

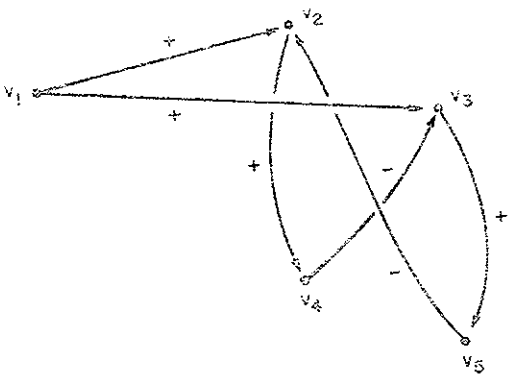

FIG. 5

states and is therefore called the set of *unobservable states* of $\mathcal{M}$. We also suppose that all inputs from E to $\mathcal{M}$ are excitatory, by analogy with the case of excitatory outstars; that is,

$$I_i^-(t) \equiv 0, \qquad i = 1, 2, 3.$$

The parameters of $\mathcal{M}$ are chosen in a "homogeneous" way. By this we mean two things. First, we think of the points in V_O and V_C as belonging to (possibly) two distinct "cell types." Mathematically speaking, this means that all "local" parameters of points in V_O and (separately) of points in V_C are the same (e.g., $\beta_1^+ = \beta_2^+ = \beta_3^+$ and $\beta_4^- = \beta_5^-$). Second, we impose a "cylindrical symmetry" on $\mathcal{M}$ by requiring that all parameters remain unchanged when points in V^+ are interchanged and then corresponding points in V^- are interchanged (e.g., $p_{12}^+ = p_{13}^+$ and $p_{42}^- = p_{52}^-$). This condition says that the geometry of $\mathcal{M}$ does not bias the learning of either v_1v_2 or v_1v_3. In all, the following constraints on parameters are imposed.

a. Rate constants

$$\alpha_i^+ = \alpha_1, \quad \alpha_i^- = \alpha_1^-, \quad \beta_i^+ = \beta_1, \quad \beta_i^- = \beta_1^-, \qquad i = 1, 2, 3,$$

and

$$\alpha_4^- = \alpha_5^-, \quad \alpha_4^+ = \alpha_5^+, \quad \beta_4^+ = \beta_5^+, \quad \beta_4^- = \beta_5^-.$$

b. Interaction coefficients

$$p_{12}^+ = p_{13}^+, \quad p_{24}^+ = p_{35}^+, \quad p_{43}^- = p_{52}^-$$

and

$$\gamma_{12}^+ = \gamma_{13}^+, \quad \gamma_{24}^+ = \gamma_{35}^+.$$

All other coefficients p_{ij}^+ and p_{ij}^- are zero. It therefore suffices to specify the following thresholds, time lags, and equilibrium constants.

c. Thresholds

$$\Gamma_{12}^+ = \Gamma_{13}^+, \quad \Gamma_{24}^+ = \Gamma_{35}^+, \quad \Gamma_{43}^- = \Gamma_{52}^-;$$
$$\Omega_{12}^+ = \Omega_{13}^+, \quad \Omega_{24}^+ = \Omega_{35}^+;$$

and

$$\Lambda_2^- = \Lambda_3^-, \quad \Lambda_4^- = \Lambda_5^-.$$

d. Time lags

$$\tau_{12}^+ = \tau_{13}^+, \quad \tau_{24}^+ = \tau_{35}^+, \quad \tau_{43}^- = \tau_{52}^-.$$

e. Equilibrium constants

$$P_1 = P_2 = P_3, \quad P_4 = P_5$$

and

$$Q_{12} = Q_{13}, \quad Q_{24} = Q_{35}.$$

To conveniently write out the (unbounded) equations for $\mathcal{M}$, we let

$$D_{jk}(t) = \alpha_j^-[P_j - x_j(t)]^+ - \alpha_j^-[x_j(t) - P_j]^+$$

and

$$D_{jklm}(t) = u_{jk}^-[Q_{jk} - z_{lm}(t)]^+ - u_{jk}^-[z_{lm}(t) - Q_{jk}]^+.$$

Then

$$\dot{x}_1(t) = D_{111}(t) + I_1^-(t), \tag{47}$$

$$\begin{aligned}\dot{x}_i(t) = {} & D_{1i}(t) + I_i^-(t) \\ & + \beta_1^- \gamma_{12}^- p_{12}^-[x_1(t - \tau_{12}^-) - \Gamma_{12}^-]^+[z_{1i}(t) - \Omega_{12}^-] \\ & - \beta_1^- \gamma_{13}^- p_{13}^-[x_{i+2}(t - \tau_{13}^-) - \Gamma_{13}^-]^+, \qquad i = 2, 3.\end{aligned} \tag{48}$$

$$\dot{x}_i(t) = D_{1i}(t) + \beta_2^- p_{23}^-[x_{i-2}(t - \tau_{23}^-) - \Gamma_{23}^-]^+[z_{i-2,i}(t) - \Omega_{23}^-]^+, \qquad i = 4, 5;$$

$$\dot{z}_{1i}(t) = D_{1211}(t) + \beta_1^- \gamma_{12}^- p_{12}^-[x_1(t - \tau_{12}^-) - \Gamma_{12}^-]^+[x_i(t) - \Lambda_2^-], \qquad i = 2, 3; \tag{49}$$

and

$$\dot{z}_{i\ 3}(t) = D_{21i,i-3}(t) + \beta_2^- \gamma_{23}^- p_{23}^-[x_i(t - \tau_{23}^-) - \Gamma_{23}^-]^+[x_{i\ 3}(t) - \Lambda_3^-], \qquad i = 2, 3. \tag{50}$$

The outputs from V_Q to E have the form

$$O_i(t) = \mu_E^-[x_i(t) - \Gamma_E^-]^+, \qquad i = 1, 2, 3. \tag{51}$$

We now show how the inhibitory interactions recapture some properties of the transformations $z_{ij}(t) \to y_{ij}(t)$ and $x_j(t) \to O_j^{(\Gamma)}(t)$, and improve upon these properties.

A. Unique Dominant $x_2(t)$

Suppose in the excitatory outstar of Fig. 4 that $x_2(t) \gg x_3(t) \geqq 0$. Then the modified outputs have the form

$$O_2^{(\Gamma)}(t) \cong [x_2(t) - \Gamma_2^-]^+$$

and $O_3^{(\Gamma)}(t) \cong 0$ [9]. The corresponding statement in $\mathcal{M}$ is that $O_2(t)$ is uninfluenced by inhibition from v_3, whereas $O_3(t)$ is inhibited to zero by v_2. This is readily seen to happen by considering the following thought experiment.

Let $\mathcal{M}$ be in equilibrium until time $t = 0$. That is, $x_j(t) = P_j$ and $z_{jk}(t) = Q_{jk}$ for all $t \leqslant 0$. Suppose that only the input $I_2^-(t)$ to v_2 is ever positive thereafter. It is then obvious by the homogeneous choice of parameters in (47)–(50) that only v_2 ever sends out signals to other points. The signal from v_2 to v_4 is excitatory, but its only effect is to create an inhibitory signal from v_4 to v_3. This signal drives $x_3(t)$ below its equilibrium

value, and in particular below the threshold $\Gamma_E^- > P_3$. Thus by (51), $O_3(t) \equiv 0$, as expected. Moreover, since $\Gamma_{34}^- > P_3$, v_3 never sends an inhibitory signal to v_2 via v_5, so that $O_3(t)$ is indeed never influenced by inhibition from v_3.

Notice that these conclusions do not depend on the numerical values of the parameters, but only on their "homogeneity."

B. Equal Border Values

Given an excitatory outstar with a common border value $x_2(t) = x_3(t)$, no matter how large, then $O_2^{(\Gamma)}(t) = O_3^{(\Gamma)}(t) \equiv 0$. This fact is translated in a signed outstar as follows: equal reciprocal inhibition between v_2 and v_3 keeps $x_2(t)$ and $x_3(t)$ so small that $O_2(t) = O_3(t) \equiv 0$. The following thought experiment shows how this inhibition takes place.

Let $\mathcal{M}$ be in equilibrium until time $t = 0$, and suppose that equal inputs $I_2^-(t)$ and $I_3^-(t)$ are received by v_2 and v_3 thereafter. By homogeneity, $x_2(t) \equiv x_3(t)$ for all $t \geqslant 0$, and in particular $O_2(t) \equiv O_3(t)$. We wish to choose the parameters of $\mathcal{M}$ to guarantee that $O_2(t) \equiv 0$, which by (51) is the same as

$$x_2(t) \leqslant \Gamma_E^-, \qquad t \geqslant 0. \tag{52}$$

If $I_2^-(t)$ has very small values, this is easily done. By "very small," we mean the following. Consider the thought experiment of Section 15, A, in which only $I_2^-(t)$ is positive: $I_2^-(t)$ is very small if Γ_E^- is larger than the maximum of $x_2(t)$ in that experiment. The only nontrivial case arises when $I_2^-(t)$ is sufficiently large to drive $x_2(t)$ above the fixed threshold Γ_E^- at some time. Suppose that such an $I_2^-(t)$ occurs in the present experiment, along with an equal $I_3^-(t)$. Inequality (52) can then be achieved for all $t \geqslant 0$ only if inhibitory signals between v_2 and v_3 take effect before the inputs drive $x_2(t)$ above Γ_E^-. Various constraints must be imposed on the parameters of $\mathcal{M}$ to guarantee that this occur. In order to conveniently discuss these constraints, let

T_1 = minimum time needed for $I_2^-(t)$ to drive $x_2(t)$ above Γ_E^- if $I_3^-(t) \equiv 0$,

T_2 = minimum time needed for $I_2^-(t)$ to create an excitatory signal in e_{21}^-,

and

T_3 = minimum time needed for the excitatory signal at v_4 from e_{24}^- to drive $x_4(t)$ above Γ_{43}^-.

An inhibitory signal from v_3 will arrive at v_2 before $x_2(t)$ exceeds Γ_E^- if and only if

$$T_2 + \tau_{21}^- + T_3 + \tau_{43}^- < T_1. \tag{53}$$

Since the expression

$$T_4 \equiv T_2 + \tau_{24}^+ + T_3 + \tau_{43}^+$$

is the total time needed for a signal to be created in v_3 *and* to reach v_2, (53) implies

$$\Gamma_{21}^+ < \Gamma_E^+. \tag{54}$$

Otherwise, no signal could leave v_3 until $x_3(t)$ reached Γ_E^+, by homogeneity between v_2 and v_3.

Inequality (53) cannot be guaranteed for an arbitrary input $I_2^+(t)$. By (54), it suffices to show that there exists an $I_2^+(t)$ such that

$$\tau_{24}^+ + \tau_{43}^+ \geqslant T_1, \tag{55}$$

given fixed parameters for $\mathcal{M}$. This is readily done by letting $I_2^+(t)$ equal a constant I_2^+ for all $t \geqslant 0$, and then choosing this constant so large that (55) is satisfied. The following properties of (48) show that this can always be done.

Before an inhibitory signal reaches v_2 from v_1, (48) implies

$$\dot{x}_2(t) \geqslant D_{12}(t) + I_2^+, \tag{56}$$

since

$$p_{12}^+ p_{12}^- [x_1(t - \tau_{12}^+) - \Gamma_{12}^-]^+ [z_{12}(t) - \Omega_{12}^-]^+$$

is nonnegative, and

$$p_{13}^+ p_{13}^- [x_1(t - \tau_{13}^+) - \Gamma_{13}^-]^+$$

equals zero. Since $\mathcal{M}$ starts out in equilibrium and no inhibition has reached v_2 in (56), (56) implies

$$\dot{x}_2(t) \geqslant \alpha_1^-(P_1 - x_2(t)) + I_2^+,$$

which readily yields

$$x_2(t) \geqslant X_2(t)$$

where

$$X_2(t) = P_1 + \frac{I_2^+}{\alpha_1^-}[1 - \exp(-\alpha_1^+ t)].$$

If I_2^+ is chosen greater than $\alpha_1^-(\Gamma_E^+ - P_1)$, then $X_2(t)$ achieves the value Γ_E^+ at time

$$T_5 \equiv -\frac{1}{\alpha_1^-}\log\left[1 - \frac{\alpha_1^-(\Gamma_E^+ - P_1)}{I_2^+}\right],$$

and $X_2(t) > \Gamma_E^+$ for all $t > T_5$. Since $\lim_{I_2^+ \to \infty} T_5 = 0$, $x_2(t)$ can be made to exceed any fixed threshold Γ_E^+ in an arbitrarily short time simply by increasing I_2^+. In fact, it suffices to choose any I_2^+ such that

$$I_2^+ > \alpha_1^-(\Gamma_E^+ - P_1)$$

and

$$\tau_{24}^+ + \tau_{43}^+ \geqslant -\frac{1}{\alpha_1^-}\log\left[1 - \frac{\alpha_1^-(\Gamma_E^+ - P_1)}{I_2^+}\right]$$

to violate (53).

A necessary condition for (53) to hold is thus that there exists a fixed finite number $N_2^{(0)}$ such that

$$I_2^+(t) \leqslant N_2^{(0)} < \infty, \tag{57}$$

for all $t \geqslant 0$ and all inputs I_2^+ that ever perturb v_2. Inequality (57) is always fulfilled if $I_2^+(t)$ is a signal created by a point whose state function has a finite maximum, as in Section 14.

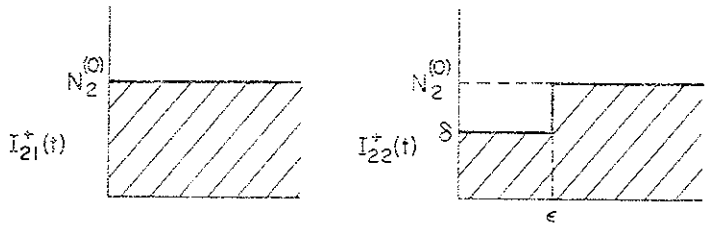

FIG. 6

It can also readily be seen that (57) must be supplemented by a condition of the form

$$\frac{d}{dt} I_2^+(t) \leqslant N_2^{(1)} < \infty \tag{58}$$

for all $t \geqslant 0$ and all inputs I_2^+ that ever perturb v_2. Figure 6 illustrates the need for (58); it describes two different inputs $I_{21}^+(t)$ and $I_{22}^+(t)$ delivered to v_2 in two identical copies $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, of the signed outstar $\mathcal{M}$. The input $I_{21}^+(t)$ to $\mathcal{M}_1$ is a rectangular pulse taking on the value $N_2^{(0)}$ for all $t \geqslant 0$, and thus satisfies (57). The parameters in $\mathcal{M}_1$ are chosen to keep $x_2(t) \leqslant \Gamma_E^+$ for all $t \geqslant 0$, but in such a way that $T_{11}/T_{11} \cong 1$. That is, the inhibitory signals arrive just before $x_2(t)$ reaches Γ_E^+ and just barely manage to keep $x_2(t)$ below Γ_E^+.

$\mathcal{M}_2$ has the same parameters as $\mathcal{M}_1$, but the inhibitory signal created in $\mathcal{M}_1$ at time T_{22} is smaller than the signal created in $\mathcal{M}_2$ at time T_{21} because $\delta < N_1^{(0)}$. This inhibitory signal can be made as small as we please by choosing δ so that $x_2(t) \cong \Gamma_{24}^+$ for $t \cong T_{22}$, as (49) shows. ϵ can be chosen in such a way that the inhibitory signal arrives when $I_{22}^+(t) = N_2^{(0)}$, and thus, by (48), when $x_2(t)$ is growing rapidly. The inhibitory signal is

often too weak to overcome the large excitatory signal, and thus the inequality $x_2(t) \leqslant \Gamma_{12}^+$ is violated for some $t \geqslant 0$. This difficulty arises because the small inhibitory signal created by v_2 at a given time is not large enough to overcome the large excitatory input that exists at v_3 when it arrives after transit through e_{21}^+ and e_{13}^-. Inequality (58) overcomes this difficulty by guaranteeing that the early values of $I_2^-(t)$, which create inhibitory signals, are not too much smaller than the later values of $I_2^+(t)$ with which the inhibitory signals compete; (58) holds whenever $I_2^+(t)$ is the output signal from a bounded state.

For any particular choice of bounds $N_2^{(0)}$ and $N_2^{(1)}$ on admissible inputs, our task is to find parameters of $\mathcal{M}$ that create and deliver inhibitory signals of sufficient strength and with sufficient rapidity to overcome the effects of prolonged excitation. As $N_1^{(0)}$ and $N_2^{(0)}$ are allowed to increase, the parameters T_2, τ_{21}^-, T_3, and τ_{13}^- must be chosen smaller. T_2 and T_3 can be decreased, for example, by decreasing the differences $\Gamma_{21}^+ - P_1$ and $\Gamma_{13}^+ - P_1$ between thresholds and equilibrium values. To guarantee that a sufficiently strong inhibitory signal arrives at v_2 from v_1, we can *amplify* the signal as much as we please by choosing arbitrarily large values of β_1^-. This extra degree of freedom in amplifying inhibitory signals is a major advantage of using two cell types V_O and V_C.

C. *Learning*

We now show that large values at the states need not cause any changes in the associations unless these values represent a learning experiment.

For example, let $\mathcal{M}$ be in equilibrium until time $t = 0$. Let any admissible input $I_1^+(t)$ occur at v_1 (i.e., A is said to $\mathcal{M}$). Equal excitatory signals are sent to v_2 and v_3, but since z_{12} and z_{13} begin in equilibrium, the signals reaching v_2 and v_3 are zero, by (48) and (49), and z_{12} and z_{13} never leave equilibrium. In short, saying A teaches us no list AB or AC, so the associations do not change.

More generally, let $z_{12}(t) = z_{13}(t) > \Omega_{13}^+$ until time $t = 0$. Let A occur once again. Then equal and *positive* signals reach v_2 and v_3 from v_1, but these signals create inhibitory signals between v_2 and v_3 that keep $x_2(t)$ and $x_3(t)$ small. In fact, by choosing A_2^+ sufficiently large, we can guarantee, as in the previous section, that $x_2(t)$ and $x_3(t)$ never exceed A_2^+. Then

$$[x_2(t) - A_2^+]^+ = [x_3(t) - A_2^+]^+ \equiv 0,$$

and by (49),

$$\dot{z}_{12}(t) = \dot{z}_{13}(t) \leqslant 0$$

for all $t \geqslant 0$. Even a large signal to A alone need not make the associations grow.

These examples show, parethetically, that by choosing sufficiently large initial $z_{12}(0)$ and $z_{13}(0)$ values, even a small input to v_1 can create signals at v_2 and v_3 that violate (57) and (58) for fixed $N_1^{(0)}$ and $N_2^{(0)}$. This difficulty need never arise if $z_{12}(t)$ and $z_{13}(t)$ obey bounded equations.

Now let $\mathcal{M}$ be in equilibrium until time $t = 0$ and present only the letter B to $\mathcal{M}$; that is, only $I_2^+(t)$ is ever positive. Then only $x_2(t)$ ever exceeds a threshold, and in particular,

$$[x_1(t) - \Gamma_{12}^+]^+ = 0, \qquad t \geqslant 0;$$

$z_{12}(t)$ and $z_{13}(t)$ once again remain at equilibrium for all $t \geqslant 0$.

Finally let A be presented to $\mathcal{M}$ at time t_A, and let B be presented to $\mathcal{M}$ at time t_B, $t_B > t_A$. That is, inputs occur at v_1 and v_2 with a time separation of $t_B - t_A$. If t_B is so much greater than t_A that the signal created by $I_1^+(t)$ from v_1 to v_2 traverses e_{12}^+ and decays before $I_2^+(t)$ becomes positive, then once again no learning occurs, as we see by a straightforward application of the preceding remarks. If, however, this signal arrives at N_{12}^+ as $I_2^+(t)$ becomes positive, then both

$$[x_1(t - \tau_{12}^+) - \Gamma_{12}^+]^+ \quad \text{and} \quad [x_2(t) - A_2^+]^+$$

will be positive, and so $z_{12}(t)$, and *only* $z_{12}(t)$, grows.

16. LEARNING DECREASES REACTION TIME

Daily life amply illustrates that familiar behavior sequences can often be emitted more rapidly than unfamiliar ones. This also happens in our machines, as we illustrate in the simplest possible case of Fig. 7. Let the equations of this machine $\mathcal{M}$ be unbounded, for simplicity. Then

$$\dot{x}_1(t) = D_{11}(t) + I_1^+(t), \tag{59}$$

$$\dot{x}_2(t) = D_{22}(t) + I_2^+(t) + \beta_1^+ p_{12}^+ [x_1(t - \tau_{12}^+) - \Gamma_{12}^+]^+ [z_{12}(t) - \Omega_{12}^+]^+, \tag{60}$$

$$\dot{z}_{12}(t) = D_{1212}(t) + \beta_1^+ \gamma_{12}^+ p_{12}^+ [x_1(t - \tau_{12}^+) - \Gamma_{12}^+]^+ [x_2(t) - A_2^+]^+ \tag{61}$$

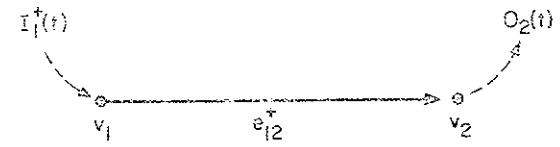

FIG. 7

and

$$O_2(t) = \mu_E^-[x_2(t) - \Gamma_E^+]^+. \tag{62}$$

Let v_1 and v_2 be in equilibrium until time $t = 0$; that is, $x_i(t) = P_i$, $t \leqslant 0$, $i = 1, 2$. Suppose that $z_{12}(0) > \Omega_{12}^+$, so that signals from v_1 reach v_2. Let a positive input $I_1^-(t)$ perturb v_1 at time $t = 0$, and let $I_2^+(t) \equiv 0$. The *reaction time* $\tau \equiv \tau(I_1^+)$ of $\mathcal{M}$ to the input I_1^+ is defined as the total time that elapses from the onset of I_1^- at time $t = 0$ until the output $O_2(t)$ from v_2 becomes positive for the first time, which by (62) is

$$\tau = \inf\{t: x_2(t) > \Gamma_E^-\}.$$

τ is the sum of three factors:

$$\tau = T_1 + \tau_{12}^- + T_2$$

where T_1 = minimum time needed for I_1^- to drive $x_1(t)$ above the threshold Γ_{12}^+ and to thereby create a signal in e_{12}^-; τ_{12}^- = time required for the signal to traverse e_{12}^-; and T_2 = minimum time needed for the signal from e_{12}^- to v_2 to drive $x_2(t)$ above the threshold Γ_E^- after the signal reaches v_2 at time $T_1 + \tau_{12}^-$. That is,

$$T_1 = \inf\{t: x_1(t) > \Gamma_{12}^+\}$$

and

$$T_2 = \inf\{t - T_1 - \tau_{12}^-: x_2(t) > \Gamma_E^-\}.$$

Whereas τ_{12}^- is constant, T_1 is a functional of I_1^+, and T_2 is a functional of I_1^- and z_{12}. It often suffices to approximate $z_{12}(t)$ by its initial value $z_{12}(0)$, since T_2 depends on the values of the slowly fluctuating $z_{12}(t)$ only in the time interval $[0, T_1 + \tau_{12}^- + T_2]$. We therefore suppose for simplicity that $z_{12}(t) \cong z_{12}(0)$ in all of the following remarks.

The qualitative behavior of T_1 and T_2 is easily found in the special case that $I_1^-(t) = I_1^+ =$ constant, $t \geqslant 0$. Then (59) readily shows that T_1 is a monotone decreasing function of I_1^+, since $x_1(t)$ is a monotone increasing function of I_1^+ for all $t \geqslant 0$. The signal from v_1 to v_2 is therefore also monotone increasing in I_1^+, and thus, by (60), T_2 is a monotone decreasing function of I_1^-. In all, both T_1 and T_2 are monotone decreasing functions of I_1^+, so that also the reaction time τ decreases as the input (or "energy") increases.

To study the effects of learning on reaction time, we need merely consider several copies $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ of the machine $\mathcal{M}$, each with identical initial data and input $I_1^-(t) = I_1^+$, and differing only in that $\mathcal{M}_{i+1}$ has a larger $z_{12}(0)$ value than $\mathcal{M}_i$ has. That is, $\mathcal{M}_{i+1}$ knows the

transition from r_1 to r_2 better than $\mathcal{M}_i$ does. Each of the machines $\mathcal{M}_i$ has the same $x_1(t)$ function, since they all have the same initial data and inputs. In particular, T_1 is the same in all the machines. But the *signal* from v_1 to v_2 is larger in $\mathcal{M}_{i+1}$ than in $\mathcal{M}_i$, as (60) shows, because $z_{12}(0)$ is larger in $\mathcal{M}_{i+1}$ than in $\mathcal{M}_i$. Thus, $x_2(t)$ grows faster in $\mathcal{M}_{i+1}$ than in $\mathcal{M}_i$, so that T_2 is smaller in $\mathcal{M}_{i+1}$ than in $\mathcal{M}_i$. Since $\tau = T_1 + \tau_{12}^- + T_2$, τ is also smaller in $\mathcal{M}_{i+1}$ than in $\mathcal{M}_i$. That is, learning decreases the reaction time.

The dependence of τ on $I_1^-(t)$ and $z_{12}(t)$ is easily found in the special case that $I_1^-(t) = I =$ constant, $z_{12}(t) \cong z_{12}(0)$, and $\alpha_1^- = \alpha_2^- = \alpha$. Then τ is the smallest root of the equation

$$x_2(t) = \Gamma_E^- \tag{63}$$

subjected to the constraint

$$\dot{x}_2(t) > 0. \tag{64}$$

Since

$$T_1 = \begin{cases} \infty & \text{if } I \leqslant \alpha(\Gamma_{12}^+ - P_1) \\ \dfrac{1}{\alpha}\log\left[\dfrac{I}{I - \alpha(\Gamma_{12}^+ - P_1)}\right] & \text{if } I > \alpha(\Gamma_{12}^+ - P_1) \end{cases}$$

$\tau = \infty$ unless $I > \alpha(\Gamma_{12}^+ - P_1)$. If $I > \alpha(\Gamma_{12}^+ - P_1)$, then by (59) and (60), (63) implies the equation

$$x + A = Be^{\alpha x}$$

where

$$x = \tau - \tau_{12}^-,$$

$$A = \frac{1}{\alpha}\left\{1 - \log\left[\frac{I}{I - \alpha(\Gamma_{12}^+ - P_1)}\right]\right\},$$

and

$$B = \frac{1}{\alpha} + \frac{1}{I}\left[P_1 - \Gamma_{12}^+ - \frac{\alpha(\Gamma_E^- - P_2)}{\beta_1^- p_{12}^-(z_{12}(0) - \Omega_{12})}\right].$$

(64) implies

$$\tau > \tau_{12}^- - \frac{1}{\alpha}\log\left\{1 + \frac{\alpha}{I}\left[P_1 - \Gamma_{12}^+ + \frac{\alpha(P_2 - \Gamma_E^-)}{\beta_1^- p_{12}^-(z_{12}(0) - \Omega_{12})}\right]\right\}.$$

These equations for the reaction time of $\mathcal{M}$ in terms of I and $z_{12}(0)$ must, of course, be modified in more realistic situations, where many components of the type illustrated in Fig. 7 mutually excite and inhibit one another before a peripheral input can give rise to a peripheral output. The next section describes some of the additional possibilities that arise as a result of inhibitory interactions between points.

17. SPATIOTEMPORAL MASKING

The speeding up of reaction time due to prior learning helps to eliminate response interference, to spontaneously improve the memory of prior learning, and to create context effects.

Consider, for example, the homogeneous signed outstar $\mathcal{M}$ of Fig. 5. Suppose that the sequence r_1r_2 has been substantially better learned than r_1r_3 before time $t = 0$; that is, $z_{12}(0) \gg z_{13}(0)$. Let all $x_i(t) = P_i$ for $t \in [-\tau_{12}^-, 0]$, and suppose that a recall trial occurs at time $t = 0$; that is, $I_1^+(t)$ becomes positive in an interval of the form $(0, \lambda_1)$, and all other inputs are identically zero. To avoid trivialities, suppose $I_1^+(t)$ is sufficiently large to guarantee that x_1 and x_2 eventually exceed the thresholds Γ_{12}^+ and Γ_{21}^+, respectively.

By the homogeneity of $\mathcal{M}$, $I_1^+(t)$ creates equal signals from v_1 along the excitatory edges e_{12}^+ and e_{13}^+ to the arrowheads N_{12} and N_{13}. Since $z_{12}(0) \gg z_{13}(0)$, the input from N_{12} to v_2 created by this signal is substantially larger than the input from N_{13} to v_3. Therefore, $x_2(t)$ grows at a faster rate than $x_3(t)$ does, and reaches larger values. In particular, $x_2(t)$ generates an inhibitory signal from v_2 to v_3 before $x_3(t)$ can generate an inhibitory signal from v_3 to v_2. If $z_{12}(0)$ sufficiently exceeds $z_{13}(0)$, then this inhibitory signal from v_2 can reach v_3 in force before $x_3(t)$ generates an inhibitory signal to v_2, and can keep $x_3(t)$ below the inhibitory threshold Γ_{33}^-. Under these circumstances, $O_3(t) \equiv 0$ since the inhibitory threshold Γ_{33}^- at v_3 is smaller than the output threshold Γ_3^+.

In short, speeding up the reaction time of well-learned behavior sequences *inhibits* the output from lesser learned sequences to subthreshold values.

The same argument shows that for $z_{12}(0)/z_{13}(0)$ sufficiently large, $x_3(t)$ can be kept below Λ_3^-, and thus $\dot{z}_{13}(t) \leqslant 0$ for all $t \geqslant 0$, whereas there exist times t for which

$$[x_1(t - \tau_{12}^+) - \Gamma_{12}^+]^+[x_2(t) - \Lambda_2^-] > 0$$

during which $z_{12}(t)$ grows.

In short, speeding up the reaction time of well-learned behavior sequences tends to preserve $\mathcal{M}$'s memory of these sequences.

These arguments do not require that $z_{13}(0)$ be at equilibrium, but only that $z_{12}(0) \gg z_{13}(0)$. In other words, some memory of an r_1r_3 transition can exist within $\mathcal{M}$, but this memory *never gives rise to an observable output signal* $O_3(t)$ if $z_{12}(0) \gg z_{13}(0)$, because $x_2(t)$ grows faster than

$x_3(t)$ grows, and can therefore inhibit $x_3(t)$ to small values before $O_3(t)$ becomes positive. Since a $z_{13}(t)$ memory exists, but never appears in $\mathcal{M}$'s overt behavior, we say that the inhibitory signal from v_2 to v_3 *spatially masks* the $z_{13}(0)$ memory by cutting off the output from v_3. Since this masking depends on the relative timing of the input signals received by v_2 and v_3, it is really more proper to call the masking process *spatiotemporal masking*.

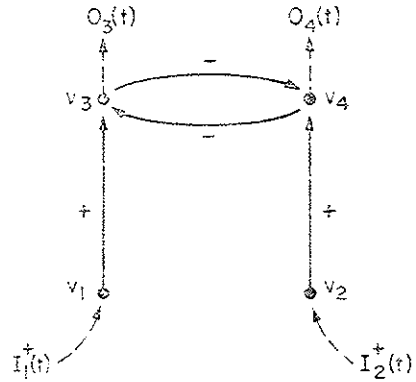

FIG. 8

The foregoing example of spatiotemporal masking depends on unequal $z_{ij}(0)$ values. Section 16 suggests the existence of another form of masking due to unequal inputs $I_i^+(t)$. Consider, for example, the machine $\mathcal{M}$ of Fig. 8. Choose the parameters *and* the initial data of $\mathcal{M}$ "homogeneously"; for example, $\alpha_1^+ = \alpha_2^+$, $p_{12}^+ = p_{24}^+$, and $p_{34}^+ = p_{43}^+$. In particular, let $x_1(t) = x_2(t) = P_1$, $t \leqslant 0$. For simplicity, let the inputs $I_1^+(t)$ and $I_2^+(t)$ have the graphs given in Fig. 9. That is, $I_1^+(t)$ is a rectangular

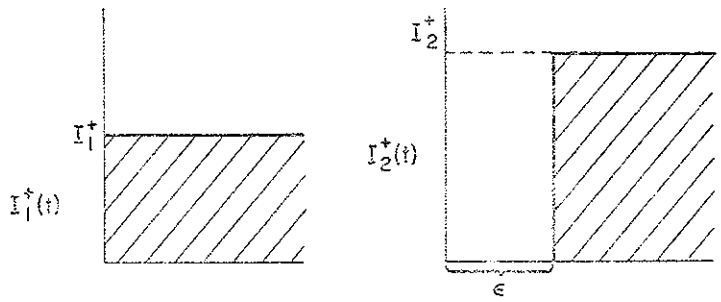

FIG. 9

input pulse with onset time $t = 0$, and $I_2^-(t)$ is a rectangular input pulse with onset time $t = \epsilon > 0$.

Although $I_1^+(t)$ perturbs $\mathcal{M}$ before $I_2^-(t)$ does, we can often find a time lag ϵ and an I_2^- that is sufficiently larger than I_1^+ that the output $O_3(t) = 0$ for all $t \geqslant 0$, whereas $O_4(t)$ becomes positive for some t. That is, the input $I_2^-(t)$ masks the *earlier* input $I_1^+(t)$! The reasoning behind this conclusion is the same as in the case that $z_{12}(0) \gg z_{13}(0)$.

Since

$$\dot{x}_i(t) = D_{i1}(t) + I_i^+(t), \qquad i = 1, 2,$$

with $x_i(t) = P_1$ for all $t \leqslant 0$, we find

$$x_i(t) = P_1 + \exp(-\alpha_1^- t) \int_0^t \exp(\alpha_1^- v) I_i^+(v)\, dv, \qquad i = 1, 2.$$

In particular,

$$x_1(t) = P_1 + \frac{I_1^+}{\alpha_1^-}[1 - \exp(-\alpha_1^- t)], \qquad t \geqslant 0,$$

whereas

$$x_2(t) = \begin{cases} P_1, & 0 \leqslant t \leqslant \epsilon, \\ P_1 + \dfrac{I_2^+}{\alpha_1^-}\{1 - \exp[-\alpha_1^-(t - \epsilon)]\}, & \epsilon \leqslant t. \end{cases}$$

Letting $T_i^+ =$ minimum time needed for $x_i(t)$ to generate an excitatory signal in $e_{i,i+2}^+$, $i = 1, 2$, we find that

$$T_1^+ = \begin{cases} \infty & \text{if } I_1^+ \leqslant \alpha_1^-(\Gamma_{12}^+ - P_1) \\ \dfrac{1}{\alpha_1^-} \log\left[\dfrac{I_1^+}{I_1^+ - \alpha_1^-(\Gamma_{12}^+ - P_1)}\right] & \text{if } I_1^+ > \alpha_1^-(\Gamma_{12}^+ - P_1) \end{cases} \tag{65}$$

and

$$T_2^+ = \begin{cases} \infty & \text{if } I_2^+ \leqslant \alpha_1^-(\Gamma_{12}^+ - P_1) \\ \epsilon + \dfrac{1}{\alpha_1^-} \log\left[\dfrac{I_2^+}{I_2^+ - \alpha_1^-(\Gamma_{12}^+ - P_1)}\right] & \text{if } I_2^+ > \alpha_1^-(\Gamma_{12}^+ - P_1). \end{cases} \tag{66}$$

x_2 creates a signal in e_{24}^+ before x_1 can create a signal in e_{13}^+ if and only if

$$T_{11}^+ > T_{12}^+,$$

which by (65) and (66) is equivalent to

$$\frac{I_1^+[I_2^+ - \alpha_1^-(\Gamma_{12}^+ - P_1)]}{I_2^+[I_1^+ - \alpha_1^-(\Gamma_{12}^+ - P_1)]} > \exp(\epsilon\alpha_1^-), \tag{67}$$

given also that $I_1^+ > \alpha_1^-(\Gamma_{12}^+ - P_1)$. Although v_1 receives an input before v_2 does, v_2 gives rise to a signal before v_1 does if I_2^- is so much larger than I_1^- that (67) is satisfied.

If (67) is satisfied, then $x_3(t)$ begins growing before $x_3(t)$ does, and v_4 sends out an inhibitory signal to v_3 before v_3 sends an inhibitory signal to v_4. If I_2^- is sufficiently large, the inhibitory signal from v_4 to v_3 can keep v_3 below the thresholds Γ_{31}^+ and Γ_{34}^+. Consequently, $O_3(t) \equiv 0$, and $x_4(t)$, which receives no inhibition from v_3, creates a positive $O_4(t)$ signal for some $t > 0$. Again spatiotemporal masking has occurred.

Exciting the *single* point v_2 a great deal can also create a masking effect, which we also call a "warm-up," "practice," or "context" effect. This happens because exciting v_2 more than v_1 creates the inequality $z_{24}(t) > z_{13}(t)$. If equal inputs are then presented to v_1 and v_2, equal outputs $O_3(t)$ and $O_4(t)$ will *not* be created, because the inhibitory signal from v_4 to v_3 will be larger than the inhibitory signal from v_3 to v_4.

The two ways of achieving masking in a homogeneous machine, namely, through inhomogeneous choices of inputs or of associations, can be mixed in more complicated machines to achieve some remarkably subtle effects. Masking can also be guaranteed by constructing machines whose *geometry* is not homogeneous.

18. REMEMBRANCES OF EVENTS LONG SUPPRESSED

We now briefly discuss an important case of spatiotemporal masking due to both the choice of inputs and of associations. Memories of experiences (say) from childhood that never come into consciousness in our customary adult environment can be triggered suddenly and with remarkable clarity by a fragment of our childhood environment. How can memories be stored with such clarity over such long time intervals without ever influencing our overt behavior or our conscious thoughts in a new environment? Spatiotemporal masking suggests a way, simply by pointing out that a learned transition whose output is inhibited by one distribution of inputs can nonetheless be strongly facilitated by a different distribution of inputs.

We will consider a very simple and highly idealized machine to illustrate this fact. Consider the machine $\mathcal{M}$ of Fig. 10, which differs from the signed outstar of Fig. 9 only by the addition of a point v_6, an excitatory edge e_{63}^-, and an end bulb N_{63}^- impinging on v_3. We can apply all of the remarks concerning learned spatiotemporal masking from Section 17 to

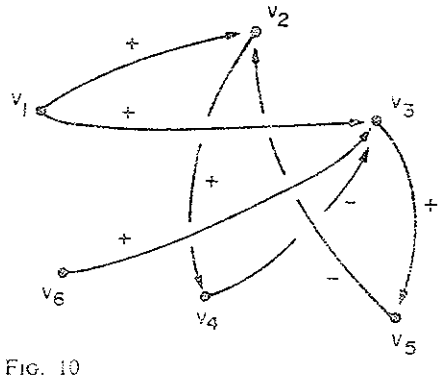

FIG. 10

the signed outstar part of $\mathcal{M}$, which we denote by $\mathcal{M}^*$. The "new environment" of $\mathcal{M}$ consists of inputs only to $\mathcal{M}^*$, whereas the "old environment" also permits inputs to v_6. Observable learned transitions within $\mathcal{M}$ can occur along e_{12}^-, e_{13}^+, and e_{63}^-. The inputs that activate these transitions are therefore concentrated at v_1 and v_6. In the new environment, only inputs to v_1 can activate a transition, and since $z_{12} \gg z_{13}$, $x_3(t)$ is always inhibited by $x_2(t)$ in the new environment, and only the output from $O_2(t)$ ever becomes positive. In the old environment, inputs to both v_1 and v_6 can become positive. In particular, if I_1^+ and I_2^+ are simultaneously large, then even though the input from v_1 to v_2 exceeds the input from v_1 to v_3 because $z_{12} \gg z_{13}$, the *total* input from v_1 and v_6 to v_3 can easily exceed the input received by v_2 from v_1. Therefore, $x_3(t)$ can inhibit $x_2(t)$, and we can easily guarantee that only the output from v_3 becomes positive. Thus the transition along e_{13}^-, which is not strong enough to create any output whatsoever in the new environment, can nonetheless overcome the stronger transition along e_{12}^- in the old environment with the help of the transition along e_{63}^-. Once this qualitative point is clearly understood, the reader can easily construct for himself more complicated and realistic instances of spatiotemporal masking due to a mixture of learned transitions and a fluctuating input environment.

19. REDUCING HIGHER-ORDER ASSOCIATIONS TO SIMPLE ASSOCIATIONS

Section 10 suggested that a signed embedding field can compute products $\prod_{k\in J} x_k(t)$ of states even for large sets J of indices without introducing higher-order associations. We now introduce the simplest signed field that can (approximately) accomplish this.

Consider the machine $\mathcal{M}$ of Fig. 11. We will show that $\mathcal{M}$ can be constructed so that the simple correlation $z_{34}(t + \tau_{13}^- + \tau_{34}^+)$ computes (approximately) the product $x_1(t)x_2(t)$.

For reasons that will soon be apparent, let $\mathcal{M}$ obey *bounded* equations. For simplicity, let the parameters of $\mathcal{M}$ be invariant under permutation of v_1 and v_2: for example, let $\tau_{13}^- = \tau_{23}^-$, $\Gamma_{13}^+ = \Gamma_{23}^+$, and $P_1 = P_2$. Also

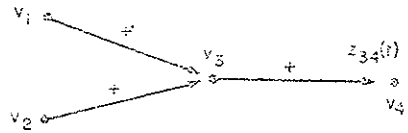

FIG. 11

suppose that the only inputs reaching $\mathcal{M}$ are $I_1(t)$, $I_2^-(t)$, and $I_4(t)$. Then $\mathcal{M}$ obeys the equations

$$\dot{x}_i(t) = \alpha_i^-(M_i - x_i(t))(\gamma_i^- + I_i^-(t)) - \alpha_i^-\gamma_i^-(x_i(t) - m_i), \qquad i = 1, 2. \tag{68}$$

$$\dot{x}_3(t) = \alpha_3^-(M_3 - x_3(t))\{\gamma_3^- + \beta_1^+ p_{13}^+[x_1(t - \tau_{13}^+) - \Gamma_{13}^+]^+[z_{13}(t) - \Omega_{13}^+]^+ + \beta_1^+ p_{13}^-[x_2(t - \tau_{13}^+) - \Gamma_{13}^+]^+[z_{23}(t) - \Omega_{13}^+]^+\} - \alpha_3^-\gamma_3^-(x_3(t) - m_3), \tag{69}$$

$$\dot{x}_4(t) = \alpha_4^-(M_4 - x_4(t))\{\gamma_4^- + I_4^-(t) + \beta_3^+ p_{34}^-[x_3(t - \tau_{34}^+) - \Gamma_{34}^-]^+[z_{34}(t) - \Omega_{34}^-]^+\} - \alpha_4^-\gamma_4^-(x_4(t) - m_4), \tag{70}$$

$$\dot{z}_{i3}(t) = (M_{i3} - z_{i3}(t))\{u_{i3}^- + \beta_1^+\gamma_{i3}^- p_{i3}^-[x_i(t - \tau_{i3}^-) - \Gamma_{i3}^-]^+[x_3(t) - \Lambda_3^-]^+\} - u_{i3}^-(z_{i3}(t) - m_{i3}), \qquad i = 1, 2, \tag{71}$$

and

$$\dot{z}_{34}(t) = (M_{34} - z_{34}(t))\{u_{34}^+ + \beta_3^+\gamma_{34}^+ p_{34}^+[x_3(t - \tau_{34}^+) - \Gamma_{34}^-]^+[x_4(t) - \Lambda_4^+]^+\} - u_{34}^-(z_{34}(t) - m_{34}). \tag{72}$$

Let the initial data of $\mathcal{M}$ be invariant under permutations of v_1 and v_2, and suppose that equal inputs $I_1^-(t)$ and $I_2^-(t)$ excite v_1 and v_2 for all $t \geqslant 0$. Then $x_1(t) \equiv x_2(t)$ and $z_{13}(t) \equiv z_{23}(t)$ for $t \geqslant 0$, by symmetry. In particular, the total input received by v_3 from v_1 and v_2 at time t, namely,

$$\beta_1^+ p_{13}^+[x_1(t - \tau_{13}^+) - \Gamma_{13}^+]^+[z_{13}(t) - \Omega_{13}^+]^+ + [x_2(t - \tau_{13}^+) - \Gamma_{13}^+]^+[z_{23}(t) - \Omega_{13}^+]^+),$$

equals

$$2\beta_1^+ p_{13}^+[x_1(t - \tau_{13}^+) - \Gamma_{13}^+]^+[z_{13}(t) - \Omega_{13}^+]^+. \tag{73}$$

Since the *maximum* input received by v_3 from either v_1 or v_2 is

$$K_{13} \equiv p_{13}^{+}(M_1 - \Gamma_{13}^{-})(M_{13} - \Omega_{13}^{-}). \tag{74}$$

(73) implies that the *maximum* input received by v_3 from v_1 and v_2 is $2K_{13}$.

To guarantee that z_{31} responds only to the *joint* excitation of v_1 and v_2, choose the threshold Γ_{34}^{-} controlling the excitatory signal from v_3 to v_4 sufficiently large that a rectangular input to v_3 of size K_{13} cannot drive $x_3(t)$ above Γ_{31}^{+}, but sufficiently small that a rectangular input of size $2K_{13}$ can drive $x_3(t)$ above Γ_{34}^{-}. Any choice of Γ_{34}^{-} such that

$$\frac{z_{13}^{-}M_3(\gamma_{13}^{-} + K_{13}) + z_{23}^{-}\gamma_{23}^{-}m_3}{z_{13}^{-}\gamma_{13}^{-} + z_{23}^{-}\gamma_{23}^{-} + K_{13}} \leqslant \Gamma_{31}^{+}$$

and

$$\Gamma_{34}^{-} < \frac{z_{13}^{-}M_3(\gamma_{13}^{-} + 2K_{13}) + z_{23}^{-}\gamma_{23}^{-}m_3}{z_{13}^{-}\gamma_{13}^{-} + z_{23}^{-}\gamma_{23}^{-} + 2K_{13}}$$

accomplishes this goal, as (69) readily shows. For such a Γ_{34}^{-}, v_3 does not send a signal to v_4 unless both v_1 and v_2 send signals to v_3. By (72), $z_{31}(t)$ cannot grow unless both $x_1(\xi)$ and $x_2(\xi)$ are large at approximately time $\xi = t - \tau_{13}^{-} - \tau_{31}^{-} - T_1 - T_3$ where T_1 is the minimum time needed for $x_1(t)$ and $x_2(t)$ to exceed the threshold Γ_{13}^{-} and T_3 is the minimum time needed for $x_3(t)$ to exceed the threshold Γ_{31}^{-} after the signals from N_{13} and from N_{23} reach v_3. It is in this sense that $z_{31}(t)$ computes the product $x_1(t - \tau_{13}^{-} - \tau_{34}^{-} - T_1 - T_3)x_2(t - \tau_{13}^{-} - \tau_{31}^{-} - T_1 - T_3)$.

In order that the argument above hold, $z_{13}(t - \tau_{13}^{-} - T_1)$ and $z_{23}(t - \tau_{13}^{-} - T_1)$ must also have large values due to frequent prior excitation of v_1 and v_2. A machine in which this warm-up effect does not occur can be constructed simply by eliminating the $[z_{i3} - \Omega_{13}^{-}]^{+}$ terms from (69), $i = 1, 2$. The foregoing argument also holds in an *unbounded* signed field whose inputs are subjected to (57), just so as z_{13} and z_{23} are not too large.

"Products" of the three or more components can be computed by a simple extension of the idea given above. Computing $x_1(t)x_2(t)x_3(t)$ requires only a bounded signed field $\mathcal{M}$ of the form pictured in Fig. 12, where we have chosen the time lags to satisfy $\tau_{13}^{-} = \tau_{23}^{-}$ and $\tau_{45}^{+} = \tau_{13}^{+} + \tau_{35}^{+}$. The thresholds Γ_{35}^{+} and Γ_{56}^{+} are chosen such that v_3 sends a signal to v_5 only if both v_1 and v_2 send signals to v_3, and v_5 sends a signal to v_6 only if both v_3 and v_4 send a signal to v_5. Thus v_5 sends a signal to v_6 only if *all* the values $x_1(t)$, $x_2(t)$, and $x_3(t)$ are large at a common prior time. The size of $z_{56}(t)$ therefore measures the frequency with which all these values have been large in the past.

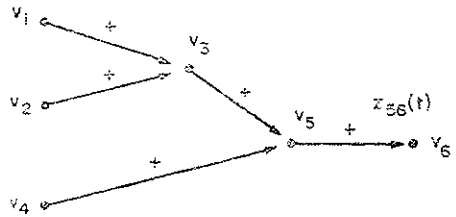

FIG. 12

When a list (say, $r_1r_2r_4$) is presented to a machine $\mathcal{M}$ with time lag w between successive symbols, $x_1(t)$ becomes large w time units before $x_2(t)$ does, and $x_2(t)$ becomes large w time units before $x_3(t)$ does, so that it is desirable that $\mathcal{M}$ compute the product

$$x_1(t)x_2(t + w)x_3(t + 2w)$$

rather than $x_1(t)x_2(t)x_3(t)$. This can be done by simply choosing the time lags τ_{ij}^{+} of $\mathcal{M}$ differently. For example, it suffices to let

$$\tfrac{1}{2}\tau_{13}^{-} = \tau_{23}^{-} = \tau_{35}^{+} = \tau_{45}^{+}$$

and to choose $w \geqq \tau_{23}^{-}$, as the reader can easily check.

The idea of summing quantities and then truncating them by a threshold is not new, although it arises in our machines as a consequence of more fundamental considerations rather than as an end in itself. The introduction of correlations that measure these operations and control the size of future signals using these measurements is new, however, and opens up qualitatively new possibilities for the use of this old idea.

20. THE EXCITATORY–INHIBITORY DUALISM

Daily experience abounds in examples of a dualism between positive versus negative, or excitatory versus inhibitory, factors. This dualism is mathematically represented in our number system by such classes as positive versus negative numbers; in the physical description of nature in such examples as particles versus antiparticles; in our philosophies as such concepts as the yang versus the yin; and in our ethics in such values as good versus evil. A special case of an excitatory versus inhibitory dualism is also visible in the excitatory versus inhibitory interactions of a signed embedding field. We will treat this dualism as a fundamental principle for deepening our understanding of signed fields, rather than merely as an amusing formal consequence of our previous *gedanken*

experiments. This point of view seems inevitable if we ever wish to bring the concepts stated herein into harmony with the results of other disciplines.

Invoking excitatory–inhibitory dualism (or EID) in the present context has far-reaching concrete consequences as well as philosophical appeal. It will, for example, lead us by simple formal manipulations to equations that can, in a natural way, be interpreted to imply

(a) the existence of inhibitory as well as excitatory transmitter substances,

(b) the existence of two quantities that are formal caricatures of Na^+ and K^+ concentrations inside the cell membrane,

(c) the creation by excitatory transmitters of an inward flow of Na^+ that induces an outward flow of K^+ at suprathreshold values, whereas only an outward flow of K^+ is created by an inhibitory transmitter,

(d) the existence of two quantities that are formal caricatures of Ca^{2+} and Mg^{2+} concentrations inside the cell membrane,

(e) the binding of Na^+, K^+, Ca^{2+}, Mg^{2+} within the end bulb in complexes of varying strength to produce transmitter production and release rates that are sensitive to prior presynaptic and postsynaptic levels of membrane potential, and related facts and predictions.

Some qualitative insight also emerges concerning such fundamental problems as the way in which a nerve cell's functions in learning determines its shape, and the way in which a nerve cell "knows" how much it must produce to meet extracellular demands upon it. This article merely introduces some of the formal machinery needed to derive these results and insights. Later papers will investigate these and related topics in greater detail.

21. THE COUPLING OF FORMAL Na^+ AND K^+ TO FORMAL EXCITATORY AND INHIBITORY TRANSMITTERS

This section uses EID to show how the simplest features of (a)–(c) in the preceding section can be derived. A *bounded* signed embedding field, as in Section 14, will always be considered for definiteness.

Our derivation begins with the observation that a "sin of omission" occurred when we passed from (8) to (9) by adjoining + superscripts to the parameters. Should not + superscripts be adjoined to the *processes* x_i and z_{ij} as well? The answer surely is yes, since x_i and z_{ij} must *grow* in order to create a signal from v_i to v_j, and this property imparts to x_i and

z_{ij} an "excitatory polarity." We therefore seek an equation for the bounded variable x_i^+.

We find this equation, at least formally, in (43) if we merely adjoin an extra superscript + to all expressions therein. Then (43) becomes

$$\dot{x}_i^+(t) = \alpha_i^{+\,+}(M_i^+ - x_i^+(t))(\gamma_i^{+\,+} + J_i^{+\,+}(t) + I_i^{+\,+}(t)) - \alpha_i^{+\,-}(x_i^+(t) - m_i^+)(\gamma_i^{+\,-} + J_i^{+\,-}(t) + I_i^{+\,-}(t)). \tag{75}$$

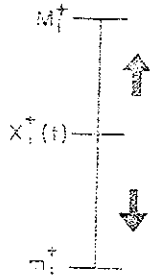

FIG. 13

It remains only to determine the input functions $J_i^{+\,+}(t)$ and $J_i^{+\,-}(t)$: $J_i^{+\,+}(t)$ is readily found by adjoining sufficiently many + superscripts to $J_i^+(t)$ in (24). Then

$$J_i^{+\,+}(t) = \sum_{m=1}^{n} \beta_m^{+\,+}[x_m^{+\,+}(t - \tau_{mi}^{+\,+}) - \Gamma_{mi}^{+\,+}]^+ p_{mi}^{+\,+}[z_{mi}^{+\,+}(t - \Omega_{mi}^{+\,+})]^+. \tag{76}$$

Input $J_i^{+\,-}(t)$ is determined from $J_i^{+\,+}(t)$ by invoking EID. This we do in the most heuristic way possible to keep the physical meaning of the results clear.

Compare the excitatory part

$$\alpha_i^{+\,+}\gamma_i^{+\,+}(M_i^+ - x_i^+(t)) \tag{77}$$

of the decay term

$$\alpha_i^{+\,+}\gamma_i^{+\,+}(M_i^+ - x_i^+(t)) - \alpha_i^{+\,-}\gamma_i^{+\,-}(x_i^+(t) - m_i^+)$$

of (75) with its inhibitory part

$$\alpha_i^{+\,-}\gamma_i^{+\,-}(x_i^+(t) - m_i^+). \tag{78}$$

In (77), $x_i^+(t)$ is compared with its maximum M_i^+, whereas in (78) $x_i^+(t)$ is compared with its minimum m_i^+. We illustrate this situation in Fig. 13 in a suggestive way: the figure shows that $x_i^+(t)$ has an excitatory polarity in the excitatory part (77), which we denote by $\uparrow$, and an inhibitory polarity in the inhibitory part (78), which we denote by $\downarrow$. Guided by this fact,

we invoke EID by requiring that all expressions in $J_i^{+-}(t)$ have the *opposite polarity* to that of the corresponding expression in $J_i^{++}(t)$. For example, the expression $[x_m^+(t-\tau_{mi}^{+-}) - \Gamma_{mi}^{+-}]^+$ in $J_i^{+-}(t)$ corresponds to the expression $[\Gamma_{mi}^{++} - x_m^+(t-\tau_{mi}^{++})]^+$ in $J_i^{++}(t)$. In all, we find, by reversing polarities in (76), that

$$J_i^{+-}(t) = \sum_{m=1}^{n} \beta_m^{+-}[\Gamma_{mi}^{+-} - x_m(t-\tau_{mi}^{+-})]^+ p_{mi}^{+-}[\Gamma_{mi}^{+-} - z_{mi}^{+-}(t)]. \tag{79}$$

Given the existence of a process $x_i^+(t)$, EID requires that there also exists a process $x_i^-(t)$, whose equation follows from (75) by changing the + superscripts corresponding to the + in x_i^+ to − superscripts. Then (75) becomes

$$\begin{aligned} \dot{x}_i^-(t) = {} & \alpha_i^{--}(M_i^- - x_i^-(t))(\gamma_i^{--} + J_i^{--}(t) + I_i^{--}(t)) \\ & - \alpha_i^{--}(x_i^-(t) - m_i^-)(\gamma_i^{--} + J_i^{--}(t) + I_i^{--}(t)); \end{aligned} \tag{80}$$

$J_i^{--}(t)$ and $J_i^{-+}(t)$ are also found by changing + superscripts corresponding to x_i^+ to − superscripts in (76) and (79). We find

$$J_i^{--}(t) = \sum_{m=1}^{n} \beta_m^{--}[x_m^-(t-\tau_{mi}^{--}) - \Gamma_{mi}^{--}]^+ p_{mi}^{--}[z_{mi}^{--}(t) - \Omega_{mi}^{--}]^- \tag{81}$$

and

$$J_i^{-+}(t) = \sum_{m=1}^{n} \beta_m^{-+}[\Gamma_{mi}^{-+} - x_m^-(t-\tau_{mi}^{-+})]^+ p_{mi}^{-+}[\Omega_{mi}^{-+} - z_{mi}(t)]^-. \tag{82}$$

Equations (75) and (80) satisfy EID formally, and therefore seem to describe a complete symmetry between excitatory and inhibitory processes. Actually, as we will now show, these equations do not give rise to a sensible learning process unless the *coefficients* p_{ij}^{++}, p_{ij}^{+-}, p_{ij}^{--}, and p_{ij}^{-+} are constrained in an *asymmetric* way. This "symmetry breaking" within the formal symmetry of (75) and (80) is needed to guarantee the "evolutionary trend" with an excitatory bias that we call "learning" in these systems. (Is it possible that symmetry breaking of ostensibly symmetric systems in various other physical disciplines occurs to guarantee analogous evolutionary trends?)

The need for symmetry breaking is clearly seen by considering $J_i^{--}(t)$ in (79). $x_m^+(t)$ creates an inhibitory signal along e_{mi}^{--} whenever $\Gamma_{mi}^{+-} > x_m^+(t)$ and $p_{mi}^{--} > 0$. Since $\Gamma_{mi}^{+-} > P_m^+$, an inhibitory signal is created whenever $x_m^+(t)$ is at equilibrium and $p_{mi}^{+-} > 0$. This conclusion is absurd, since we have introduced the threshold Γ_{mi}^{+-} to guarantee that only large $x_m^+(t)$ values create signals. Inhibitory signals at equilibrium would ultimately

wash away all learned associations. To eliminate this catastrophic possibility, we let

$$p_{mi}^{+-} = 0, \tag{83}$$

and thus $J_i^{+-}(t) \equiv 0$. By analogy, we also let $I_i^{+-}(t) \equiv 0$. Equation (75) then becomes

$$\begin{aligned} \dot{x}_i^+(t) = {} & \alpha_i^{++}(M_i^+ - x_i^+(t))(\gamma_i^{++} + J_i^{++}(t) + I_i^{++}(t)) \\ & - \alpha_i^{++}\gamma_i^{++}(x_i^+(t) - m_i^+). \end{aligned} \tag{84}$$

Equation (84) still obeys EID *formally*, but the constraint (83) gives the equation an asymmetric appearance.

Since no actual inhibitory inputs are permitted in (84), the very concept of inhibition can only be salvaged if inhibitory inputs are permitted in (80). As in all our previous discussions, we must require that an inhibitory signal from v_m into e_{mi} at time t is created by an increase in $x_m^+(t)$ above its inhibitory threshold Γ_{mi}^{+-}, and therefore has the form

$$\beta_m^{+-}[x_m^+(t) - \Gamma_{mi}^{+-}]^+. \tag{85}$$

One of the two output expressions

$$\beta_m^{+-}[x_m^+(t) - \Gamma_{mi}^{+-}]^- \tag{86}$$

or

$$\beta_m^{+-}[\Gamma_{mi}^{+-} - x_m^+(t)]^- \tag{87}$$

in $J_{mi}^{+-}(t)$ and $J_{mi}^{--}(t)$, respectively, must therefore equal (85), for every $m = 1, 2, \ldots, n$. Suppose (86) equals (85). Then, as in (84), we must impose the constraint $p_{mi}^{+-} = 0$ for every $m = 1, 2, \ldots, n$. Thus $J_i^{+-}(t) \equiv 0$, and by analogy, $I_i^{+-}(t) = 0$. $x_i^+(t)$ and $x_i^-(t)$ then obey equations of identical form in which no inhibitory signals occur. Inhibitory signals are only possible if (85) equals (87). Then

$$\beta_m^{+-}[x_m^+(t) - \Gamma_{mi}^{+-}]^+ = \beta_m^{+-}[\Gamma_{mi}^{+-} - x_m^+(t)]^+, \tag{88}$$

$$p_{mi}^{+-} = 0, \qquad i, m = 1, 2, \ldots, n \tag{89}$$

and $I_i^{-+}(t) \equiv 0$, $i = 1, 2, \ldots, n$. Equation (80) becomes

$$\dot{x}_i^-(t) = \alpha_i^{--}\gamma_i^{--}(M_i^- - x_i^-(t)) - \alpha_i^{--}(x_i^-(t) - m_i^-)(\gamma_i^{--} + J_i^{--}(t) + I_i^{--}(t)). \tag{90}$$

The important relation (88) implies that whenever $x_m^+(t)$ has a suprathreshold value, then

$$\begin{aligned} \beta_m^{+-}x_m(t) + \beta_m^{--}x_m^-(t) &= \beta_m^{--}\Gamma_{mi}^{+-} + \beta_m^{--}\Gamma_{mi}^{-+} \\ &= \text{constant}. \end{aligned} \tag{91}$$

That is, $x_m^{--}(t)$ and $x_m^{-}(t)$ are linearly, but *antagonistically*, coupled when $x_m^{+}(t)$ is driven to suprathreshold values by the excitatory input $J_m^{+*}(t)$ in (84). The coupling (88) cannot be valid when $x_m^{-}(t)$ is driven to small values by the inhibitory input $J_m^{--}(t)$ in (90), or else we must accept the absurd conclusion that a large inhibitory input to v_m creates a large excitatory output from v_m. The coupling (88) is therefore *broken* by an inhibitory input signal. This situation can be formally expressed in terms of the function

$$\chi(w) = \begin{cases} 1 & \text{if } w > 0 \\ 0 & \text{if } w \leqslant 0. \end{cases}$$

Then we replace (88) by

$$\chi([x_m^{-}(t) - \Gamma_{mi}^{--}]^{+})\beta_m^{-}[x_m^{-}(t) - \Gamma_{mi}^{--}]^{+} - \beta_m^{--}[\Gamma_{mi}^{--} - x_m^{-}(t)]^{+} = 0, \quad (92)$$

so that the coupling (88) holds only if

$$\chi([x_m^{-}(t) - \Gamma_{mi}^{--}]^{+}) = 1,$$

which is the same as saying that $x_m^{-}(t)$ has been driven to suprathreshold values by an excitatory input.

The foregoing simple manipulations can be summarized in the following way. The equation (43) contains a latent symmetry between its excitatory and inhibitory interactions. The effort to make this symmetry explicit shows that it describes a process $x_m^{+}(t)$ with *positive polarity* and a process $x_m^{-}(t)$ with *negative polarity*, along with four possible correlational (or associational) processes $z_{mi}^{++}(t)$, $z_{mi}^{+-}(t)$, $z_{mi}^{-+}(t)$, and $z_{mi}^{--}(t)$. The four new processes $x_m(t)$, $z_{mi}^{++}(t)$, $z_{mi}^{+-}(t)$, and $z_{mi}^{--}(t)$ can be thought of as new formal degrees of freedom that must be coupled to the old variables $x_m^{+}(t)$ (or $x_m(t)$) and $z_{mi}^{+}(t)$ (or $z_{mi}(t)$) in such a way that the *dynamics of learning* (or the evolutionary trends) in our original psychologically derived equations are not lost. The mechanism needed for learning is, however, manifestly *asymmetric* with respect to excitatory and inhibitory interactions. We hereby find the equations (84), (90), and (92), which exhibit an *approximate symmetry* that must nonetheless sometimes be *broken*. In this sense, the modest symmetry breaking within the very special situation of (84), (90), and (92) reconciles, at least formally, two pervasive but not manifestly compatible tendencies within nature; namely, (i) the creation of systems whose excitatory and inhibitory interactions are as symmetric as possible, and (ii) the creation of systems that can benefit from their experience, and can thereby evolve to ever more efficient and elaborate levels of organization. We will presently see that this symmetry

breaking seems to occur *in vivo* in the response of intracellular Na^+ and K^+ fluxes to excitatory and inhibitory transmitter substances.

Before establishing this fact, we remark that another antagonistic coupling is suggested by EID. Certainly the associations $z_{mi}(t)$ that occur in the expression $J_i^{--}(t)$ should have the notation $z_{mi}^{--}(t)$. By (82) we therefore find that

$$p_{mi}^{--}[z_{mi}^{--}(t)\,\Omega_{mi}^{--}]^{+} = p_{mi}^{--}[\Omega_{mi}^{+} - z_{mi}^{--}(t)]^{+}, \quad (93)$$

which means that at suprathreshold values of $z_{mi}^{--}(t)$,

$$p_{mi}^{--}z_{mi}^{--}(t) + p_{mi}^{--}z_{mi}^{--}(t) = \text{constant}. \quad (94)$$

EID has hereby created the following expectations.

1. Two processes $x_i^{+}(t)$ and $x_i^{-}(t)$ exist in every cluster of "cell bodies" v_i.

2. These processes are antagonistically coupled at suprathreshold values, such that:

3. An excitatory signal to v_i at time t causes an increase in $x_i^{+}(t)$ within v_i, which in turn at suprathreshold values causes a decrease in $x_i^{-}(t)$ within v_i;

4. An inhibitory signal to v_i at time t merely causes a decrease in $x_i^{-}(t)$ within v_i;

5. A process $z_{mi}^{++}(t)$ occurs in every excitatory "end bulb" cluster N_{mi}^{++} and contributes to the excitatory signal discussed in property 3;

6. A process $z_{mi}^{--}(t)$ occurs in every inhibitory end bulb cluster N_{mi}^{--} and contributes to the inhibitory signal discussed in property 4.

Properties 1–6 suggest an obvious neural interpretation of the quantities $x_i^{+}(t)$, $x_i^{-}(t)$, $z_{mi}^{++}(t)$, and $z_{mi}^{--}(t)$. We introduce this interpretation herein in a qualitative way because our formal work must still be extended to achieve a quantitative connection. Let

$x_i^{+}(t)$ = amount of Na^+ inside the v_i "cell membrane" at time t,
$x_i^{-}(t)$ = amount of K^+ inside the v_i "cell membrane" at time t,
$z_{mi}^{++}(t)$ = amount of excitatory transmitter inside the end bulbs N_{mi}^{++},
$z_{mi}^{--}(t)$ = amount of inhibitory transmitter inside the end bulbs N_{mi}^{--}.

The conclusions 1–6 now read:

1′. Each cell body cluster v_i contains significant amounts of Na^+ and K^+.

2′. An antagonistic coupling between the amounts of Na^+ and K^+ in v_i exists at suprathreshold values, such that:

3′. Excitatory transmitter causes an inward flow of Na^+, which in turn at suprathreshold values causes an outward flow of K^+;

4′. Inhibitory transmitter merely causes an outward flow of K^+.

Moreover, the process by which the signal $[x_m^+(t) - \Gamma_{mi}^+]^+$ traverses the axons e_{mi}^+ and e_{mi}^- now reads:

5′. The spikes along e_{mi}^+ are due to an inward flow of Na^+ coupled to an outward flow of K^+.

All of these phenomena have been experimentally reported [3, 4, 27–30]. A number of other interesting phenomena, some new, sit in the equations waiting to be interpreted. This will be done in a later article, which will also describe the equations for $z_{mi}^{++}(t)$ and $z_{mi}^{--}(t)$.

It is certain that our formalism is at best a rough description of neural events. In fact, later papers will show how to extend the formalism considerably in a rational way. Nonetheless it is gratifying that such highly nontrivial experimental properties as 1′–5′ should have formal analogs in a theoretical picture that is derived in a simple and rather inevitable way from such basic principles as locality, principle of sufficient reason, and excitatory–inhibitory dualism, along with an elementary analysis of what we mathematically mean by learning. If the leap from mathematical to neural variables is accepted, then the results 1′–5′ are *consequences* of these principles, and in this sense, we know "why" 1′–5′ occur and how these experimental properties contribute to learning. Most important, some of the basic facts about Na^+ and K^+ fluxes, which heretofore have been thought of as a part of the repetitive, and therefore *stationary*, responses of nerves to input signals are now implicated in the dynamics of neural learning, which is a *nonstationary* phenomenon.

REMARK. Equations (84), (90), and (92) cannot possibly be in their final form. We can see this by analogy with (**), in which two kinds of transformations occur: "differential" transformations that describe the rate of change of a process using a differential equation, and "algebraic" transformations such as $z_{ij} \to y_{ij}$, which describe a very fast process in an approximate way. Equations (84) and (90) are examples of differential transformations, whereas (92) is a very fast process approximately described. Equation (92) manifestly describes a coupling of $x_i^+(t)$ and $x_i^-(t)$ *within* the cell bodies v_i, but thus far v_i has no "cellular interior" in which to study the details of the coupling. Each point must be blown up into an extended cell body before (92) can be replaced by a differential

transformation. How this can be done will be shown in a later paper, where dendritic and other effects of cell shape will also be discussed. Clearly a first step in this modification is to weaken the coupling between $x_i^+(t)$ and $x_i^-(t)$ in (90) and (92), so that an outward flux of K^+ can be created by a linear mixture of inward excitatory transmitter-induced Na^+ fluxes and of directly applied inhibitory transmitter at different spatial loci on the blown-up boundary of v_i. Otherwise (90) and (92) are in general incompatible as they stand.

ACKNOWLEDGMENTS

The preparation of this work was supported in part by the National Science Foundation (GP 9003) and the Office of Naval Research (N00014-67-A-0204-0016).

REFERENCES

1 S. Grossberg, Embedding fields: A new theory of learning with physiological implications, *J. Math. Psychol.* 6(1969).

2 J. C. Eccles, *The physiology of nerve cells*, The Johns Hopkins Press, Baltimore, Maryland, 1957.

3 J. C. Eccles, *The physiology of synapses*, Academic Press, New York, 1964.

4 A. L. Hodgkin, *The conduction of the nervous impulse*, Charles C Thomas, Springfield, Illinois, 1964.

5 F. Ratliff, *Mach bands: Quantitative studies on neural networks in the retina*, Holden-Day, San Francisco, California, 1965.

6 S. Grossberg, Nonlinear difference-differential equations in prediction and learning theory, *Proc. Natl. Acad. Sci. USA* 58(1967), 1329–1334.

7 S. Grossberg, Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity, *Proc. Natl. Acad. Sci. USA* 59(1968), 368–372.

8 S. Grossberg, On the serial learning of lists, *Math. Biosci.* 4(1968), 201–253.

9 S. Grossberg, A prediction theory for some nonlinear functional-differential equations: I, II; *J. Math. Anal. Appl.* 21(1968), 643–694; 22(1968), 490–522.

10 S. Grossberg, On the global limits and oscillations of a system of nonlinear differential equations describing a flow on a probabilistic network, *J. Diff. Eqs.* (1969).

11 S. Grossberg, On the variational systems of some nonlinear difference-differential equations, *J. Diff. Eqs.* (1969).

12 A. I. Khinchin, *Mathematical foundations of information theory*, Dover, New York, 1957.

13 C. E. Osgood, *Method and theory in experimental psychology*, Oxford Univ. Press, London and New York, 1953.

14 J. M. Brookhart, in *Handbook of physiology*, Vol. II, *Neurophysiology* (J. Field, ed.), pp. 1245–1280. Amer. Physiol. Soc., Washington, D.C., 1960.

15 H. H. Jasper, in *Handbook of physiology*, Vol. II, *Neurophysiology* (J. Field, ed.), pp. 1307–1321. Amer. Physiol. Soc., Washington, D.C., 1960.

16 D. H. Barron and B. H. C. Matthews, *J. Physiol. (London)* **92**(1938), 276.
17 B. Katz, *J. Physiol. (London)* **111**(1950), 261.
18 C. A. Terzuolo and Y. Washizu, *J. Neurophysiol.* **25**(1962), 56.
19 T. H. Bullock, *Science* **129**(1959), 997.
20 M. G. F. Fuortes, *Amer. Natur.* **43**(1959), 213.
21 W. A. H. Rushton, in *Sensory communication* (W. A.Rosenblith, ed.), pp. 169–181. M.I.T. Press, Cambridge, Massachusetts, 1959.
22 E. D. P. DeRobertis, *Histophysiology of synapses and neurosecretion*, Macmillan, New York, 1964.
23 B. Katz, *Nerve, muscle, and synapse*, McGraw-Hill, New York, 1966.
24 H. K. Hartline and F. Ratliff, *J. Gen. Physiol.* **40**(1957), 357.
25 H. K. Hartline and F. Ratliff, *J. Gen. Physiol.* **41**(1958), 1049.
26 F. Ratliff, in *Sensory communication* (W. A. Rosenblith, ed.), pp. 183–203. M.I.T. Press, Cambridge, Massachusetts, 1959.
27 A. L. Hodgkin and A. F. Huxley, *J. Physiol. (London)* **166**(1952), 449.
28 A. L. Hodgkin and A. F. Huxley, *J. Physiol. (London)* **116**(1952), 473.
29 A. M. Shanes, *Pharmacol. Rev.* **10**(1958), 59.
30 A. M. Shanes, *Pharmacol. Rev.* **10**(1958), 165.