Reprinted from:
Journal of Cybernetics, 1971, 1, 1, pp. 28–50

Embedding Fields: Underlying Philosophy, Mathematics, and Applications to Psychology, Physiology, and Anatomy[1]

Stephen Grossberg†
Massachusetts Institute of Technology
and
The University of Warwick

Abstract

This article reviews results on a learning theory that can be derived from simple psychological postulates and given a suggestive neurophysiological, anàtomical, and biochemical interpretation. The neural networks described can discriminate, learn, simultaneously remember, and perform individually upon demand any number of space-time patterns of essentially arbitrary complexity. A general theorem expressing this fact is stated in the language of nonlinear functional–differential systems. Applications of the theory to various empirical problems are mentioned; e.g., serial learning, stimulus sampling, lateral inhibition, energy–entropy dependence, reaction time, transmitter production and release, spatiotemporal masking, operant and respondant conditioning, influences of under- or over-arousal on learning.

1. Introduction

This paper reviews some results concerning various networks, or machines, that can discriminate, learn, simultaneously remember, and perform individually upon demand any number of space-time patterns (e.g., motor sequences, reflexes, internal perceptual representations) of essentially arbitrary complexity. These networks are part of a nonstationary prediction, or learning, theory which is called the theory of *embedding fields.* Speaking mathematically, the machines of embedding field type are described by cross-correlated flows on signed directed networks. These flows obey systems of nonlinear functional-differential equations, and the mathematician's task is to globally analyze the limiting and oscillatory behavior of these systems.

[1] The preparation of this work was supported in part by the National Science Foundation (GP 9003), the Office of Naval Research (N00014-67-A-0204-0051), and the A. P. Sloan Foundation (71609).

†Requests for reprints should be sent to Dr. Stephen Grossberg, Department of Mathematics, M.I.T., Cambridge, Mass. 02139.

The embedding field equations can be plausibly derived from simple psychological facts that are familiar to us all from daily life. For example, one makes rigorous the statement that predicting the letter *B*, given the letter *A*, can be accomplished after practicing the list *AB* sufficiently often. Once these mathematical equations are available, one can proceed in at least four directions.

First, one can perturb the systems with inputs that represent complicated experiences, and check in a variety of cases that the mathematical behavior is qualitatively similar to psychological data that has been gathered in analogous experiments. To the extent that the psychological interpretation of these mathematical systems is valid, one can then say that the more complicated psychological behavior is merely the net effect, under complicated initial and boundary conditions, of such simple ideas as: practicing *AB* allows one to predict *B* given *A*. Some examples of this reduction will be mentioned below. Through exercises of this kind, the theory tries to show that some seemingly complex and unrelated psychological phenomena are manifestations of a few simple and familiar behavioral facts.

Second, one observes that the mathematical equations are already in a form that suggests a plausible neurophysiological, anatomical, and in some cases biochemical interpretation. After labelling the mathematical variables in this interdisciplinary fashion, one invokes the psychologically derived laws governing these variables; and checks to see whether the resulting neurophysiological, anatomical, and biochemical statements conform to known data. In some cases, one finds qualitative or even quantitative agreement. In other cases, one is led to new predictions, for example concerning possible mechanisms regulating changes in transmitter production rates.

Third, one investigates the purely mathematical problem of generalizing to the farthest possible extent the psychologically derived equations. One does this for two reasons. First, one thereby introduces a large class of nonlinear functional-differential systems which can be globally analyzed, and are therefore interesting in their own right. Second, by noting closely related alternative mathematical possibilities, one sees more clearly the formal advantages of the psychologically derived equations, including their stability under perturbations.

Fourth, one recognizes that a psychologically derived theory cannot at the outset incorporate many of the microscopic physiological and biochemical interactions that are important in brain studies. It is altogether proper that this be so, since our psychological experience does not include an intuitive awareness of the complex biochemical interactions that subserve it. In other words, our psychological experience averages over individual neural and biochemical events. One can, however, refine in successive steps the spatial and temporal scales of the equations, and thereby pick up in a natural way a number of these finer interactions, such as possible transients in the production and release of transmitter substances.

The remainder of this paper will pursue the following strategy. Some results will be stated in heuristic terminology to illustrate the interdisciplinary breadth of the theory. These results should be interpreted in the following way. They colorfully describe mathematical properties of our networks using empirical terminology. There exists no necessary connection between the mathematical variables and their empirical interpretation. This limitation holds for every physical theory, however, and the success of the formalism in so many areas suggests that some degree of truth ascribes to the particular empirical interpretation herein.

After stating these results and related references, a terse derivation of the theory's equations will be given to illustrate the contact between psychological and mathematical variables. Then aspects of the physiological and anatomical interpretation will be noted. Thereafter, a simple way to learn space-time patterns will be presented, and then the paper will conclude by stating a general theorem that illustrates the contribution of the theory to functional-differential systems.

2. Review of Some Results

A) *Serial Learning.* Consider the problem of learning a long list of similar behavioral events, say the alphabet $ABC \ldots XYZ$, by presenting this list several times consecutively to a learning subject and requiring the subject to guess the proper successor of each letter. A large experimental literature exists concerning such tasks, and our networks qualitatively exhibit some of the remarkable properties of the data found in these experiments. For example, bowing occurs: the middle of the list is harder to learn than the beginning or end. Backward learning occurs: practicing AB also partially teaches the list BA. On the other hand, if ABC occurs, then the association $B \rightarrow C$ ultimately inhibits the association $B \rightarrow A$: a global "arrow in time" is induced. Anchoring occurs: the order in which items are learned proceeds both in the forward and backward direction around the "anchor" stimulus A; for example AB, then YZ, then BC, then XY, and so on, might be learned. Chunking occurs: simple behavioral units are gradually aggregated into composite units as practice continues. After sufficient practice, the composite units can be performed without difficulty much as the simple units were at the outset. One can also study such phenomena as all-or-none vs. gradualist learning, the distribution of anticipatory vs. perseverative errors as a function of list position, accumulation of inhibition near the list's middle, and the dependence of various other learning properties on intratrial interval, intertrial interval, list length, list position, and reaction time.

Underlying these observations is the basic dynamical fact that the geometry of a list, as it is learned in tine, is a space-time geometry which is not isomorphic with the geometry of the list as a row of symbols on paper. For example, let the alphabet $ABC \ldots XYZ$ be presented to a learning subject with a time spacing of

w between successive letters. What is the earliest time at which the subject knows what Z is the end of the list? It is not until w time units have passed *after* Z is presented that the subject can know that Z is the end of the list. Until that time, Z is more properly in the (dynamical) middle of the list. In other words, a "time reversal" relative to the time scale of the psychological experimentalist then changes the internal dynamics of the machine. Z can be in the (dynamical) middle or end, or even the beginning of the list at different times; Reference [1] discusses the applications of the theory to these phenomena.

Two papers ([2], [3]), jointly with James Pepe of M.I.T., study the dependence of serial phenomena on the size of spiking thresholds and arousal level. It has been proved, that very low thresholds (overarousal) yield easier learning at the end of the list than at the beginning (recency vs. primacy), whereas higher thresholds yield the more normal easier learning at the beginning of the list than at the end (primacy vs. recency). Lowering the threshold has the effect of ultimately flooding the network with background noise due to past inputs. Since the network dynamics loses all but the most recent inputs in background noise, the network is unable to effectively use any but the most recent inputs to determine its next behaviors. In other words, the ability of the network to "pay attention" decreases as the signal thresholds decrease. Analogous pathological effects seem to occur in certain manics and schizophrenics.

B) *Stimulus Sampling.* The statistical learning theory known as stimulus sampling theory describes changes in the transition probabilities of an organism's responses in time as a result of reinforcing events, using the formalism of finite Markov chains [4]. The theory has succeeded in describing various experimental data, but is weakened by the lack of a concrete physical interpretation for the abstract stimulus sampling operation. The present theory behaves much like the stimulus sampling formalism in simple cases, and provides an entirely concrete psychological, physiological, anatomical, and even biochemical interpretation of stimulus sampling probabilities and the sampling operation. This interpretation involves, for example, laws for the potentiation of transmitter production at specific synaptic knobs, and for the spatiotemporal distribution of positive spiking frequencies in the axons of these synaptic knobs. See [5] for a discussion of these facts.

C) *Towards Resolution of Psychological Controversies.* The history of psychology can profitably be viewed as a history of controversies, many of them never resolved; for example, all-or-none learning theorists vs. gradualist learning theorists, peripheralists (or contiguity theorists) vs. gestaltists, etc. Why do so many controversies exist, and why do they persist? Each school of thought is supported by some convincing data, and it is my impression that various controversies arose because experimentalists tacitly interpreted their data concerning stimuli and responses using mechanisms which are mathematically too linear and too local. This was quite natural, since almost all macroscopic physical theorizing up to the time this data appeared used linearity and locality in a basic way.

The present theory is not always linear and local on the state space of stimuli and responses, but it often has a more linear and local behavior when the data of a particular psychological school is considered. Aspects of data from each of these schools is not unfamiliar in our networks due to the inherent nonlinearity and nonlocality of our systems in the large. From this vantage point, each pair of schools in a controversy picked out two polar extremes in a particular continuum of experimental possibilities within the phase space of initial data and experimental inputs. No two controversies picked out the same continuum, and no one controversy could be resolved using a linear and local dynamical mechanism. It is a credit to the remarkable intuition and integrity of these experimentalists that these controversies and their dynamical lessons have endured. See [1] and [6] for further discussion.

D) *Lateral Inhibition (Hartline-Ratliff).* Inhibitory interactions are abundant in the nervous system. An empirical equation that describes steady-state inhibitory interactions in the *Limulus* retina is the Hartline-Ratliff equation [7]. Given the special anatomy found in this retina, our equations reduce formally to the Hartline-Ratliff equation and provide theoretical formulas for the empirical coefficients of that equation; see [8].

New aspects of inhibitory dynamics have also been found; for example, inhibition contributes to the phenomenon of "enhancement of associational strengths" or "spontaneous improvement of memory", closely related to "contour enhancement" due to lateral inhibition [8]. This phenomenon shares many properties with the Ward-Hovland phenomenon, or "reminiscence", and is an example of self-improving properties in the networks. Other uses for inhibitory interactions are stated in the next paragraphs.

E) *Pattern Discrimination.* It is well-known that individual receptor cells in various sensory modalities can respond to many different input patterns; for example, a single retinal cell can be activated by many different visual scenes. On the other hand, the nervous system as a whole can discriminate one pattern from another. How can "local nonspecificity" and "global specificity" of cellular response be reconciled? Related to this question is the data of Hubel and Wiesel and their colleagues ([9]-[11]), which has shown the existence of nerve cells that respond most vigorously to specified patterns at the sensory periphery of perhaps great complexity.

We have introduced cellular configurations that can discriminate any number of space-time patterns of essentially arbitrary complexity in a way that permits these discriminations to create appropriate responses based on past experience, and which include cells with specific output preferences in a natural way [12]. This construction uses very few cells, except for the cells in the receptor mosaic, and can be used to discriminate any patterns, whatever be their sensory interpretation.

The construction yields some formal insights concerning the following phenomena: uses of nonrecurrent inhibitory interneurons for temporal or spatial discrimination tasks which recurrent inhibitory interneurons cannot carry out; mechanisms of temporal generalization whereby the same cells control performance of a given act at variable speeds; tendency of cells furthest from the sensory periphery to have the most discriminative response modes, and the least ability to follow sensory intensities (e.g., on-off and bimodal responses are common); uses of nonrecurrent on-off cellular fields whose signals arrive in waves forming "interference patterns," with the net effect of rapidly choosing at most one behavioral mode from any number of competitive modes, or of nonspecifically arousing or suppressing cells which can sample and learn ongoing internal patterns of cellular activity (cf., operation of the reticular formation [13]); uses of specific vs. nonspecific inhibitory interneurons, axon hillock inhibition, presynaptic inhibition, equal smoothing of excitatory and inhibitory signals, possible production of both excitatory and inhibitory transmitter in a single synaptic knob, blockade of postsynaptic potential response, logarithmic transduction of inputs to spiking frequencies, and saturation of cell body response in nonrecurrent on-off fields for purposes of pattern discrimination.

Since the same cellular configuration can in principle be used to discriminate patterns from any sensory modality, the following basic question arises. Why are the anatomies of pathways in different sensory modalities so different? The answer seems to lie in the following direction: the sensory anatomies of higher animals structurally contain provisions for guaranteeing the particular perceptual constancies of that modality, as well as provisions for making "operant" as well as passive discriminations.

It is important to note that knowing the anatomy of a given cellular configuration does not determine its capabilities as an input filter. One must also know such physiological parameters as the relative strengths and onset times of excitatory and inhibitory signals at a given cell, the relative speeds of exponential averaging at different cells, the spatial distribution of spiking threshold values at all cells, etc.

F) *One Cell One Pattern?* The following question is of considerable interest. What is the *minimal* number of cells needed to encode the memory of a space-time pattern, such as a piano sonata? The answer in our networks is "one"! Unfortunately, such a network exhibits a profound liability: performance of the pattern is always ritualistic, or by rote, and once performance of the pattern begins, it is hard to stop.

The cells which can learn in this way have profusely branching bushes of axon collaterals. Insects also have some large cells which control performance of large sectors of important reflexes; e.g., feeding or withdrawal [14]. They also pay the price of ritualistic performance.

It thus appears that a more subtle performance, tuned to feedback from prior events, requires the encoding of any given pattern in many different cells, no one capable of irrevocably eliciting the entire pattern. Such an encoding creates another problem; namely, since many different cells will then fire to common motor control cells in rapid succession, background noise can easily build up at these control cells and thereby impair performance. To keep the background noise down, inhibitory interactions are needed which can rapidly be excited by the excitatory control cells, and thereby inhibit the excitatory signals shortly after they occur. Consider for example the special, but important, case in which the sensory and motor modalities have a linear ordering, at least in first approximation; for example, the fingers, successive joints on arms and legs, the spine, the tonotopic representation of the auditory system, etc. In the linearly ordered case a plausible construction of the inhibitory mechanism yields a cellular configuration that is strongly reminiscent of cerebellar neocortex, and the inhibitory cells are then interpreted as cerebellar Purkinje cells [15].

G) *Energy-Entropy Dependence.* The Second Law of Thermodynamics teaches us that the Universe is heading inexorably towards a maximal entropic doom. On the other hand, daily experience with living creatures assures us that powerful forces yielding ever greater order exist in Nature; for example, Evolution. In our networks, a similar preference for order arises in a special case.

We note first the desirability of having organisms whose complexity of response is appropriate to the complexity of stimulus demand; for example, a sharp pin prick more often elicits rapid withdrawal than poetry recitation. Speaking roughly, simple demands elicit less cellular processing than complicated demands. One important parameter that influences whether or not many cells will be activated at any time is the amount of energy that is supplied to the sensory periphery which reaches internal cells. In special cases of interest, we have proved that peripheral inputs with minimal entropy maximize the total energy transfer to these internal cells, and inputs with maximal entropy minimize the total energy transfer [16]. In other words, the most complicated demands elicit the most complicated internal response, energetically speaking. This preference for order can be traced to the learning, or "evolutionary," mechanism in our open systems.

H) *Energy-Learning Dependence: Reaction Time.* It is well-known that increasing the energy of an input can speed up the performance of a suitable output, and that an input of fixed energy can yield faster performance if the input-output connection is a familiar one. Model systems are now available in which the dependence of reaction time on peripheral energy and on the degree of prior learning can be studied in a unified fashion [8].

I) *Transmitter Production and Cellular Control.* Our theory suggests that in cells capable of learning, presynaptic transmitter production in jointly controlled by presynaptic spiking frequency and postsynaptic potential [17]. This control is presumed to be effected by the interaction of the pairs (Na^+, K^+) and (Ca^{++}, Mg^{++}) of antagonistic ions whose binding properties to intracellular sites and enzymes set various cellular production levels. It is suggested that nerve cells are capable of learning as "chemical dipoles". A qualitative behavioral rationale can be set forth for such phenomena as the following: joint inward fluxes of Na^+ and Ca^{++} due to membrane excitation; distribution of mitochondria and synaptic vesicles near the synaptic cleft; sensitivity of RNA activation to Mg^{++} concentration; stronger binding of Ca^{++} relative to K^+ within the synaptic knobs; mobilization and depletion of transmitter by presynaptic spiking; post-tetanic potentiation; excitatory transients in transmitter release after a rest period; feedback inhibition of transmitter onto a late stage of transmitter production; transport down the axon of some lighter molecules produced in the cell body; proportionality of cell body membrane area to nuclear volume; intracellular tubules as faithful transport mechanisms between nerve cell body and nucleus, and from nucleus along axon to synaptic knobs; and division of cell shape into a cell body, axon, and synaptic knobs as a structural manifestation of the underlying chemical dipole.

J) *Phase Transitions in Memory: Myelinization.* For suitable choices of anatomy and physiology, there exists a division of the network's rate parameters into two regimes, R_1 and R_2. In one anatomy, for example, if the parameters fall into R_1, then all memories are eventually forgotten; if the parameters fall into R_2, then all memories persist with a precision that depends on the numerical value of a composite of these parameters. In a system with slightly different anatomy, again all memories persist if the parameters fall into R_2; if they fall into R_1, however, then only spatial patterns are remembered, and temporal discriminations are eventually forgotten. The dividing boundary between the regimes R_1 and R_2 is sharp, and in this sense the memory of the system goes through a kind of phase transition when it passes from one regime to the next.

It turns out that by speeding up the signal velocity in the network axons (or edges) one can, given fixed (but appropriate) choices of other network parameters, carry the system from regime R_1 to regime R_2. That is, speeding up signals can "rigidify" the memory. In vivo, there exists a way to speed up signals in axons; namely, encase the axon in a myelin sheath. To the extent that these examples have relevance to the physiological case, one can therefore contemplate the possibility that myelinization helps to rigidify the memory of past experiences. One must be very careful in making this proposal, however, since there exist yet another anatomies—with the very same local dynamics—for which no phase transitions in memory occur when the signal velocity is varied. See [18] and [19] for further details.

K) *"Wave-Particle" Dualism: Hidden Inhibitory Interactions.* One can simultaneously intrepret the theory statistically and deterministically. Deterministically, one studies the evolution through time of specific inputs and outputs. Statistically, one studies the evolution of transition probabilities that are formally associated with the theory in a natural way, and which one can think of as waves of excitation passing from one state to many other states through time.

This simultaneous interpretation is possible because the wave of excitation is later acted upon by inhibitory interactions that pick out a perfectly definite output in response to each input, if any output whatever occurs. The inhibitory interactions themselves execute a transformation on the system that is reminiscent of the informational functional, and that has the effect of fulfilling the "principle of sufficient reason": outputs occur only if they represent distinguishable paths in the machine. See [8] and [20] for further discussion.

L) *Spatiotemporal Masking and Consolidation.* Suppose that two spatially disjoint points of light are shined on a retina in succession. If the latter light point is much more intense than the former, then it can totally mask the former light if the spatial separation of the two light sources is not too great. In another direction, we know that memories which have lain dormant for many years can suddenly be triggered with remarkable clarity by suitable fragments of past experiences. These and related aspects of masking can be studied in simplified idealizations of the theory [8].

One factor that contributes to masking is the consolidation, after sufficiently many practice trials, of the memory trace into a relatively compact, rapidly evokable, cluster of cells. Tendencies to reach consolidation can also be studied in the theory [15].

It is hoped that the above heuristic remarks help to illustrate the interdisciplinary flavor of embedding field theory. To be sure, the theory will probably not have the last word on any of the above phenomena. Yet definite progress towards their understanding seems to have been achieved using the theory, and it is especially gratifying that all of these phenomena can at least partially be studied from a unified formal point of view. Below, to keep our discussion brief, we tersely list some other mathematical facts about learning in embedding fields.

M) *Pavlovian Conditioning.* Learning occurs by respondant, or Pavlovian, conditioning in our simplest networks, and this conditioning paradigm can be shown to be mathematically the same as practicing AB to predict B given A.

N) *Practice Makes Perfect.* The more often the behavior is practiced, the better will be the prediction. All-or-none learning effects are also possible, however.

O) *Memory.* A network exists whose memory is perfect even during recall trials; no overt or covert practice is needed to ensure perfect memory [20]. Another network exists whose memory decays exponentially at a rate that can be chosen arbitrarily small; cf., the Ebbinghaus forgetting curve [5]. In yet another network, memory is perfect until recall trials occur. During recall, memory can "extinguish" unless it is "rewarded" [5]. In the last two cases, "spontaneous recovery" of memory is possible, and can occur by a process analogous to "post-tetanic potentiation". Also, spontaneous improvement of memory, or "reminiscence", occurs after a moderate amount of practice in all three cases [5, 20].

Each case above uses a slightly different physiological mechanism. Our goal is to classify the behavioral possibilities and some of their physiological correlates. A large variety of other memory phenomena can also be achieved. For example, "pattern completion" is readily found.

P) *Errors.* All errors can be corrected in suitable networks. Also, if any number of patterns have already been learned by the networks, another pattern can also be learned without interfering with the memory of the other patterns, up to some finite upper bound [18, 15, 20].

Q) *Operant Conditioning.* The theory suggests that operant and respondant conditioning are dynamically very similar; they differ primarily in terms of the total distribution of excitation and inhibition that occurs in different conditioning paradigms, rather than in terms of the local learning equations at individual cells. One can give precise formal examples of such phenomena as "internal drive states", "nonspecific arousal stimuli", "paying attention", "goals", "novelty", "habituation", and "incompatible" vs. "facilitatory" behavioral modes in the networks. These involve nothing more than specific anatomical constructions using the same dynamical laws. Some of these constructions can, however, be rather subtle ([18], [22]).

3. Derivation of Some Networks

The derivation to be given will occur in story-book form to emphasize its intuitive basis. We begin with an experimentalist $\mathcal{E}$ who interacts with a machine $\mathfrak{M}$ to teach $\mathfrak{M}$ to predict B given A by practicing AB. Suppose for simplicity that the experimentalist can present the letters of the alphabet $A, B, C, \ldots, X, Y, Z$ to $\mathfrak{M}$ one at a time at prescribed instants of time. How can we represent the presentation of the letter A at time t_A in $\mathfrak{M}$?

A) *Each Letter Seems Simple.* In daily speech and listening, a letter is never decomposed into two parts. To maintain close contact with experience, we assume that a single state v_A in $\mathfrak{M}$ corresponds to A.

This assumption does not mean that one cell corresponds to A. As the theory is refined, one sees that a complicated trajectory of excitation and inhibition over many cells corresponds to hearing the letter A. How to reach this conclusion would not be clear, however, without first making the simplifying assumption. This assumption has a deep dynamical significance which focuses upon the fact that certain behaviors which seem very complex and spread out in space and time before learning, seem to be simple and quite instantaneously performed after learning [6]. The assumption means essentially that $\mathfrak{M}$ already knows the letters separately at the time learning begins, even though $\mathfrak{M}$ does not know any lists of letters.

In a similar fashion, let v_B correspond to B, v_C to C, etc. We designate each v_i by a point, or vertex, as in Figure 1.

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

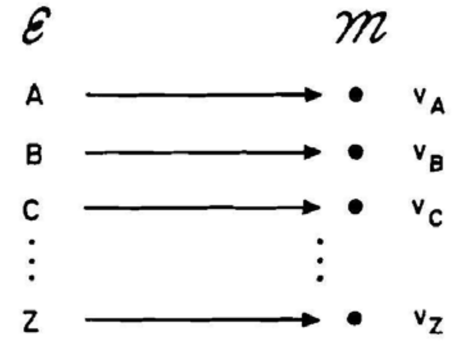

FIG. 1.

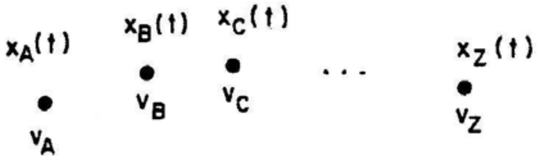

FIG. 2.

C) *Continuous Vertex Functions.* The functions $x_A(t), \ldots, x_Z(t)$ will be chosen continuous, and in fact differentiable. Several reasons for this exist, the most general being that all macroscopic theories have been cast in a continuous setting. Moreover, our daily experience has a manifestly continuous aspect and it is this that we seek to understand. More specifically, consider the following question. What follows *ABC*? It is tempting to say *D*, but really the problem is ill-defined if the letters are presented one at a time with time spacing w between successive letters. If indeed w is small, say $w \simeq 2$ seconds, then *D* might well be the correct response, but if $w \simeq 24$ hours then to the sound *C*(= "see") one can also reply "see what?" That is, as w varies from small to large values, the influence of *A* and *B* on the prediction following *C* gradually wears off. Since $x_A(t)$ and $x_B(t)$ describe the relevance at time t of *A* and *B* in $\mathfrak{M}$, we conclude that these functions also vary gradually in time.

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

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

Shortly after A is presented, A no longer is heard by $\mathfrak{M}$. That is, $x_A(t)$ gradually returns to the value signifying no recent presentation of A, namely 0.

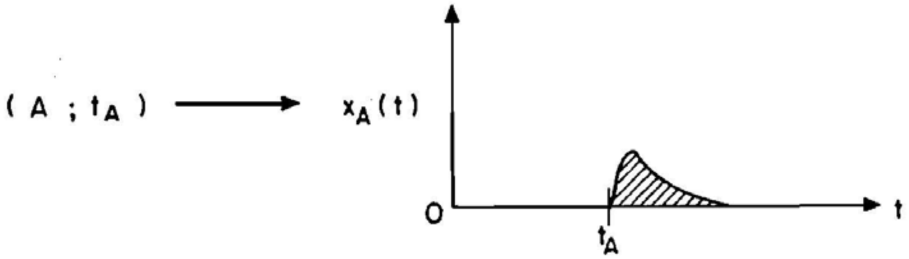

FIG. 3.

See Figure 3. In a similar fashion, if A is presented at times $t_A^{(1)} < t_A^{(2)} < \cdots < t_A^{(N_A)}$, then we find the graph of Figure 4. The same construction holds true for all letters. In this way, we have translated the presentation of any letters $A, B, C, \ldots$ in the alphabet at prescribed times into a definite sequence of perturbations of the vertex functions $x_A(t), x_B(t), x_C(t), \ldots$.

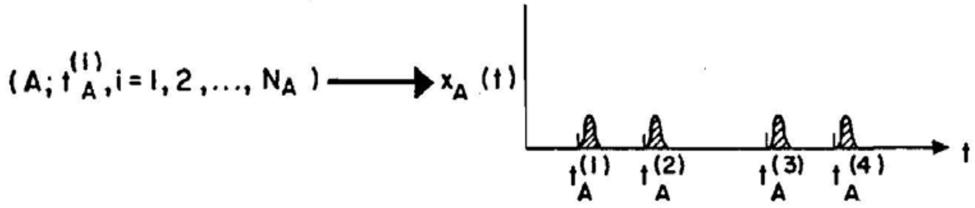

FIG. 4.

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

What is the simplest way to translate Figure 4 into mathematical terms? Since we are constructing a system whose goal is to adapt with as little bias as possible to its environment, we are strongly advised to make the system as linear as possible. The simplest linear way to write Figure 4 is in terms of the equations

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

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

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

where $J_i(t)$ is some nonnegative and continuous function that is positive in an interval of the form $(0, \lambda_i)$.

F) *After Learning.* In order that $\mathfrak{M}$ be able to predict B given A after practicing AB, interactions between the vertices v_i must exist. Suppose for example that $\mathfrak{M}$ has already learned AB, and that A is presented to $\mathfrak{M}$ at time t_A. We expect $\mathfrak{M}$ to respond with B after a short time interval, say at time $t = t_A + \tau_{AB}$, where $\tau_{AB} > 0$. τ_{AB} is called the reaction time from A to B. Let us translate these expectations into graphs for the functions $x_A(t)$ and $x_B(t)$. We find Figure 5. The input $I_A(t)$ controlled by $\mathfrak{E}$ gives rise to the perturbation of $x_A(t)$. The internal mechanism of $\mathfrak{M}$ must give rise to the perturbation of $x_B(t)$. In other words, after AB is learned, $x_B(t)$ gets large τ_{AB} units after $x_A(t)$ gets large.

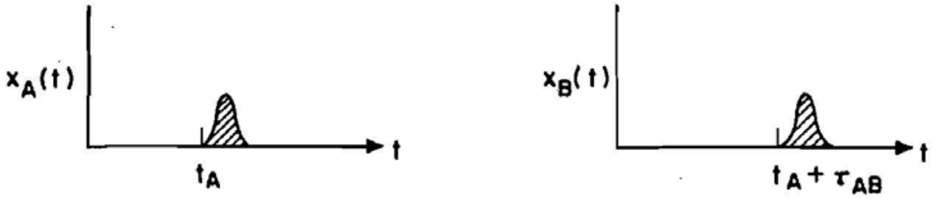

FIG. 5.

There exists a linear and continous way to say this; namely, v_A sends a linear signal to v_B with time lag τ_{AB}. Then (1) with $i = B$ is replaced by

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

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

$$\dot{x}_j(t) = -\alpha x_j(t) + I_j(t) + p_{ij} x_i(t - \tau_{ij}) . \tag{2}$$

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

G) *Directed Paths.*The signal $p_{ij} x_i(t - \tau_{ij})$ from v_i to v_j in (2) is carried along some pathway at a finite velocity, or else the locality of the dynamics would be violated. Denote this pathway by e_{ij}. The pathways e_{ij} and e_{ji} are distinct because the lists r_ir_j and r_jr_i are distinct. To designate the direction of flow in e_{ij}, we draw e_{ij} as an arrow from v_i to v_j whose arrowhead N_{ij} touches v_j, as in Figure 6.

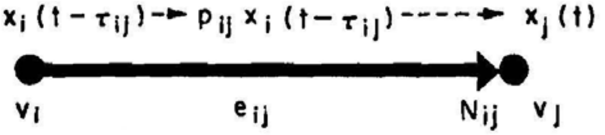
$x_i(t-\tau_{ij}) \to p_{ij}x_i(t-\tau_{ij}) \dashrightarrow x_j(t)$
v_i
e_{ij}
N_{ij}
v_j

FIG. 6.

H) *Before Learning.* Before any learning occurs, if A leads only to B, then learning would have already occurred. A must therefore also be able to lead to C, D, or some other letters, as in Figure 7. Thus the process of learning can be viewed as elimination of the incorrect pathways AC, AD, etc., while the correct pathway AB endures, or is strengthened.

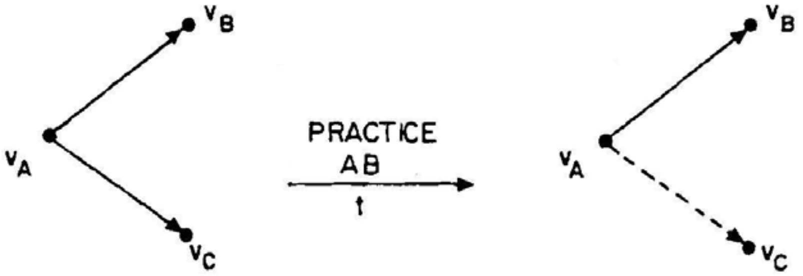

FIG. 7.

I) *Distinguishing Order.* How does $\mathfrak{M}$ know that AB and not AC is being learned? By Figure 5, practicing AB means that x_A and then x_B become large several times. Saying A alone, or B alone, or neither A nor B should yield no learning. This can be mathematically stated most simply as follows. If AB occurs with a time spacing w, then the product $x_A(t-w)x_B(t)$ is large at suitable times $t \simeq t_A^{(i)} + w$, $i = 1,2,\ldots,N_A$. We therefore seek a process in $\mathfrak{M}$ that can compute products of past $x_A(v)$ values $(v < t)$ and present $x_B(t)$ values. Denote this process by $z_{AB}(t)$. Note that $z_{AB} \neq z_{BA}$.

Where in $\mathfrak{M}$ do past values of $x_A(v)$ and present values of $x_B(t)$ come together, so that $z_{AB}(t)$ can compute them? By Figure 6, this happens only in the arrowhead N_{AB}. Thus $z_{AB}(t)$ takes place in N_{AB}. But then the past $x_A(v)$ value received by N_{AB} at time t is the signal $p_{AB}x_A(t-\tau_{AB})$. The most linear and continuous way to express this rule for $z_{AB}(t)$ is the following.

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

with β and λ positive constant, or more generally for $r_i r_j$, we find in N_{ij} the process

$$\dot{z}_{ij}(t) = -\beta z_{ij}(t) + \lambda p_{ij} x_i(t - \tau_{ij}) x_j(t) . \tag{3}$$

J) *Gating Outputs.* The $z_{ij}(t)$ function can distinguish whether or not $r_i r_j$ is practiced. But more is desired. Namely, if $r_i r_j$ is practiced, presenting r_i should yield a delayed output from v_j. If $r_i r_j$ is not practiced, presenting r_i should not yield an output from v_j. And even if $r_i r_j$ is practiced, no output from v_j should occur if r_i is not presented. In other words, $x_j(t)$ should become large only if $x_i(t - \tau_{ij})$ *and* $z_{ij}(t)$ are large. Again a product is called for, and (2) is changed to

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

K) *Independence of Lists in First Approximation.* Consider Figure 8. If B is not presented to $\mathfrak{M}$, then in first approximation CA should be learnable without interference from B. (Not so in second approximation, since a signal could travel from C to B to A.) Similarly if C is not presented to $\mathfrak{M}$, then BA should be learnable without interference from C, in first approximation. Mathematically speaking, this means that all signals to each v_j combine additively at v_j. Thus (4) becomes

$$\dot{x}_j(t) = -\alpha x_j(t) + I_j(t) + \sum_{i=1}^{n} x_i(t - \tau_{ij}) p_{ij} z_{ij}(t) . \tag{5}$$

The system (3) and (5) is a mathematically well-defined proposal for a learning machine that uses only such general notions as linearity, continuity, and locality, and a mathematical analysis of how a machine can learn to predict B given A on the basis of practicing AB.

L) *Thresholds.* One further modification of systems (3) and (5) is convenient; namely, the introduction of signal thresholds. Here we introduce this modification directly to keep background noise down. A more fundamental analysis would introduce it by first analyzing the need in complex learning situations for inhibitory interactions, and then by pointing out that learning becomes difficult without signal thresholds if inhibitory interactions exist.

A possible difficulty in (3) and (5) is this. Small signals can possibly be carried round-and-round the network thereby building up background noise and interfering with the processing of behaviorally important inputs. We therefore seek to eliminate the production of signals in response to small $x_i(t)$ values, in the

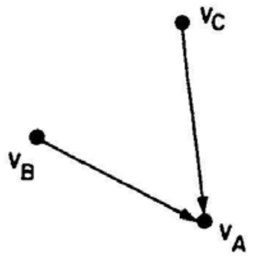

FIG. 8.

most linear possible way. Thresholds do this for us. Letting $[\xi]^+ = \max(\xi, 0)$, we replace (3) and (5) by

$$\dot{x}_i(t) = -\alpha x_i(t) + \sum_{m=1}^{n} [x_m(t - \tau_{mi}) - \Gamma_{mi}]^+ p_{mi} z_{mi}(t) + I_i(t) \tag{6}$$

and

$$\dot{z}_{jk}(t) = -\beta z_{jk}(t) + \lambda p_{jk} [x_j(t - \tau_{jk}) - \Gamma_{jk}]^+ x_k(t) , \tag{7}$$

where all Γ_{mi} are positive thresholds, and $i, j, k = 1, 2, \ldots, n$. Systems (6) and (7) complete the derivation of this paper. More general systems, which include inhibitory interactions and related decay and interaction laws, are discussed in [5] and [16].

4. Empirical Interpretation

The following empirical labels can be assigned to the mathematical variables.

a) v_i = ith cell body (or cell body cluster);
b) e_{ij} = axon(s) from v_i to v_j;
c) N_{ij} = synaptic knob(s) of e_{ij};
d) gap between N_{ij} and v_j = (i, j) th synapse;
e) $x_i(t)$ = average cell body potential of v_i at time t
= stimulus trace of r_i at time t;
f) $z_{ij}(t)$ = average amount of available excitatory transmitter (e.g., ACh) in N_{ij} at time t
= associational strength from r_i to r_j at time t;
g) Γ_{ij} = spiking threshold of e_{ij};
h) p_{ij} = axonal path weight of e_{ij};
i) $[x_i(t) - \Gamma_{ij}]^+ \simeq$ spiking frequency emitted from v_i into e_{ij} in the time interval $[t, t + dt]$.

Using these labels for $x_i(t)$ and $z_{ij}(t)$, for example, one can translate psychological statements into physiological statements, and vice versa. The new definitions of "stimulus trace" and "associational strength" give a rigorous dynamical description of the more heuristic terminology of Hull [21]. Note that equations (1) and (2) suggest a law for transmitter production that requires joint pre- and post-synaptic influences.

5. Space-Time Pattern Learning

This section describes the simplest anatomy for space-time pattern learning in our networks. We will describe this network heuristically, since the theorem in Section 6 includes its behavior, and that of vastly more complex situations, as special cases.

First we study the problem of spatial pattern learning in the smallest possible network. A spatial pattern on a grid of cells $v_i, i = 2, 3, \ldots, n$, is defined as a vector function $I_i(t) = \theta_i I(t)$, $i = 2, 3, \ldots, n$, of inputs delivered to these cells, where the θ_i's form a probability distribution $\left(\theta_i \geq 0 \text{ and } \sum_{k=2}^{n} \theta_k = 1\right)$, and $I(t)$ is a nonnegative and continuous function. θ_i is the relative intensity of the pattern at v_i, and $I(t)$ is the total intensity of the pattern at time t.

We seek a network having a minimal number of cells that can learn any spatial pattern by respondant conditioning. The spatial pattern is the unconditioned stimulus (US), and we can think of the cells v_i, $i \neq 1$, as controlling, for example, a given collection of muscle groups. Clearly we need at least one more cell v_1 to which the conditioned stimulus (CS) will be delivered, and whose excitation can ultimately reproduce the pattern on the cells v_i, $i \neq 1$. This cell v_1 must be able to send signals to each v_i, $i \neq 1$, or else not all patterns could be reproduced. That is, the edges e_{1i}, $i \neq 1$, exist. The situation is diagrammed in Figure 9. Figure 9a shows that v_1 sends axon collaterals to all v_i, $i \neq 1$. Figure 9b idealizes this situation. The network in Figure 9b is called an *outstar* with *source vertex* v_1, *sink vertices* v_i, $i \neq 1$, and *border* $B = \{v_i\,;\ i \neq 1\}$.

The main mathematical objects of study are the probabilities

$$X_i(t) = x_i(t)\left[\sum_{k=2}^{n} x_k(t)\right]^{-1}$$

and

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

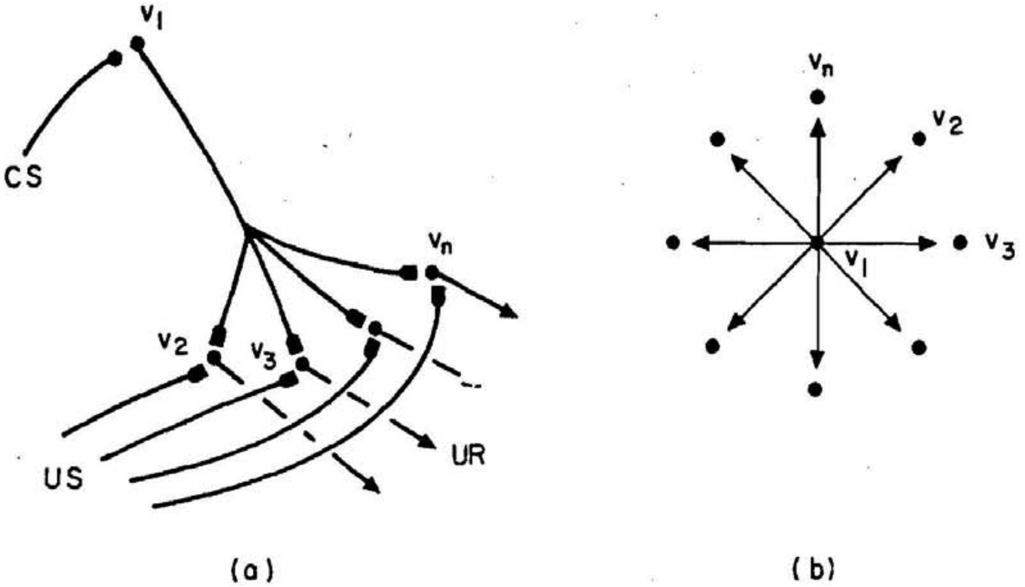

FIG. 9.

$i = 2, 3, \ldots, n$. Suppose for example that the system starts in a state of "maximal ignorance" and the CS and US are presented consecutively for a sufficient amount of time. Then

(1) *Limiting behavior.*

$$\lim_{t\to\infty} X_i(t) = \lim_{t\to\infty} y_{li}(t) = \theta_i .$$

That is, the pattern is eventually learned.

(2) *Oscillations.* The functions $f_i(t) = X_i(t) - \theta_i$, $g_i(t) = y_{li}(t) - X_i(t)$, and $\dot{y}_{li}(t)$ never change sign, so that the approach by $y_{li}(t)$ to θ_i is monotonic (corresponding to our impression that we are learning better and better with increasing practice) even though the inputs to the vertices might fluctuate wildly in time.

(3) *Memory and recall.* If the US is not presented after time $t = T$, then for $t \geq T$,

$$X_i(t) \in [m_i(T), M_i(T)]$$

and

$$y_{li}(t) \in [m_i(T), M_i(T)] ,$$

where

$$m_i(T) = \min[X_i(T), y_{1i}(T)]$$

and

$$M_i(T) = \max[X_i(T), y_{1i}(T)] .$$

That is, if $X_i(T) \simeq \theta_i \simeq y_{1i}(T)$, then the relative associational strength $y_{1i}(t)$ is remembered even during recall trials for all $t \geq T$.

(4) *Stimulus sampling.* If $x_1(t - \tau) \leq \Gamma$, where $\tau = \tau_{1i}$ and $\Gamma = \Gamma_{1i}$ for all $i \neq 1$, then $\dot{y}_{1i}(t) = 0$. That is, no learning occurs unless the spiking frequency at N_{1i} is positive.

To learn the space-time pattern with functions $I_i(t)$, define the weights

$$\theta_i(t) = I_i(t)I^{-1}(t) , \text{ where } I(t) = \sum_{k=2}^{n} I_k(t) .$$

Since $\theta_i(t)$ is a continuous function, it can be approximated by the sequence of its values

$$\{\theta_i(k\xi) : k = 1, 2, \ldots, N_\xi\}$$

if $\xi > 0$ is sufficiently small. Because of the property (4) of stimulus sampling, a sequence of outstars can approximately learn the pattern with weights $\theta_i(t)$ as in Figure 10. In Figure 10, the cell v_1 gives off successive clusters of axon collaterals

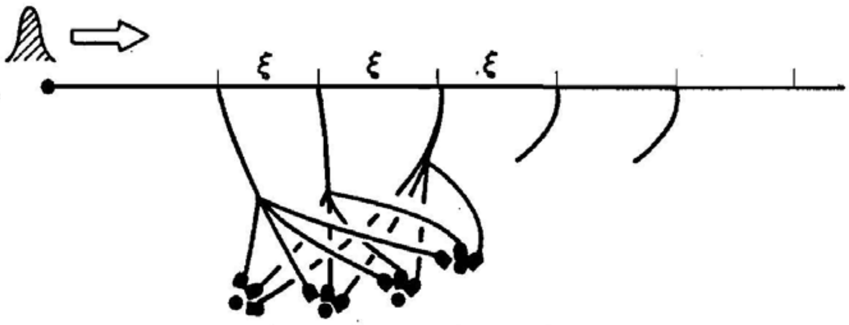

FIG. 10.

to the cells v_i, $i \neq 1$. Each successive cluster is excited ξ time units after the previous cluster. If a brief signal is emitted from v_1, the first cluster learns $\theta(\xi)$, the second cluster $\theta(2\xi)$, and so on. Such a nerve is called an *outstar avalanche.* See [5] and [18] for details.

The stimulus sampling property in general allows thousands of synaptic knobs to end on the same cell body, each synaptic knob perhaps encoding at entirely different time intervals the potentials playing on the cell through time.

6. A General Global Theorem

The following theorem illustrates two anatomies and a large class of closely related dynamical systems in which any number of cells can effectively learn by Pavlovian conditioning. Once these two anatomies are understood, the potentialities for learning of a large collection of related anatomies can be analyzed.

Given any finite sets of indices I and J, such that $I = J$ or $I \cap J = \phi$, consider the following class of functional-differential systems.

$$\dot{x}_i(t) = A(W_t, t)\, x_i(t) + \sum_{k \in J} B_k(W_t, t)\, z_{ki}(t) + C_i(t)\, , \tag{8}$$

and

$$\dot{z}_{ji}(t) = D_j(W_t, t)\, z_{ji}(t) + E_j(W_t, t)\, x_i(t)\, , \tag{9}$$

where $i \in I$ and $j \in J$. W denotes the vector function

$$W = (x_i, z_{ji} : i \in I,\ j \in J)\, ,$$

and W_t designates a functional dependence on all entries of W evaluated possibly at all times $v \leq t$. Thus $A(W_t, t)$ in (8) designates a possibly nonlinear functional of all past values of the system, and perhaps of independent functions of t. $A(W_t, t)$ is however independent of i, and clearly generalizes the exponential decay $-\alpha$ in (6). Were $A(W_t, t)$ to depend on i, we would first decompose the cells v_i with $i \in I$ into maximal subsets such that the A's in each subset were independent of i. Without some restrictions on indices of our functionals, any functional-differential system could be written as in (8) and (9), yielding absurd conclusions.

$B_k(W_t, t)$ in (8) is also possibly a nonlinear functional of W evaluated at all past times, and of independent functions of time. For example, one can choose

$$B_k(W_t, t) = [x_k(t - \tau_k) - \Gamma_k]^+$$

or

$$B_k(W_t, t) = [x_k(t - \tau_k) - \Gamma_k]^+ \left[\sum_{i \in I} z_{ki}(t)\right]^{-1},$$

etc. B_k is chosen independent of i to achieve unbiased learning. This generality in the spiking frequency term allows, for example, such effects as absolute and relative refractory periods in spiking, spontaneous buildup of potential, etc., to be considered. Similar remarks hold for the functionals D_j and E_j. The cells v_i with $i \in I$ thus receive synchronous signals from each fixed cell v_j, $j \in J$. All source cells v_i can, however, be mutually out of phase, even when they mutually interact in the case $I = J$.

The system (8) and (9) is further constrained as follows, and these constraints can readily be verified for the special case $A = -\alpha$; $B_j = [x_j(t - \tau_j) - \Gamma_j]^+ p_j$;

$$D_j = -u_j \quad \text{or} \quad -u_j[x_j(t - \tau_j) - \Gamma_j] \, ;$$

and $E_j = [x_j(t - \tau_j) - \Gamma_j]^+ q_j$, except for condition (4) below, which requires estimates of the numerical parameters.

1) All B_j, E_j, and C_i are nonnegative;
2) $\int_0^\infty B_j(W_t, t)dt = \infty$ only if $\int_0^\infty E_j(W_t, t)dt = \infty$;
3) All functionals A, B_j, E_j, D_j, and inputs C_i are continuous as functions of t;
4) the system is bounded;
5) $C_i(t) = \theta_i C(t)$ with θ_i a fixed probability distribution;

$$\int_T^{t+T} C(v) \exp\left[\int_v^{t+T} A(W_\xi, \xi)d\xi\right] dv \geq K_1$$

if $t \geq K_2$ for all $T \geq 0$ and some positive K_1 and K_2; and

$$\int_0^\infty C(v)dv = \infty \, .$$

Then for arbitrary nonnegative and continuous initial data, the functions

$$X_i(t) = x_i(t)\left[\sum_{k \in I} x_k(t)\right]^{-1}$$

and

$$y_{ji}(t) = z_{ji}(t)\left[\sum_{k \in I} z_{jk}(t)\right]^{-1},$$

defined for $i \in I$ and $j \in J$, have limits Q_i and P_{ji}, respectively, as $t \to \infty$. Moreover, $Q_i = \theta_i$ and, if

$$\int_0^\infty E_j(W_t, t)dt = \infty ,$$

then also $P_{ji} = \theta_i$.

The oscillations of all $X_i(t)$ and $y_{ji}(t)$ can also be completely classified during practice, memory, and recall intervals. The conditions (1)–(5) seem to hold in all known applications, and are clearly very weak. See [16] for further details.

7. Concluding Remarks

The above discussion hopes to suggest that embedding fields have an interesting range of interdisciplinary applications, which embraces various empirical as well as mathematical and philosophical phenomena. This interdisciplinary development permits empirical facts to suggest new mathematical constructions, and conversely. It also brings into the explicit constructs of rigorous science various philosophical observations whose consequences have previously lain dormant or relatively unexplored in our daily lives.

REFERENCES

[1] S. Grossberg, "On the serial learning of lists," *Math. Biosci.*, vol. 4, pp. 201-253, 1969.

[2] S. Grossberg and J. Pepe, "Schizophrenia: possible dependence of associational span, bowing, and primacy vs. recency on spiking threshold, *Behav. Sci.*, vol. 15 (4), pp. 359-362, 1970.

[3] S. Grossberg and J. Pepe, Spiking threshold and overarousal effects in serial learning, *J. of Stat. Physics*, June, 1971.

[4] R. C. Atkinson and W. K. Estes, "Stimulus sampling theory," in: *Handbook of Mathematical Psychology*, New York: John Wiley & Sons (R. D. Luce, R. R. Bush and E. Galanter, eds.), vol. II, chapters 9-14, pp. 121-268, 1963.

[5] S. Grossberg, "Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Studies in Applied Math.*, vol. XLIX (2), pp. 135-166, 1970.

[6] S. Grossberg, "Embedding fields: a theory of learning with physiological implications," *J. Math. Psych.*, vol. 6 (2), pp. 209-239, 1969.

[7] F. Ratliff, *Mach Bands: Quantitative Studies of Neural Networks in the Retina.* San Francisco: Holden-Day, 1965.

[8] S. Grossberg, "On learning, information, lateral inhibition, and transmitters," *Math. Biosci.*, vol. 4, pp. 255-310, 1969.

[9] D. H. Hubel and T. N. Wiesel, *J. Neurophys.*, vol. 28, pp. 229-289, 1965.

[10] H. R. Maturana, J. Y. Lettvin, W. S. McCulloch and W. H. Pitts, *J. Gen. Physiol.*, vol. 43 (Pt. 2), pp. 129-176, 1960.

[11] P. Sterling and B. G. Wickelgren, *J. Neurophys.*, vol. 32, pp. 1-15, 1969.

[12] S. Grossberg, "Neural pattern discrimination," *J. Theoret. Biol.*, vol. 27, pp. 291-337, 1970.

[13] W. L. Kilmer, "The reticular formation," Part II: "The biology of the reticular formation." Interim. Sci. Report No. 3, Division of Engineering Research, Michigan State, 1969.

[14] V. G. Dethier, *Physiology of Insect Senses.* London: Methuen and Co., 1963.

[15] S. Grossberg, "On learning of spatiotemporal patterns by networks with ordered sensory and motor components, 1. Excitatory components of the cerebellum," *Studies in Applied Math.*, vol. XLVIII (2), pp. 105-132, 1969.

[16] S. Grossberg, "On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks," *J. Stat. Physics,* vol. 1 (2), pp. 319-350, 1969.

[17] S. Grossberg, "On the production and release of chemical transmitters and related topics in cellular control," *J. Theoret. Biol.*, vol. 22, pp. 325-364, 1969.

[18] S. Grossberg, "Some networks that can learn, remember, and reproduce any number of complicated space-time patterns I." *J. Math. and Mech.*, vol. 19 (1), pp. 53-91, 1969.

[19] S. Grossberg, "On the global limits and oscillations of a system of nonlinear differential equations describing a flow on a probabilistic network," *J. Diff. Eqns.*, vol. 5 (3), pp. 531-563, 1969.

[20] S. Grossberg, A prediction theory for some nonlinear functional-differential equations," I. *J. Math. Anal. and Applics.*, vol. 21 (3), pp. 643-694, 1968.

[21] C. E. Osgood, Method and Theory in Experimental Psychology. London: Oxford Univ. Press, 1953.

[22] S. Grossberg, "On the dynamics of operant conditioning," *J. Theoret. Biol.*, (in press).