Chapter 9

ASSOCIATIVE AND COMPETITIVE PRINCIPLES OF LEARNING AND DEVELOPMENT: THE TEMPORAL UNFOLDING AND STABILITY OF STM AND LTM PATTERNS

Preface

This Chapter reviews some of the mathematical systems and theorems that arose through analyses of data in a variety of behavioral and neural experimental paradigms. Included are systems and theorems about associative list learning; associative spatial pattern learning; absolutely stable cooperative-competitive decision making and memory storage; and feature discovery, category learning, and pattern recognition by competitive learning networks. A historical mode of exposition has been adopted in order to help bridge the years of specialized investigations in which these theorems were embedded.

Such theorems provide a secure foundation for using in specialized circuits the general learning, recognition, and decision-making mechanisms that they characterize. At the time that the theorems were proved, few workers in the mind and brain sciences appreciated the important role which a mathematical analysis plays in the identification of good physical laws. No one who has succeeded in proving a global limit or oscillation theorem about a nonlinear feedback system with arbitrarily many variables takes the system whose nice properties have admitted such an analysis for granted. By now, mathematical studies of mind and brain have led to the discovery of one of the larger sets of global theorems about nonlinear feedback systems in any number of variables of any scientific discipline. This fact argues against the oft-made claim that neural network theory is still in its infancy.

Armed with such a mathematical understanding, a theorist is much better equipped to successfully carry out quantitative computer simulations of difficult behavioral and neural data. A mathematical understanding also quickly enables one to spot points of weakness in models which seem to work well in a limited domain of applications. Finally, a mathematical understanding enables one to distinguish those model properties which are essentially model-independent or well-understood from properties which add key new elements to the modeling literature.

A study of this mathematical literature reveals that a surprising number of neural models, concepts, and methods which are introduced today as major new contributions are recapitulating results which have been well-known in the neural modeling literature for some time. Since many more people are introducing neural network models today than in the past, this may be an unpopular conclusion, but it is true nonetheless and must be reckoned with to properly appreciate the proven richness of this field, and to enable truly new discoveries to be easily identified and developed.

Competition and Cooperation in Neural Nets
S. Amari and M. Arbib (Eds.)
Lecture Notes in Biomathematics 45: 295–341 (1982)

ASSOCIATIVE AND COMPETITIVE PRINCIPLES OF LEARNING AND DEVELOPMENT: THE TEMPORAL UNFOLDING AND STABILITY OF STM AND LTM PATTERNS

Stephen Grossberg†

1. Introduction: Brain, Behavior, and Babel

This article reviews some principles, mechanisms, and theorems from my work over the past twenty-five years. I review these results here to illustrate their interconnectedness from a recent perspective, to indicate directions for future work, and to reaffirm an approach to theorizing on problems of mind and brain that is still not fashionable despite growing signs that it needs to become so soon.

I say this because, despite the explosive growth of results on the fundamental issues of mind and brain, our science remains organized as a patchwork of experimental and theoretical fiefdoms which rarely interact despite the underlying unity of the scientific problems that they address. The territorial lines that bound these fiefdoms often seem to be as sacrosanct as national boundaries, and for similar cultural and economic reasons. A theorist who succeeds in explaining results from distinct experimental preparations by discovering their unifying mechanistic substrates may, through repeated crossings of these territorial boundaries, start to feel like a traveler without a country, and will often be treated accordingly. My own intellectual travels have repeatedly left me with such a feeling, despite the reassuring belief that theory had provided me with an international passport. To quickly review how some of these territorial passages were imposed by the internal structure of my theory, I will use a personal historical format of exposition, since the familiar territories do not themselves provide a natural descriptive framework.

2. From List Learning to Neural Networks: The Self-Organization of Individual Behavior

My scientific work began unexpectedly in 1957–58 while I was an undergraduate psychology major at Dartmouth College. A great deal of structured data and classical theory about topics like verbal learning, classical and instrumental conditioning,

---

† Supported in part by the Air Force Office of Scientific Research (AFOSR 82-0148) and the National Science Foundation (NSF IST-80-00257).

perceptual dynamics, and attitude change were then available. It struck me that the revolutionary meaning of these data centered in issues concerning the self-organization of individual behavior in response to environmental pressures. I was exhilarated by the dual problems of how one could represent the emergence of behavioral units that did not exist before, and how one could represent the environmental interaction that stimulated this emergence even before the units emerged that would ultimately stabilize this interaction. I soon realized that various data which seemed paradoxical when viewed in terms of traditional concepts seemed inevitable when viewed in a network framework wherein certain laws hold. In fact, the same laws seemed to hold, in one version or another, in all the learning data that I studied. This universality suggested an important role for mathematics to quantitatively classify these various cases, which is why I sit in a mathematics department today. Although the laws were derived from psychological ideas, once derived they readily suggested a neurophysiological interpretation. In fact, that is how I learned my first neurophysiology, and crossed my first major experimental boundaries. To a nineteen-year-old, these heady experiences were motivationally imprinting, and they supplied enough energy to face the sociological difficulties that my blend of psychology, physiology, and mathematics tends to cause. I might add that this interdisciplinary penetration of boundaries by my laws has prevented them from being widely studied by psychologists to the present time, despite the fact that their manifestations have appeared in a vast array of data and specialized models during the past decade.

3. Unitized Nodes, Short Term Memory, and Automatic Activation

The network framework and the laws themselves can be derived in several ways (Grossberg, 1969a, 1974). My first derivation was based on classical list learning data (Grossberg, 1961, 1964) from the serial verbal learning and paired associate paradigms (Dixon and Horton, 1968; Jung, 1968; McGeogh and Irion, 1952; Osgood, 1953; Underwood, 1966). List learning data force one to confront the fact that new verbal units are continually being synthesized as a result of practice, and need not be the obvious units which the experimentalist is directly manipulating (Young, 1968). All essentially stationary concepts, such as the concept of information itself (Khinchin, 1957) hereby become theoretically useless. I therefore find the recent trend to discuss results about human memory in terms of "information processing" misleading (Klatsky, 1980; Loftus and Loftus, 1976; Norman, 1969). Such approaches either implicitly or explicitly adopt a framework wherein the self-organization of new behavioral units cannot be intrinsically characterized. Because these appraoches miss processing constraints that control self-organization, they often construct special-purpose models to explain experiments in which the formation of new units is not too important, or deal indirectly with the self-organization problem by using computer models that would require a homunculus to carry out their operations in a physical setting. I will clarify these assertions as I go along.

By putting the self-organization of individual behavior in center stage, I realized that the phenomenal simplicity of familiar behavioral units, and the evolutionary aggregation of these units into new representations which themselves achieve phenomenal simplicity through experience, should be made a fundamental property of my theory. To express the phenomenal simplicity of familiar behavioral units, I represented them by indecomposable internal representations, or unitized nodes, rather than as composites of phonemes or as individual muscle movements. The problem of how phonemic, syllabic, and word-like representations might all coexist with different importance in different learning contexts was hereby vividly raised.

Once unitized nodes were conceived, it became clear that experimental inputs can activate these nodes via conditionable pathways. A distinction between sensory activation (the input source) and short term memory (the node's reaction) hereby became natural, as well as a concept of "automatic" activation of a node by its input. These network concepts have become popular in psychology during the past decade under the pressure of recent data (e.g., Schneider and Shiffrin, 1976), but they were already needed to analyze classical list learning data that are currently out of fashion.

The following properties of list learning helped to constrain the form of my associative laws. To simplify the discussion, I will only consider associative interactions within a given level in a coding hierarchy, rather than the problem of how coding hierarchies develop and interact between several levels. All of my conclusions can be, and have been, generalized to a hierarchical setting (Grossberg, 1974, 1978a, 1980a).

4. Backward Learning and Serial Bowing

Backward learning effects and, more generally, error gradients between nonadjacent, or remote, list items (Jung, 1968; McGeogh and Irion, 1952; Murdock, 1974; Osgood, 1953; Underwood, 1966) suggest that pairs of nodes v_i and v_j can interact via distinct directed pathways e_{ij} and e_{ji} over which conditionable signals can travel. Indeed, an analysis of how any node v_i can know where to send its signals reveals that no local information exists at the node itself whereby such a decision can be made. By the principle of sufficient reason, the node must therefore send signals towards all possible nodes v_j with which it is connected by directed paths e_{ij}. Some other variables must exist which can discriminate which combination of signals should reach their target nodes based on past experience. These auxiliary variables turn out to be the long term memory traces. The concept that each node sends out signals to all possible nodes has recently appeared in models of *spreading activation* (Collins and Loftus, 1975; Klatsky, 1980) to explain semantic recognition and reaction time data.

The form that the signaling and conditioning laws should take is forced by data about serial verbal learning. A main paradox about serial learning concerns the form of the bowed serial position curve which relates cumulative errors to list positions (Figure 1a). This curve is paradoxical for the following reason. If all that happens during serial learning is a build-up of various types of interference at each list position due to the occurrence of prior list items, then the error curve should be monotone increasing (Figure 1b). Because the error curve is bowed, and the degree of bowing depends on the length of the intertrial interval between successive list presentations, the *nonoccurrence* of list items after the last item occurs somehow improves learning across several prior list items.

5. The Inadequacy of Rehearsal as an Explanatory Concept

Just saying that rehearsal during the intertrial interval causes this effect does not explain it, because it does not explain why the middle of the list is less rehearsed. Indeed the middle of the list has more time to be rehearsed than does the end of the list before the next learning trial occurs. In the classical literature, one reads that the middle of the list experiences maximal proactive interference (from prior items) and retroactive interference (from future items), but this just labels what we have to explain (Jung, 1953; Osgood, 1953; Underwood, 1966). In the more recent literature, rehearsal is given a primary role in determining the learning rate (Rundus, 1971) although it is believed that only certain types of rehearsal, called *elaborative* rehearsal, can accomplish this (Bjork, 1975; Craik and Watkins, 1973; Klatsky, 1980). Notwithstanding the type of rehearsal used, one still has to explain why the list middle is rehearsed less than the list end in the serial learning paradigm.

The severity of such difficulties led the serial learning expert Young (1968) to write: "If an investigator is interested in studying verbal learning processes ... he would do well to choose some method other than serial learning" (p.146). Another leading verbal learning expert Underwood (1966) realized the magnitude of the difficulties, but also that they would not go away by ignoring them, when he wrote: "The person who

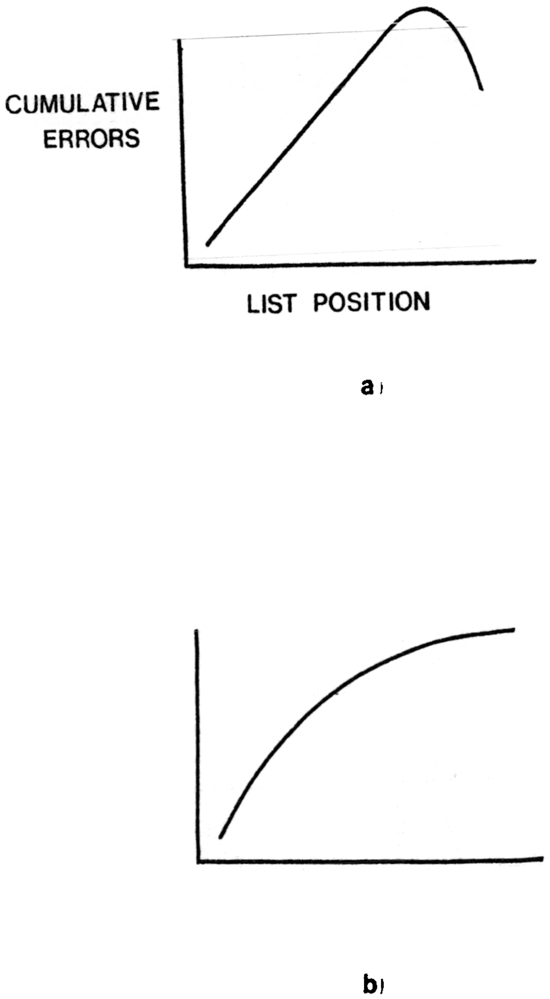

Figure 1. (a) The cumulative error curve in serial verbal learning is a skewed bowed curve. Items between the middle and end of the list are hardest to learn. Items at the beginning of the list are easiest to learn. (b) If position-dependent difficulty of learning were all due to interference from previously presented items, the error curve would be monotone increasing.

originates a theory that works out to almost everyone's satisfaction will be in line for an award in psychology equivalent to the Nobel prize" (p.491).

Most recent investigators have followed Young's advice. They have turned to paradigms like free recall (Bower, 1977; Murdock, 1974) wherein single trial presentations minimize self-organizing effects and subject-determined recall strategies simplify the interactions between item recall and retrieval probes. However, analysis of the free recall paradigm has not proved deep enough to explain the serial bowed curve. In particular, one cannot resort to the type of free recall explanations which are used to explain the bowed effects in that paradigm (Atkinson and Shiffrin, 1968), since the improvement in recall at the end of a serially learned list is due to long term memory rather than to short term memory. Indeed, I have elsewhere argued that popular free recall theories contain internal problems of a homuncular nature, cannot explain some critical free recall data concerning primacy effects in STM which are not supposed to exist, and cannot even explain how a telephone number can be accurately repeated out of STM, because they do not address various issues which are also raised by serial learning data (Grossberg, 1978b).

6. The Inadequacy of Programmatic Time

The massive backward effect that causes the bowed serial curve forces the use of a real-time theory that can parameterize the temporal unfolding of both the occurrences and the nonoccurrences of events. The bowed curve hereby weakens the foundations of all theories whose time variable is counted in terms of computer program operations, no matter how successful these theories might be in simulating data via homuncular constructions (Anderson and Bower, 1973). The existence of facilitative effects due to nonoccurring items also shows that traces of prior list occurrences must endure beyond the last item's presentation time, so they can be influenced by the future nonoccurrences of items. This fact leads to the concept of stimulus traces, or short term memory (STM) traces, $x_i(t)$ at the nodes v_i, $i = 1, 2, \ldots, n$, which are activated by inputs $I_i(t)$, but which decay at a rate slower than the input presentation rate.

Thus in response to serial inputs, *patterns* of STM activity are set up across the network's nodes. By sufficient reason, each supraliminally activated node also sends signals along all its directed pathways. The combination of serial inputs, distributed internodal signals, and spontaneous STM changes at each node changes the STM pattern as the experiment proceeds. A major task of learning theory is to characterize the rules whereby these STM patterns evolve through time. Indeed, a major mathematical task is to learn how to think in terms of pattern transformations, rather than just in terms of feature detectors or other local entities.

7. Network versus Computer Parsing: Distinct Error Gradients at Different List Positions

The general philosophical interest of the bowed curve can be better appreciated by asking: What is the first time a learning subject can possibly know that item r_n is the last list item in a newly presented list $r_1 r_2 \ldots r_n$, given that a new item is presented every w time units until r_n occurs? The answer obviously is: not until at least w time units *after* r_n has been presented. Only after this time passes and no item r_{n+1} is presented can r_n be correctly reclassified from the list's "middle" to the list's "end". Since parameter w is under experimental control and is not a property of the list ordering *per se*, spatiotemporal network interactions parse a list in a way that is fundamentally different from the parsing rules that are natural to apply to a list of symbols in a computer. Indeed, increasing the intratrial interval w during serial learning can flatten the entire bowed error curve and minimize the effects of the intertrial interval (Jung, 1968; Osgood, 1953).

To illustrate the difference between computer models and my network approach, suppose that after a node v_i is excited by an input I_i, its STM trace gets smaller through time due to either internodal competition or to passive trace decay. Then in response to a serially presented list, the last item to occur always has the largest STM trace—in other words, at every time a *recency* gradient obtains in STM (Figure 2). Given this natural assumption—which, however, is not always true (Grossberg, 1978a, 1978b)—how do the generalization gradients of errors at each list position get learned (Figure 3)? In particular, how does a gradient of anticipatory errors occur at the beginning of the list, a two-sided gradient of anticipatory and perseverative errors occur near the middle of the list, and a gradient of perseverative errors occur at the end of the list (Osgood, 1953)? Otherwise expressed, how does a temporal succession of STM recency gradients generate an LTM *primacy* gradient at the list beginning but an LTM *recency* gradient at the list end? These properties immediately rule out any linear theory, as well as any theory which restricts itself to nearest neighbor associative links, unless the theory makes the homuncular assumption that the system has absolute knowledge of how to compute the list's beginning, end, and direction towards its middle (Feigenbaum and Simon, 1962).

8. Graded STM and LTM Patterns: Multiplicative Sampling and Slow Decay by LTM Traces

Figures 2 and 3 can be reconciled by positing the existence of STM traces and LTM traces that evolve according to different time scales and rules. Indeed, this reconciliation is one of the strongest arguments that I know for these rules.

Suppose, as above, that each node v_j can send out a sampling signal S_j along each directed path e_{jk} towards the node v_k, $j \neq k$. Suppose that each path e_{jk} contains a long term memory (LTM) trace z_{jk} at its terminal point, where z_{jk} can compute, using only local operations, the product of signal S_j and STM trace x_k. Also suppose that the LTM trace decays slowly, if at all, during a single learning trial. The simplest law for z_{jk} that satisfies these constraints is

$$\frac{d}{dt} z_{jk} = -c z_{jk} + d S_j x_k, \quad j \neq k. \tag{1}$$

To see how this rule generates an LTM primacy gradient at the list beginning, we need to study the LTM pattern $(z_{12}, z_{13}, \ldots, z_{1n})$ and to show that $z_{12} > z_{13} > \ldots > z_{1n}$. To see how the same rule generates an LTM recency gradient at the list end, we need to study the LTM pattern $(z_{n1}, z_{n2}, \ldots, z_{n,n-1})$ and to show that $z_{n1} < z_{n2} < \ldots < z_{n,n-1}$. The two-sided gradient at the list middle can then be understood as a combination of these effects.

By (1), node v_1 sends out a sampling signal S_1 shortly after item r_1 is presented. After rapidly reaching peak size, signal S_1 gradually decays as future list items $r_2, r_3, \ldots$ are presented. Thus S_1 is largest when trace x_2 is maximal, S_1 is smaller when both traces x_2 and x_3 are active, S_1 is smaller still when traces x_2, x_3, and x_4 are active, and so on. Consequently, the product $S_1 x_2$ in row 2 of Figure 2 exceeds the product $S_1 x_3$ in row 3 of Figure 2, which in turn exceeds the product $S_1 x_4$ in row 4 of Figure 2, and so on. Due to the slow decay of each LTM trace z_{1k} on each learning trial, z_{12} adds up to the products $S_1 x_2$ in successive rows of column 1, z_{13} adds up to the products $S_1 x_3$ in successive rows of column 2, and so on. An LTM primacy gradient $z_{12} > z_{13} > \ldots > z_{1n}$ is hereby generated. This gradient is due to the way signal S_1 multiplicatively *samples* the successive STM recency gradients and the LTM traces z_{1k} sum up the sampled STM gradients.

By contrast, the signal S_n of a node v_n at the end of the list samples a different set of STM gradients. This is because v_n starts to sample (viz., $S_n > 0$) only after

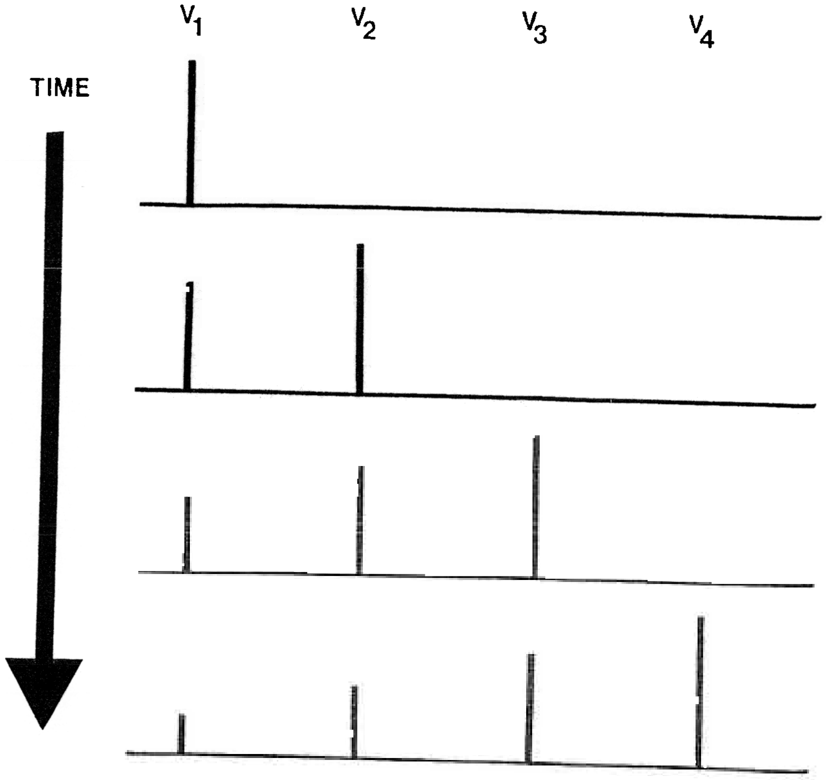

Figure 2. Suppose that items $r_1, r_2, r_3, r_4, \ldots$ are presented serially to nodes $v_1, v_2, v_3, v_4, \ldots$, respectively. Let the activity of node v_i at time t be described by the height of the histogram beneath v_i at time t. If each node is initially excited by an equal amount and its excitation decays at a fixed rate, then at every time (each row) the pattern of STM activity across nodes is described by a recency gradient.

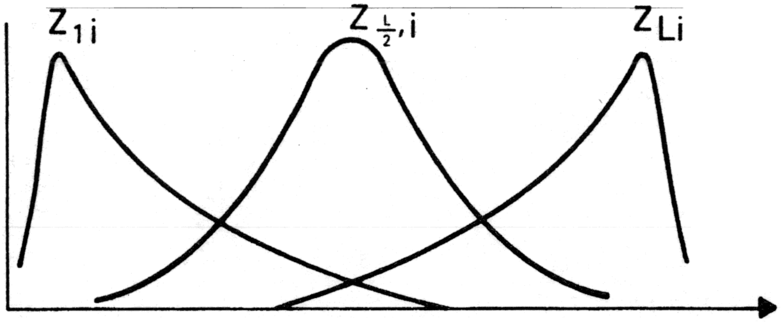

Figure 3. At each node v_j, the LTM pattern $z_j = (z_{j1}, z_{j2}, \ldots, z_{jn})$ that evolves through time is different. In a list of length $n = L$ whose intertrial interval is sufficiently long, the LTM pattern at the list beginning ($j \cong 1$) is a primacy gradient. At the list end ($j \cong L$), a recency gradient evolves. Near the list middle ($j \cong \frac{L}{2}$), a two-sided gradient is learned. These gradients are reflected in the distribution of anticipatory and perseverative errors in response to item probes at different list positions.

all past nodes $v_1, v_2, \ldots, v_{n-1}$ have already been activated on that trial. Consequently, the LTM traces $(z_{n1}, z_{n2}, \ldots, z_{n,n-1})$ of node v_n encode a recency gradient $x_1 < x_2 < x_3 < \ldots < x_{n-1}$ at *each* time. When all the recency gradients are added up through time, the total effect is a recency gradient in v_n's LTM pattern. In summary, nodes at the beginning, middle, and end of the list encode different LTM gradients because they multiplicatively sample and store STM patterns at different times.

Lest the reader who is sensitized to the functional unit issue object to these internodal feedback effects, let me reiterate that similar LTM gradients obtain if the sequences of nodes which are active at any time selectively excite higher-order nodes (chunks) which in turn sample the field of excited nodes via feedback signals (Grossberg, 1974, 1978a).

9. Binary versus Continuous Associative Laws

The LTM gradient problem illustrates why I have always avoided binary laws for STM and LTM traces. Binary laws have often attracted workers who began with the all-or-none character of individual axon spikes. However, the continuously fluctuating potentials that receive these spikes often average them in time, thereby yielding graded intercellular signaling effects. For similar reasons, population interactions often obey continuous laws. Workers like Amari (1974, 1977) and Geman (1981) have formally studied how to justify the averaging procedures that can convert binary microscopic rules into continuous macroscopic rules. Because of the psychological derivation of my networks, I have always worked with preaveraged equations from the start.

The example of continuous LTM error gradients is not the only one wherein binary and continuous models yield distinct outcomes. In fact, they usually do. For example, just changing sigmoid feedback signals to binary feedback signals in a competitive network can significantly change network dynamics (Grossberg, 1973, 1978c), notably because sigmoid signals can support infinitely many equilibrium points in competitive geometries wherein binary signals cannot.

10. Retrieval Probes and LTM Gating of STM Mediated Signals

Having shown how STM patterns may be read into LTM patterns, we now need to describe how a retrieval probe r_m can read v_m's LTM pattern back into STM on recall trials, whereupon the STM traces can be transformed into observable behavior. In particular, how can LTM be read into STM without distorting the learned LTM gradients?

The simplest rule generates an STM pattern which is proportional to the LTM pattern that is being read out, and allows distinct probes to each read their LTM patterns into STM in an independent fashion.

To achieve faithful read-out of the LTM pattern $(z_{m1}, z_{m2}, \ldots, z_{mn})$ by a probe r_m that turns on signal S_m, I let the product $S_m z_{mi}$ determine the growth rate of x_i. In other words, LTM trace z_{mi} *gates* the signal S_m along e_{mi} before the gated signal reaches v_i. The independent action of several probes implies that the gated signals $S_m z_{mi}$ are added, so that the total effect of all gated signals on v_i is $\sum_{m=1}^{n} S_m z_{mi}$. The simplest equation for the STM trace x_i that abides by this rule is

$$\frac{d}{dt} x_i = -a x_i + b \sum_{m=1}^{n} S_m z_{mi} + I_i, \tag{2}$$

where $-a$ is the STM decay rate that produces Figure 2, S_m is the mth sampling signal, z_{mi} is the LTM trace of pathway e_{mi}, and I_i is the ith experimental input.

The reaction of equations (1) and (2) to serial inputs I_i is much more complex than is their response to an isolated retrieval probe r_m. Due to the fact that STM traces may decay slower than the input presentation rate, several sampling signals S_m can be simultaneously active, albeit in different phases of their growth and decay.

11. Behavioral Choices and Competitive Feedback

Once one accepts that patterns of STM traces are evolving through time, one also needs a mechanism for choosing those activated nodes which will influence observable behavior. Lateral inhibitory feedback signals are readily implicated as a choice mechanism (Grossberg, 1968, 1969b, 1970). The simplest extension of (2) which includes competitive interactions is

$$\frac{d}{dt} x_i = -a x_i + \sum_{m=1}^{n} S_m^+ b_{mi}^+ z_{mi} - \sum_{m=1}^{n} S_m^- b_{mi}^- + I_i \tag{3}$$

where $S_m^+ b_{mi}^+$ $(S_m^- b_{mi}^-)$ is the excitatory (inhibitory) signal emitted from node v_m along the excitatory (inhibitory) pathway e_{mi}^+ (e_{mi}^-). Correspondingly, equation (1) is generalized to

$$\frac{d}{dt} z_{jk} = -c z_{jk} + d_{jk} S_j^+ x_k. \tag{4}$$

The asymmetry between terms $\sum_{m=1}^{n} S_m^+ b_{mi}^+ z_{mi}$ and $\sum_{m=1}^{n} S_m^- b_{mi}^-$ in (3) readily suggests a modification of (3) and a definition of inhibitory LTM traces analogous to (4), where such traces exist (Grossberg, 1969d).

Because lateral inhibition can change the sign of each x_i from positive to negative in (3), and thus change the sign of each z_{jk} from positive to negative in (4), some refinements of (3) and (4) are needed to prevent absurdities like the following: $S_m^+ < 0$ and

$x_i < 0$ implies $z_{mi} > 0$; and $S_m^+ < 0$ and $z_{mi} < 0$ implies $x_i > 0$. Threshold constraints accomplish this in the simplest way. Letting $[\xi]^+ = \max(\xi, 0)$, these absurdities are prevented if threshold cut-offs are imposed on signals, such as in

$$S_j^+ = [x_j(t - \tau_j^+) - \Gamma_j^+]^+ \tag{5}$$

and

$$S_j^- = [x_j(t - \tau_j^-) - \Gamma_j^-]^+, \tag{6}$$

as well as on sampled STM traces, such as in

$$\frac{d}{dt} z_{jk} = -c z_{jk} + d_{jk} S_j^+ [x_k]^+. \tag{7}$$

The equations (3), (5), (6), and (7) have been used by modelers for a variety of purposes. For example, in his seminal article on code development, Malsburg (1973) used these equations, supplemented by his synaptic conservation rule.

12. Skewing of the Bow: Symmetry-Breaking Between the Future and the Past

To explain the bowed error curve, we now need to compare the LTM patterns $z_i = (z_{i1}, z_{i2}, \ldots, z_{in})$ which evolve at all list nodes v_i. In particular, we need to explain why the bowed curve is *skewed*; that is, why the list position where learning takes longest occurs nearer to the end of the list than to its beginning (Figure 1a). This skewing effect has routinely demolished learning theories which assume that forward and backward effects are equally strong, or symmetric (Asch and Ebenholtz, 1962; Murdock, 1974). I have elsewhere argued that the symmetry-breaking between the future and the past, by favoring forward over backward associations, makes possible the emergence of a global "arrow in time," or the ultimate learning of long event sequences in their correct order (Grossberg, 1969c, 1974).

Theorem 1 below asserts that a skewed bowed curve does occur in the network, and predicts that the degree of skewing will decrease and the relative learning rate at the beginning and end of the list will reverse as the network's arousal level increases or its excitatory signal thresholds Γ_j^+ decrease to abnormal levels (Grossberg and Pepe, 1970, 1971). The arousal and threshold predictions have not yet been tested to the best of my knowledge. They are of some conceptual importance because abnormally high arousal or low thresholds can hereby generate a formal network syndrome characterized by contextual collapse, reduced attention span, and fuzzy response categories that resembles some symptoms of simple schizophrenia (Grossberg and Pepe, 1970; Maher, 1977).

To understand what is involved in my explanation of bowing, note that by equation (7), each correct LTM trace $z_{12}, z_{23}, z_{34}, \ldots, z_{n-1,n}$ may grow at a comparable rate, albeit w time units later than the previous correct LTM trace. However, the LTM patterns $z_1, z_2, \ldots, z_n$ will differ no matter when you look at them, as in Figure 3. Thus when a retrieval probe r_j reads its LTM pattern z_j into STM, the entire pattern must influence overt behavior to explain why bowing occurs. The relative size of the correct LTM trace $z_{j,j+1}$ compared to all other LTM traces in z_j will influence its success in eliciting r_{j+1} after competitive STM interactions occur. A larger $z_{j,j+1}$ relative to the sum of all other z_{jk}, $k \neq j, j+1$, should yield better performance of r_{j+1} given r_j, other things being equal. To measure the distinctiveness of a trace z_{jk} relative to all traces in z_j, I therefore define the relative LTM traces, or stimulus sampling probabilities

$$Z_{jk} = z_{jk}\Big(\sum_{m \neq j} z_{jm}\Big)^{-1}. \tag{8}$$

The appropriateness of definition (8) is strengthened by the following observation. The ordering within the LTM gradients of Figure 3 is unchanged by the relative LTM traces; for example, if $z_{12} > z_{13} > \ldots > z_{1n}$, then $Z_{12} > Z_{13} > \ldots > Z_{1n}$ because all the Z_{1k}'s have the same denominator. Thus all conclusions about LTM gradients are valid for relative LTM gradients.

In terms of the relative LTM traces, the issue of bowing can be mathematically formulated as follows. Define the *bowing function* $B_i(t) = Z_{i,i+1}(t)$. Function $B_i(t)$ measures how distinctive the ith correct association is at time t. After a list of n items is presented with an intratrial interval w and a sufficiently long intertrial interval W elapses, does the function $B_i((n-1)w + W)$ decrease and then increase as i increases from 1 to n? Does the minimum of the function occur in the latter half of the list? The answer to both of these questions is "yes."

To appreciate the subtlety of the bowing issue, it is necessary to understand how the bow depends upon the ability of a node v_i to sample incorrect future associations, such as $r_i r_{i+2}, r_i r_{i+3}, \ldots$ in addition to incorrect past associations, such as $r_i r_{i-1}, r_i r_{i-2}, \ldots$. As soon as S_i becomes positive, v_i can sample the entire past field of STM traces at $v_1, v_2, \ldots, v_{i-1}$. However, if the sampling threshold is chosen high enough, S_i might shut off before r_{i+2} occurs. Thus the sampling duration has different effects on the sampling of past than future incorrect associations. For example, if the sampling thresholds of all v_i are chosen so high that S_i shuts off before r_{i+2} is presented, then the function $B_i(\infty)$ decreases as i increases from 1 to n. In other words, the monotonic error curve of Figure 1b obtains because no node v_i can encode incorrect future associations.

Even if the thresholds are chosen so that incorrect future associations can be formed, the function $B_i((i+1)w)$ which measures the distinctiveness of $z_{i,i+1}$ just before r_{i+2} occurs is again a decreasing function of i. The bowing effect thus depends on threshold choices which permit sampling durations that are at least $2w$ in length.

The shape of the bow also depends on the duration of the intertrial interval, because before the intertrial interval occurs, all nodes build up increasing amounts of associative interference as more list items are presented. The first effect of the nonoccurrence of items after r_n is presented is the growth through time of $B_{n-1}(t)$ as t increases beyond the time nw when item r_{n+1} would have occurred in a larger list (Grossberg, 1969c). The last correct association is hereby facilitated by the absence of interfering future items during the intertrial interval. This facilitation effect is a nonlinear property of the network. Indeed, bowing itself is a nonlinear phenomenon in my theory, because it depends on a comparison of ratios of integrals of sums of products as they evolve through time.

In my review of a bowing theorem below, I will emphasize the effect of the signal threshold Γ on the degree of skewing. One can, however, also compute the effect of the intertrial interval W on skewing, as well as the role of other network parameters, such as STM decay rate and LTM growth rate.

The position of the bow has not yet been quantitatively computed although it has been qualitatively demonstrated within the full system (3), (5), (6), (7). Complete computations have been made in a related system, the *bare field*, wherein the primary effects of serial inputs on associative formation and competition are preserved (Grossberg, 1969c; Grossberg and Pepe, 1971) on a single learning trial. In the bare field, serial inputs occur with intratrial interval w:

$$I_1(t) = I_2(t + w) = \qquad = I_n(t + (n-1)w); \tag{9}$$

the STM traces decay after they are excited by their inputs:

$$\frac{d}{dt} x_i = -a x_i + I_i; \tag{10}$$

the LTM traces add up products of signals and STM traces:

$$\frac{d}{dt}z_{jk} = d[x_j(t-\tau)-\Gamma]^+ x_k, \quad j \neq k; \tag{11}$$

and the relative LTM traces, or stimulus sampling probabilities, estimate how well as given LTM trace fares after it is read into STM and STM competition takes place:

$$Z_{jk} = z_{jk}\Big(\sum_{m\neq j} z_{jm}\Big)^{-1}. \tag{12}$$

Theorem 1 (Skewing of the Bowed Curve):

(I) If the bare field is initially at rest and associatively unbiased; that is, all

$$x_i(t)=0, \quad -\tau \leq t \leq 0, \quad \text{and} \quad z_{jk}(0)=\alpha>0, \quad j\neq k; \tag{13}$$

(II) the signals S_i and inputs r_{i+1} are well-correlated; that is

$$W=\tau \tag{14}$$

(this condition is convenient but not essential);

(III) successive inputs do not overlap in time; that is, $I_1(t)$ is positive only in an interval $(0,\lambda)$ with $\lambda<\tau$ and is zero elsewhere;

(IV) the inputs are not too irregular; that is, $I_1(t)$ is continuous and grows monotonically until it reaches a maximum at time $t=T_{\max}$, after which it monotonically decreases to zero at time $t=\lambda$;

(V) at high threshold, the sampling signals don't last too long; that is, if Γ is chosen so large that v_1 first emits a signal S_1 at the time $T_{\max}$, then S_1 shuts off before r_3 occurs: if $\Gamma_0=\int_0^{T_{\max}} e^{-a(T_{\max}-v)}I_1(v)dv$, then

$$\int_0^{\lambda} e^{-a(\lambda-v)}I_1(v)dv > \Gamma_0 \geq \int_0^{2\tau} e^{-a(2\tau-v)}I_1(v)dv. \tag{15}$$

Under hypotheses (I)–(V), if the intertrial interval is infinite, then the bow occurs ($B_i(\infty)$ is minimized) at the list position closest to $M(\Gamma)$, where

A. (Overaroused Bowing)

$$M(0)=\frac{1}{2}(n-1), \tag{16}$$

B. (Skewing)

$$\frac{dM}{d\Gamma}>0, \tag{17}$$

C. (No Incorrect Future Associations)

$$M(\Gamma)=n \text{ if } \Gamma\geq\Gamma_0. \tag{18}$$

If the intertrial interval is $W<\infty$, then the bow occurs ($B_i((n-1)w+W)$ is minimized) at a list position strictly greater than $M(\Gamma)$.

The function $M(\Gamma)$ can, moreover, be explicitly computed. It satisfies the equation

$$M(\Gamma)=\frac{1}{a\tau}\log\Big[\frac{E+\sqrt{E^2+4CD}}{2D}\Big] \tag{19}$$

where

$$C=\tau E^{-1}(-\tau,0)[AB(\lambda-T_1,T_1)+A\Gamma E(T_2,T_1)+\frac{A^2}{2}E(2\lambda,2T_2)-\frac{\Gamma^2}{2a}E(-L)], \tag{20}$$

$$D=A\tau E(L)E^{-1}(0,\tau)[B(\lambda,0)+\frac{A}{2a}e^{-2a\lambda}], \tag{21}$$

$$E=\Gamma[C(\lambda,0)+\frac{A}{a}e^{-a\lambda}], \tag{22}$$

with

$$A=\int_0^{\lambda} e^{av}I_1(v)dv, \tag{23}$$

$$B(t,p)=\int_0^t e^{-2a(v+p)}\int_0^{v+p} e^{aw}I_1(w)dwdv, \tag{24}$$

$$C(t,p)=\int_0^t e^{-a(v+p)}\int_0^{v+p} e^{aw}I_1(w)dwdv, \tag{25}$$

$$E(x)=e^{-a\tau x}, \tag{26}$$

and

$$E(x,y)=\frac{1}{a}(e^{-ax}-e^{-ay}). \tag{27}$$

13. Evolutionary Invariants of Associative Learning: Absolute Stability of Parallel Pattern Learning

Many features of system (3), (5), (6), (7) are special; for example, the exponential decay of STM and LTM and the signal threshold rule. Because associative processing is ubiquitous throughout phylogeny and within functionally distinct subsystems of each individual, a more general mathematical framework is needed. This framework should distinguish universally occurring associative principles which guarantee essential learning properties from evolutionary variations that adapt these principles to specialized environmental demands. Before we can speak with confidence about variations on an evolutionary theme, we first need to identify the theme.

I approached this problem during the years 1967 to 1972 in a series of articles wherein I gradually realized that the mathematical properties that I used to globally analyze specific learning examples were much more general than the examples themselves. This work culminated in my universal theorems on associative learning (Grossberg, 1969d, 1971a, 1972a).

The theorems are universal in the following sense. They say that if certain associative laws were invented at a prescribed time during evolution, then they could achieve unbiased associative pattern learning in essentially any later evolutionary specialization. To the question: Is it necessary to re-invent a new learning rule to match every perceptual or cognitive refinement, the theorems say "no". More specifically, the universal associative laws enable arbitrary spatial patterns to be learned by arbitrarily many, simultaneously active sampling channels that are activated by arbitrary continuous data preprocessing in an essentially arbitrary anatomy. Arbitrary space-time patterns can also be learned given modest constraints on the temporal regularity of stimulus sampling. The universal theorems thus describe a type of parallel processing whereby unbiased pattern learning can occur despite mutual crosstalk between very complex feedback signals.

Such results cannot be taken for granted. They obtain only if crucial network operations, such as spatial averaging, temporal averaging, preprocessing, gating, and cross-correlation are computed in a canonical ordering. This canonical ordering constitutes a general purpose design for unbiased parallel pattern learning, as well as a criterion for whether particular networks are acceptable models for this task. The universality of the design mathematically takes the form of a classification of oscillatory and limiting possibilities that is invariant under evolutionary specializations.

The theorems can also be interpreted in another way that is appropriate in discussions of self-organizing systems. The theorems are *absolute stability* theorems. They show that evolutionary invariants obtain no matter how system parameters are changed within this class of systems. Absolute stability is an important property in a self-organizing system because parameters may change in ways that cannot be predicted in advance, notably before specialized environments act on the system. Absolute stability guarantees that the onset of self-organization does not subvert the very properties which make self-organization possible.

The systems which I considered have the form

$$\frac{d}{dt}x_i = A_i x_i + \sum_{k \in J} B_{ki} z_{ki} + C_i(t) \tag{28}$$

$$\frac{d}{dt}z_{ji} = D_{ji} z_{ji} + E_{ji} x_i \tag{29}$$

where $i \in I$, $j \in J$, and I and J parameterize arbitrarily large, not necessarily disjoint, sets of sampled and sampling cells, respectively. As in my equations for list learning, A_i is an STM decay rate, B_{ki} is a nonnegative performance signal, $C_i(t)$ is an input function, D_{ji} is an LTM decay rate, and E_{ji} is a nonnegative learning signal. Unlike the list learning equations, A_i, B_{ki}, D_{ji}, and E_{ji} are continuous functionals of the entire history of the system. Equations (28) and (29) are therefore very general, and include many of the specialized models in the literature.

For example, although (28) does not seem to include inhibitory interactions, such interactions may be lumped into the STM decay functional A_i. The choice

$$A_i = -a_i + (b_i - c_i x_i)G_i(x_i) - \sum_{k=1}^{n} H_k(x_k) d_{ki} \tag{30}$$

describes the case wherein system nodes compete via shunting, or membrane equation, interactions (Cole, 1968; Grossberg, 1973; Kuffler and Nicholls, 1976; Plonsey, 1969). The performance, LTM decay, and learning functionals may include slow threshold changes, nonspecific Now Print signals, signal velocity changes, presynaptic modulating effects, arbitrary continuous rules of dendritic preprocessing and axonal signaling, as well as many other possibilities (Grossberg, 1972a, 1974). Of special importance are the variety of LTM decay choices that satisfy the theorems. For example, an LTM law like

$$\frac{d}{dt}z_{ji} = [x_j(t - \tau_j) - \Gamma_j(y_t)]^+(-d_j z_{ji} + e_j x_i) \tag{31}$$

achieves an interference theory of forgetting, rather than exponential forgetting, since $\frac{d}{dt}z_{ji} = 0$ except when v_j is sampling (Adams, 1967). Equation (31) also allows the vigor of sampling to depend on changes in the threshold $\Gamma_j(y_t)$ that are sensitive to the prior history $y_t = (x_i, z_{ji} : i \in I, j \in J)_t$ of the system before time t.

In this generality, too many possibilities exist to as yet prove absolute stability theorems. One further constraint on system processing paves the way towards such results. This constraint still admits the above processing possibilities, but it imposes some spatiotemporal regularity on the sampling process. Indeed, if the performance signals B_{ji} from a fixed sampling node v_j to all the sampled nodes v_i, $i \in I$, were arbitrary nonnegative and continuous functionals, then the irregularities in each B_{ji} could override any regularities in z_{ji} within the gated performance signal $B_{ji}z_{ji}$ from v_j to v_i.

14. Local Symmetry and Self-Similarity in Pattern Learning and Developmental Invariance

Absolute stability does obtain even if different functionals B_j, D_j, and E_j are assigned to each node v_j, $j \in J$, just so long as the same functional is assigned to all pathways e_{ji}, $i \in I$. Where this is not globally true, one can often partition the network into maximal subsets where it is true, and then prove unbiased pattern learning in each subset. This restriction is called the property of *local symmetry axes* since each sampling cell v_j can act as a source of coherent history-dependent waves of STM and LTM processing. Local symmetry axes still permit (say) each B_j to obey different history-dependent preprocessing, threshold, time lag, and path strength laws among arbitrarily many mutually interacting nodes v_j.

When local symmetry axes are imposed on (28) and (29), the resulting class of systems takes the form

$$\frac{d}{dt}x_i = Ax_i + \sum_{k \in J} B_k z_{ki} + C_i(t) \tag{32}$$

and

$$\frac{d}{dt}z_{ji} = D_j z_{ji} + E_j x_i. \tag{33}$$

A simple change of variables shows that constant interaction coefficients b_{ji} between pairs v_j and v_i of nodes can depend on $i \in I$ without destroying unbiased pattern learning in the systems

$$\frac{d}{dt}x_i = Ax_i + \sum_{k \in J} B_k b_{ki} z_{ki} + C_i(t) \tag{34}$$

and

$$\frac{d}{dt}z_{ji} = D_j z_{ji} + E_j b_{ji}^{-1} x_i. \tag{35}$$

By contrast, the systems (34) and

$$\frac{d}{dt}z_{ji} = D_j z_{ji} + E_j b_{ji} x_i \tag{36}$$

are not capable of unbiased parallel pattern learning (Grossberg, 1972a). A dimensional analysis shows that (34) and (35) hold if action potentials transmit the network's intercellular signals, whereas (34) and (36) hold if electrotonic propagation is used.

The dimensional analysis hereby suggests that spatial biases in the b_{ji} which are due to differences in axonal diameters can be overcome by an interaction between action potentials and mass action properties of the LTM traces. Temporal biases in time lags that are due to differences in intercellular distances are overcome by the proportionality of action potential velocity to axon diameter (Katz, 1966; Ruch, Patton, Woodbury, and Towe, 1961) in cells whose axon lengths and diameters covary. Such cells are said to be *self-similar* (Grossberg, 1969f). Self-similar cell populations can preserve the learned meaning of patterns under significant developmental deformations of their mutual distances and sizes. Self-similar rules of network design also permit individual nodes to arrive at globally correct decisions from locally ambiguous data (Grossberg, 1978a). In the developmental biology literature, self-similarity is called self-regulation (Wolpert, 1969).

15. The Unit of LTM is a Spatial Pattern: Global Constraints on Local Network Design

To illustrate the global theorems that have been proved, I consider first the simplest case, wherein only one sampling node exists (Figure 4a). Then the network is called an *outstar* because it can be drawn with the sampling node at the center of outward-facing conditionable pathways (Figure 4b) such that the LTM trace z_i in the ith pathway samples the STM trace x_i of the ith sampled cell, $i \in I$. An *outstar* is thus a functional-differential system of the form

$$\frac{d}{dt}x_i = Ax_i + Bz_i + C_i(t) \tag{37}$$

$$\frac{d}{dt}z_i = Dz_i + Ex_i \tag{38}$$

where A, B, D, and E are continuous functionals such that B and E are nonnegative.

Despite the fact that the functionals A, B, D, and E can fluctuate in extremely complex system-dependent ways, and the inputs $C_i(t)$ can also fluctuate wildly through time, an outstar can learn an arbitrary spatial pattern $C_i(t) = \theta_i C(t)$ $(\theta_i \geq 0, \sum_{k \in I} \theta_k = 1)$ with a minimum of oscillations in its pattern variables $X_i = x_i(\sum_{k \in I} x_k)^{-1}$ and $Z_i = z_i(\sum_{k \in I} z_k)^{-1}$. Recall that the Z_i's are the stimulus sampling probabilities that played such a central role in my explanation of serial bowing. Because the limits and oscillations of the pattern variables have a classification that is independent of particular choices of A, B, C, D, and E, these properties are the evolutionary invariants of outstar learning.

The following theorem summarizes, albeit not in the most general known form, some properties of outstar learning. One of the constraints in this theorem is called a *local flow* condition. This constraint says that a performance signal B can be large only if its associated learning signal E is large. Local flow prevents the read-out of old LTM memories from a sampling pathway which has lost its plasticity. Such a read-out would prevent accurate registration of the UCS in STM and thus accurate LTM encoding of the UCS via STM sampling.

I should immediately remark that a plastic synapse can be dynamically buffered against recoding by global network interactions (Grossberg, 1976c, 1980a). Such a synapse can still obey the local flow condition. I should also say that the local flow condition is needed only if all sampling sources are trying to encode the same pattern without bias, as in the parallel learning of sensory expectancies (Grossberg, 1980a) or of motor synergies (Grossberg, 1978a).

If the threshold of B is no smaller than the threshold of E, then local flow is assured. Such a threshold inequality occurs automatically if the LTM trace z_{ji} is physically interpolated between the axonal signal $S_j b_{ji}$ and the postsynaptic target cell v_i. That is why I call the condition a local flow condition. Such a geometric interpretation of

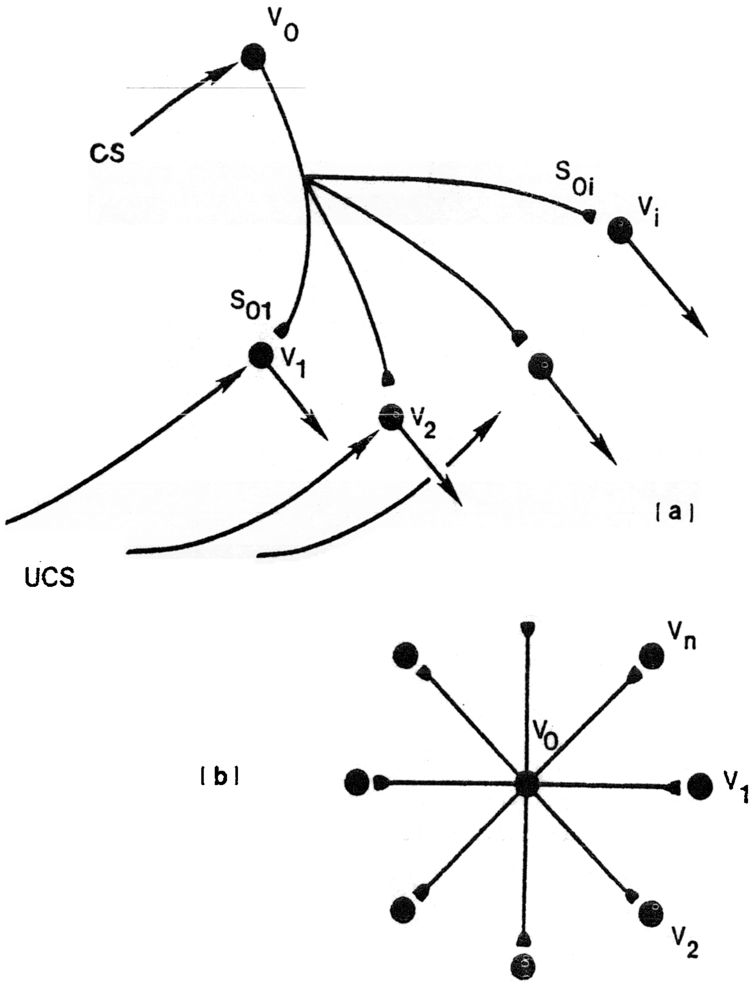

Figure 4. (a) In the minimal anatomy capable of associative learning in a classical conditioning paradigm, a conditioned stimulus (CS) excites a single node, or cell population, v_0 which thereupon sends sampling signals to a set of nodes $v_1, v_2, \ldots, v_n$. An input pattern representing the unconditioned stimulus (UCS) excites the nodes $v_1, v_2, \ldots, v_n$, which thereupon elicit output signals that contribute to the unconditioned response (UCR). The sampling signals from v_0 activate the LTM traces z_{0i} (which I denote by z_i in the text for brevity) that are computed at the synaptic knobs S_{0i}, $i = 1, 2, \ldots, n$. The activated LTM traces can learn the activity pattern across $v_1, v_2, \ldots, v_n$ that represents the UCS. (b) When the sampling structure in (a) is redrawn to emphasize its symmetry, the result is an *outstar*, whose sampling source is v_0 and whose sampled border is the set $\{v_1, v_2, \ldots, v_n\}$.

the location of the LTM trace is not forced by the psychological derivation of the associative equations, although it is the minimal anatomical realization of this derivation. Local flow gives unexpected support to the minimal realization by showing that pattern learning depends upon a mathematical constraint which automatically obtains in the minimal realization, but is at best *ad hoc* and difficult to guarantee in other anatomical interpretations of the associative equations.

Theorem 2 (Outstar Pattern Learning)

Suppose that

(I) the functionals are chosen to keep system trajectories bounded;

(II) a local flow condition holds:

$$\int_0^\infty B(t)dt = \infty \quad \text{only if} \quad \int_0^\infty E(t)dt = \infty; \tag{39}$$

(III) the UCS is practiced sufficiently often, and there exist positive constants K_1 and K_2 such that for all $T \geq 0$,

$$f(T,T+t) \geq K_1 \quad \text{if} \quad t \geq K_2 \tag{40}$$

where

$$f(U,V) = \int_U^V C(\xi)\exp\Big[\int_\xi^V A(\eta)d\eta\Big]d\xi. \tag{41}$$

Then, given arbitrary continuous and nonnegative initial data in $t \leq 0$ such that $\sum_{k \in I} z_k(0) > 0$,

(A) practice makes perfect:

The stimulus sampling probabilities $Z_i(t)$ are monotonically attracted to the UCS weights θ_i if

$$[Z_i(0) - X_i(0)][X_i(0) - \theta_i] \geq 0, \tag{42}$$

or may oscillate at most once due to prior learning if (42) does not hold, no matter how wildly A, B, C, D, and E oscillate;

(B) the UCS is registered in STM and partial learning occurs:

The limits $Q_i = \lim_{t\to\infty} X_i(t)$ and $P_i = \lim_{t\to\infty} Z_i(t)$ exist with

$$Q_i = \theta_i, \quad i \in I. \tag{43}$$

(C) If, moreover, the CS is practiced sufficiently often, then perfect learning occurs: if

$$\int_0^\infty E(t)dt = \infty, \quad \text{then} \quad P_i = \theta_i, \quad i \in I. \tag{44}$$

Remarkably, similar global theorems hold for systems (32)–(33) wherein arbitrarily many sampling cells can be simultaneously active and mutually signal each other by very complex feedback rules (Geman, 1981; Grossberg, 1969d, 1971a, 1972a, 1980b). This is because all systems of the form (32)–(33) can *factorize* information about STM and LTM pattern variables from information about how fast energy is being pumped into the system. Pattern variable oscillations can therefore be classified even if wild fluctuations in input and feedback signal energies occur through time. In the best theorems now available, only one hypothesis is not known to be necessary and sufficient (Grossberg, 1972a). It would be most satisfying if this imperfection in the theorems could be overcome.

When many sampling cells v_j, $j \in J$, can send sampling signals to each v_i, $i \in I$, the outstar property that each stimulus sampling probability $Z_{ji} = z_{ji}(\sum_{k \in I} z_{jk})^{-1}$ oscillates at most once fails to hold. This is so because the Z_{ji} of all active nodes v_j track $X_i = x_i(\sum_{k \in I} x_k)^{-1}$, while X_i tracks θ_i and the Z_{ji} of all active nodes v_j. The oscillations of the functions $Y_i = \max\{Z_{ji} : j \in J\}$ and $y_i = \min\{Z_{ji} : j \in J\}$ can, however, be classified much as the oscillations of each Z_i can be classified in the outstar case. Since each Z_{ji} depends on all z_{jk}, $k \in I$, each Y_i and y_i depends on all z_{jk}, $j \in J$, $k \in I$, and each X_i depends on all x_k, $k \in I$, the learning at each v_i is influenced by *all* x_k and z_{jk}, $j \in J$, $k \in I$. No local analysis can provide an adequate insight into the learning dynamics of these networks.

Because the oscillations of all X_i, Y_i, and y_i relative to θ_i can be classified, the following generalization of the outstar learning theorem holds.

Theorem 3 (Parallel Pattern Learning)

Suppose that

(I) the functionals are chosen to keep system trajectories bounded;

(II) every sampling cell obeys a local flow condition: for every $j \in J$,

$$\int_0^\infty B_j dt = \infty \quad \text{only if} \quad \int_0^\infty E_j dt = \infty; \tag{45}$$

(III) the UCS is presented sufficiently often: there exist positive constants K_1 and K_2 such that (40) holds.

Then given arbitrary nonnegative and continuous initial data in $t \leq 0$ such that $\sum_{k \in I} x_k(0) > 0$ and all $\sum_{k \in I} z_{jk}(0) > 0$,

(A) the UCS is registered in STM and partial learning occurs: the limits $Q_i = \lim_{t\to\infty} X_i(t)$ and $P_{ji} = \lim_{t\to\infty} Z_{ji}(t)$ exist with

$$Q_i = \theta_i, \quad i \in I. \tag{46}$$

(B) If the jth CS is practiced sufficiently often, then it learns the UCS pattern perfectly: if

$$\int_0^\infty E_j dt = \infty \quad \text{then} \quad P_{ji} = \theta_i, \quad i \in I. \tag{47}$$

Because LTM traces z_{ji} gate the performance signals B_j which are activated by a retrieval probe r_j, the theorem enables any and all nodes v_j which sampled the pattern $\theta = (\theta_i, i \in I)$ during learning trials to read it out with perfect accuracy on recall trials. The theorem does not deny that oscillations in overall network activity can occur during learning and recall, but shows that these oscillations merely influence the rates and intensities of learning and recall. Despite the apparent simplicity of these statements, the details of learning, memory, and recall can be dramatically altered by different choices of functionals. As one of many examples, phase transitions in memory can occur, and the nature of the phases can depend on a complex interaction between network rates and geometry (Grossberg, 1974).

Neither Theorem 2 nor Theorem 3 needs to assume that the CS and UCS are presented at correlated times. This is because the UCS condition keeps the baseline STM activity of sampled cells from ever decaying below the positive value K_1 in (40). For purposes of space-time pattern learning, this UCS uniformity condition is too strong. In Grossberg (1972a) I show how to replace the UCS uniformity condition by a weaker condition which guarantees that CS–UCS presentations are well enough correlated to guarantee perfect pattern learning of a given spatial pattern by certain cells v_j, even if other spatial patterns are presented at irregular times when they are sampled by distinct cells v_j.

16. The Teleology of the Pattern Calculus: Retina, Command Cell, Reward, Attention, Motor Synergy, Sensory Expectancy, Cerebellum

Three simple but fundamental facts emerge from the mathematical analysis of pattern learning: the unit of LTM is a spatial pattern $\theta = (\theta_i : i \in I)$; suitably designed neural networks can factorize invariant pattern θ from fluctuating energy; the size of a node's sampling signal can render it adaptively sensitive or blind to a pattern θ. These concepts helped me to think in terms of pattern transformations, rather than in terms of feature detectors, computer programs, linear systems, or other types of analysis. When I confronted equally simple environmental constraints with these simple pattern learning properties, the teleological pressure that was generated drove me into a wide-ranging series of specialized investigations.

What is the minimal network that can discriminate θ from background input fluctuations? It looks like a retina, and the θ's became reflectances. What is the minimal network that can encode and/or perform a space-time pattern or ordered series of spatial patterns? It looks like an invertebrate command cell. How can one synchronize CS–UCS sampling if the time intervals between CS and UCS presentations are unsynchronized? The result leads to psychophysiological mechanisms of reward, punishment, and attention. What are the associative invariants of motor learning? Spatial patterns become motor synergies wherein fixed relative contraction rates across muscles occur, and temporally synchronized performance signals read-out the synergy as a unit. What are the associative invariants of sensory learning? The potential ease of learning and reading-out complex sensory expectancies and spatial representations shows that even eidetic memory is more remarkable as a memory retrieval property than as a learning property. What is the minimal network that can bias the performance of motor acts with learned motor expectancies? It looks like a cerebellum.

An historical review of these investigations is found in the prefaces to a selection of my articles reprinted in Grossberg (1982a). Individually and collectively, these results add force to the idea that patterns rather than features are the functional units which regulate the neural designs subserving behavioral adaptation.

17. The Primacy of Shunting Competitive Networks Over Additive Networks

These specialized investigations repeatedly drove me to consider competitive systems. As just one of many instances, the same competitive normalization property which arose during my modeling of receptor-bipolar-horizontal cell interactions in retina (Grossberg, 1970a, 1972b) also arose in studies of the decision rules needed to release the right amount of incentive motivation in response to interacting drive inputs and conditioned reinforcer inputs within midbrain reinforcement centers (Grossberg, 1972c, 1972d). Because I approached these problems from a behavioral perspective, I knew what interactive properties the competition had to have. I have repeatedly found that shunting competition has all the properties that I need, whereas additive competition often does not.

As solutions to specialized problems involving competition piled up, networks capable of normalization, sensitivity changes via automatic gain control, attentional biases, developmental biases, pattern matching, shift properties, contrast enhancement, edge and curvature detection, tunable filtering, multistable choice behavior, normative drifts, traveling and standing waves, hysteresis, and resonance began to be classified within the framework of shunting competitive feedforward and feedback networks. See Grossberg (1981) for a recent review. As in the case of associative learning, the abundance of special cases made it seem more and more imperative to find an intuitive and mathematical framework within which these results could be unified and generalized. I also began to wonder whether many of the pattern transformations and STM storage properties of specialized examples were not instances of an absolute stability property of a general class of networks.

18. The Noise-Saturation Dilemma and Absolute Stability of Competitive Decision-Making

A unifying intuitive theme of particular simplicity can be recognized by considering the processing of continuously fluctuating patterns by cellular tissues. This theme is invisible to theories based on binary codes, feature detectors, or additive models. All cellular systems need to solve the *noise-saturation dilemma* which might cause sensitivity loss in their responses to both low and high input intensities. Mass action, or shunting, competition enables cells to elegantly solve this problem using automatic gain control by lateral inhibitory signals (Grossberg, 1973, 1980a). Additive competition fails in this task because it does not, by definition, possess an automatic gain control property.

A unifying mathematical theme is that every competitive system induces a decision scheme that can be used to prove global limit and oscillation theorems, notably absolute stability theorems (Grossberg, 1978c, 1978d, 1980c). This decision scheme interpretation is just a vivid way to think about a Liapunov functional that is naturally associated with each competitive system.

A class of competitive systems with absolutely stable decision schemes is the class of *adaptation level systems*

$$\frac{d}{dt}x_i = a_i(x)[b_i(x_i) - c(x)], \tag{48}$$

$i = 1, 2, \ldots n$, where $x = (x_1, x_2, \ldots, x_n)$. These systems include all shunting competitive feedback networks of the form

$$\frac{d}{dt}x_i = -A_i x_i + (B_i - x_i)[I_i + f_i(x_i)] - (x_i + C_i)[J_i + \sum_{k \neq i} f_k(x_k)] \tag{49}$$

which, in turn, are capable of many of the special properties listed above, given suitable choices of parameters and feedback signal functions. A special case of my theorem concerning these systems is the following.

Theorem 4 (Absolute Stability of Adaptation Level Systems)
Suppose that
(I) *Smoothness*: the functions $a_i(x)$, $b_i(x_i)$, and $c(x)$ are continuously differentiable;
(II) *Positivity*:

$$a_i(x) > 0 \quad \text{if} \quad x_i > 0, \quad x_j \geq 0, \quad j \neq i; \tag{50}$$

$$a_i(x) = 0 \quad \text{if} \quad x_i = 0, \quad x_j \geq 0, \quad j \neq i; \tag{51}$$

for sufficiently small $\lambda > 0$, there exists a continuous function $\bar{a}_i(x_i)$ such that

$$\bar{a}_i(x_i) \geq a_i(x) \quad \text{if} \quad x \in [0, \lambda]^n \tag{52}$$

and

$$\int_0^\lambda \frac{dw}{\bar{a}_i(w)} = \infty; \tag{53}$$

(III) *Boundedness*: for each $i = 1, 2, \ldots, n$,

$$\limsup_{x_i \to \infty} b_i(x_i) < c(0, 0, \ldots, \infty, 0, \ldots, 0) \tag{54}$$

where ∞ is in the ith entry of $(0, 0, \ldots, \infty, 0, \ldots, 0)$;

(IV) *Competition:*

$$\frac{\partial c(x)}{\partial x_i} > 0, \quad x \in \mathbf{R}^n_+, \quad i = 1, 2, \ldots, n; \tag{55}$$

(V) *Decision Hills:* The graph of each $b_i(x_i)$ possesses at most finitely many maxima in every compact interval.

Then the pattern transformation is stored in STM because all trajectories converge to equilibrium points: given any $x(0) > 0$, the limit $x(\infty) = \lim_{t \to \infty} x(t)$ exists.

This theorem intuitively means that the decision schemes of adaptation level systems are globally consistent. Globally inconsistent decision schemes can, by contrast, force almost all trajectories to persistently oscillate. This can occur even if $n = 3$ and all feedback signals are linear, as the voting paradox vividly illustrates (Grossberg, 1978c, 1980c; May and Leonard, 1975).

Adaptation level systems exclude distance-dependent interactions. To overcome this gap, Michael Cohen and I (Cohen and Grossberg, 1982) recently studied the absolute stability of the distance-dependent networks

$$\frac{d}{dt} x_i = -A_i x_i + (B_i - C_i x_i)[I_i + f_i(x_i)] - (D_i x_i + E_i)[J_i + \sum_{k=1} g_k(x_k) F_{ki}]. \tag{56}$$

Distance-dependence means that $F_{ki} = F_{ik}$. The networks (56) include examples of Volterra-Lotka systems, Hartline-Ratliff networks, Gilpin-Ayala systems, and shunting and additive networks.

In this setting, we constructed a global Liapunov function for these systems and used the LaSalle Invariance Principle, Sard's lemma, and some results about several complex variables to analyze the limiting behavior of (56). Modulo some technical hypotheses, we have proved that almost all systems of the form (56) are absolutely stable, and that systems with polynomial and sigmoid feedback signals can be directly analyzed.

These results show that adaptation level and distance-dependent competitive networks represent stable neural designs for competitive decision-making. The fact that adaptation level systems have been analyzed using Liapunov functionals whereas distance-dependent networks have been analyzed using Liapunov functions shows that the absolute stability theory of competitive systems is still incomplete. Absolute stability theorems for cooperative systems have also been recently discovered (Hirsch, 1982a, 1982b). This is an exciting area for intensive mathematical investigation.

The final sections of the article discuss code development issues wherein interactions between associative and competitive rules play a central role.

19. The Babel of Code Development Models

The experimental interest in geniculo-cortical and retino-tectal development (Gottlieb, 1976; Hubel and Wiesel, 1977; Hunt and Jacobson, 1974) has been paralleled by a vigorous theoretical interest in these basic phenomena. Perhaps in no other area of brain theory is the issue of what constitutes a new model, a new idea, or real progress so badly discussed. A literal reading of this literature might lead one to conclude that a one-to-one correspondence between articles and models exists, or at least between authors and models. A world of theoretical nomads is an anarchy, which is the antithesis of what a theoretical community should be. If we are to achieve the coherence that theory must have to be effective, then the endless numerical and experimental variations on our laws must not be confused with the invariant structure of these laws. A new model is not a change of notation, a use of a discrete instead of a continuous time variable, a different setting of numerical parameters, or a presentation of the same equations with a different input series.

When Malsburg (1973) adapted the equations which he found in Grossberg (1972b) for computer simulation and subjected them to a series of input patterns, I was delighted but not surprised by his findings. I was delighted because here was an interesting new twist in the use of the equations. I was not surprised because the results are a variant of pattern learning properties which had already been studied. Now I will review some of the relationships between code development and pattern learning, state some mathematical results on code development which computer studies missed, and make some comparative remarks about recent articles in the literature.

20. The Duality Between Code Development and Pattern Learning

In both pattern learning and code development situations, one often finds two sets, or fields, $F^{(1)}$ and $F^{(2)}$ of cells, which are not necessarily disjoint. The set of sampled cells v_i, $i \in J$, and sampling cells v_j, $j \in J$, are illustrative. Conditionable pathways e_{ji} are assumed to exist from one set to the other set of cells, and LTM traces z_{ji} are assigned to the pathways e_{ji}. Competitive interactions are assumed to occur within $F^{(1)}$ and $F^{(2)}$, if only to solve the noise-saturation dilemma at each level of pattern processing. In what, then, does the difference between a pattern learning and a code development model consist?

In a word, the answer is arrow-reversal, or duality. Whereas the conditionable pathways in a pattern learning example point from sampling cell to sampled cells, the conditionable pathways in a code development example point from sampled cells to sampling cell. Because of arrow-reversal, each sampling cell receives a sum of LTM-gated signals from sampled cells, which in turn influence the activity of the sampling cell and thus whether the sampled cells will be sampled.

If we apply the principle of sufficient reason to the arrow-reversal distinction, it becomes more ambiguous. How, after all, does an individual LTM trace z_{ji} from v_j to v_i know whether v_j is a sampling cell and v_i a sampled cell, or conversely? The answer is that it doesn't. Consequently, similar principles of pattern learning hold in both cases. Only when we ask more global questions about network design do distinctions between the two problems emerge.

For example, how do the fields $F^{(1)}$ and $F^{(2)}$ determine whether their cells will be sampling cells, sampled cells, or both? A major part of the answer lies in how sharply $F^{(1)}$ and $F^{(2)}$ contrast enhance their input patterns. To fix ideas, suppose that conditionable pathways pass between $F^{(1)}$ and $F^{(2)}$ in both directions and that both $F^{(1)}$ and $F^{(2)}$ directly receive input patterns. If $F^{(1)}$ does not sharply contrast enhance the input patterns but $F^{(2)}$ does, then $F^{(2)}$ will encode patterns across $F^{(1)}$ within the $F^{(1)} \to F^{(2)}$ LTM traces, and $F^{(2)}$ will learn patterns across $F^{(1)}$ within the $F^{(2)} \to F^{(1)}$ LTM traces. The difference between code development and pattern learning in this example thus resides in an asymmetric choice of competitive parameters within $F^{(1)}$ and $F^{(2)}$, not in a choice of new associative or competitive laws.

21. Outstars and Instars

These facts become clearer if we start with the simplest examples of pattern learning and code development, and then build up towards more complex examples. As Section 15 noted, the simplest network capable of pattern learning is an outstar (Figure 5a). By duality, the simplest network capable of code development is an instar (Figure 5b). The main difference between an outstar and an instar is that the source of an outstar excites the outstar border, whereas the border of an instar excites the instar source.

The changing efficacy with which practiced border patterns can excite the instar source constitutes code development. Because of the outstar learning theorem, it is no surprise that the LTM traces of an instar can learn a spatial pattern that perturbs its border. In an outstar, if a space-time pattern or sequence of spatial patterns plays upon its border while the source cell is sampling, then the source learns a weighted average of the sampled patterns (Grossberg, 1970b). This fact also holds in an instar for the same mathematical reason.

It is instructive to write down equations for an instar and to compare them with illustrative examples in the literature. Because an instar reverses the arrows between sampling and sampled cells, an instar with a local symmetry axis with respect to its sampling cell v_1 ($J = \{1\}$) obeys equations such as

$$\frac{d}{dt}x_1 = A_1 x_1 + \sum_{k \in I} B_k z_{k1} + C_1, \tag{57}$$

$$\frac{d}{dt}x_i = A x_i + C_i, \quad i \in I, \tag{58}$$

and

$$\frac{d}{dt}z_{i1} = D_1 z_{i1} + E_1 x_i, \tag{59}$$

$i \in I$. In (57), the sampling cell v_1 receives LTM gated signals from the sampled cells v_i, $i \in I$, in addition to a possible input C_1. In (58), the sampled cells v_i share a common STM decay functional A due to the local symmetry axis, but receive distinct inputs C_i from the input patterns $(C_i : i \in I)$. In (59), the usual LTM trace law holds with a shared LTM decay functional D_1 and a shared learning functional E_1 due to the local symmetry axis.

The article by Bienenstock, Cooper, and Munro (1982) is devoted to the study of a locally symmetric instar. These authors consider the equation

$$\frac{d}{dt}m_j = -\epsilon m_j + \phi d_j \tag{60}$$

for the jth LTM trace m_j and the jth input d_j. They define ϕ to be a functional of the past and present values of the function

$$c = \sum_j d_j m_j. \tag{61}$$

In particular, they use an average of past values of c as a threshold against which a present value of c is compared. If the present value exceeds threshold, $\phi > 0$, otherwise not. The threshold is assumed to increase as a function of past values of c.

A simple change of notation shows that equations (60) and (61) are a lumped version of an instar. In (59), let $i = j$, $z_{i1} = m_j$, $D_1 = -\epsilon$, and $E_1 = \phi$ to see that (59) subsumes (60). In (58), let A average C_1 so fast that

$$x_i \cong C_i, \quad i \in I. \tag{62}$$

In (57), let $C_1 \simeq 0$ and let A_1 rapidly average $\sum_{k \in I} B_k z_{k1}$ so fast that

$$x_1 \cong \sum_{k \in I} B_k z_{k1}. \tag{63}$$

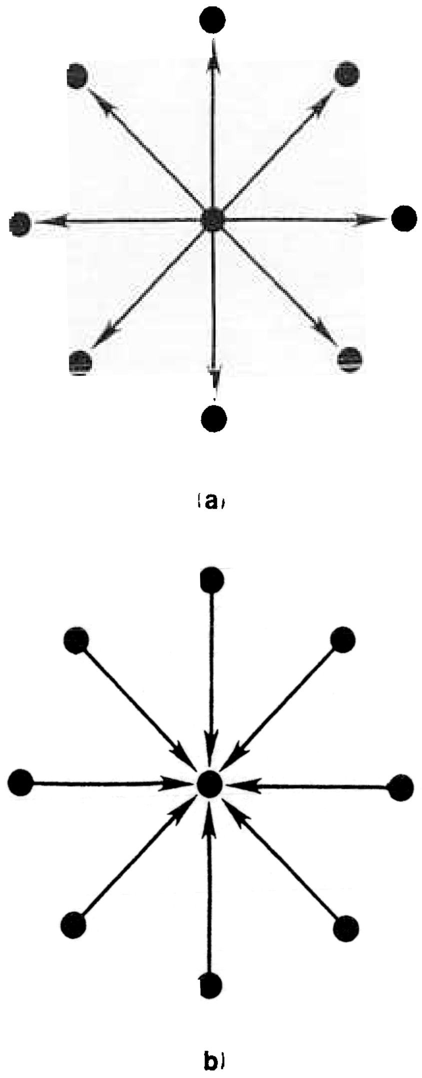

Figure 5. (a) An *outstar* is the minimal network capable of associative pattern learning. (b) An *instar* is the minimal network capable of code development. The source of an outstar excites the outstar border, whereas the border of an instar excites the instar source. In both cases, source activation is necessary to drive an LTM sampling process. Since the instar border signals are gated by LTM traces before activating the instar source, code learning changes the efficacy of source activation and is changed by it.

Letting $B_k = x_k$ shows, by (62), that

$$x_1 \cong \sum_{k \in I} C_k z_{k1}, \tag{64}$$

which is the same as c in (61), but in different notation. Now plug x_1 into E_1 and use a threshold rule as in (31) to complete the reduction.

Despite the obvious nature of this reduction, the authors make a number of claims that illustrate the present fragmentation of the theoretical community. They say that they have introduced in the threshold rule "a new and essential feature" which they call "temporal competition between input patterns." They also write that Cooper, Lieberman, and Oja (1979) were the first to introduce "the idea of such a modification scheme." They note that their equations result "in a form of *competition between incoming patterns* rather than competition between synapses" which they allege to be the conclusion of alternative theories. They also suggest that "our theory is in agreement with classical experimental results obtained over the last generation." Finally, in 1981 they feel free to "conjecture that some form of correlation modification is a very general organizational principle."

The status of some of these claims is clear from the preceding discussion. I will, however, indicate below how the threshold rule in (60) and (61) generates a temporally unstable code when more than one sampling node exists, and why this threshold rule either cannot explain critical period termination or cannot explain the results of Pettigrew and Kasamatsu (1978). Thus although equations (60) and (61) are a special case of an instar, not all choices of instar functionals are equally good for purposes of stable code development.

22. Adaptive Filtering of Spatial Pattern Sequences

The comparison between pattern learning and code development becomes more interesting when a space-time pattern, or sequence of spatial patterns, is to be parsed by pattern learning or code development mechanisms. In either case, the fact that the LTM unit is a spatial pattern is fundamental, and the task is to show how individual spatial patterns, or subsequences of spatial patterns, can be differentially processed. To do this, one needs to show how distinguishable sampling sources, or subsets of sources, can be sequentially activated by the spatial patterns in the pattern sequence (Figure 6).

In the simplest pattern learning examples, pre-wired sequentially activated sampling nodes can learn an arbitrary space-time pattern (Grossberg, 1969e, 1970b). The price paid for such a ritualistic encoding is that the order of pattern performance, although not its velocity, is rigidly constrained. This *avalanche* type of anatomy is isomorphic to the anatomies of certain invertebrate command cells (Grossberg, 1974; Stein, 1971), and illustrates that complex acts can be encoded by small numbers of cells if ritualistic performance is acceptable. In examples wherein the order with which sampling nodes will be activated is not prewired into the network, serial learning mechanisms—notably associative and competitive interactions such as those utilized in Section 12—are needed to learn the correct ordering as practice of the ordering proceeds (Grossberg, 1969c, 1974, 1978a). In examples wherein the filtering rules whereby individual sampling nodes are selected are not prewired into the network, we are confronted with a problem in code development, notably the problem of how *sequences* of events adaptively select the nodes that will elicit the most accurate predictive commands within their sequential context (Grossberg, 1978a). Most code development models consider special cases of this general problem.

If we generalize the instar equations (57) and (59) to include the possibility that many sampling (encoding) cell indices occur in J, we find equations

$$\frac{d}{dt} x_j = A_j x_j + \sum_{k \in I} B_k z_{kj} + I_j \tag{65}$$

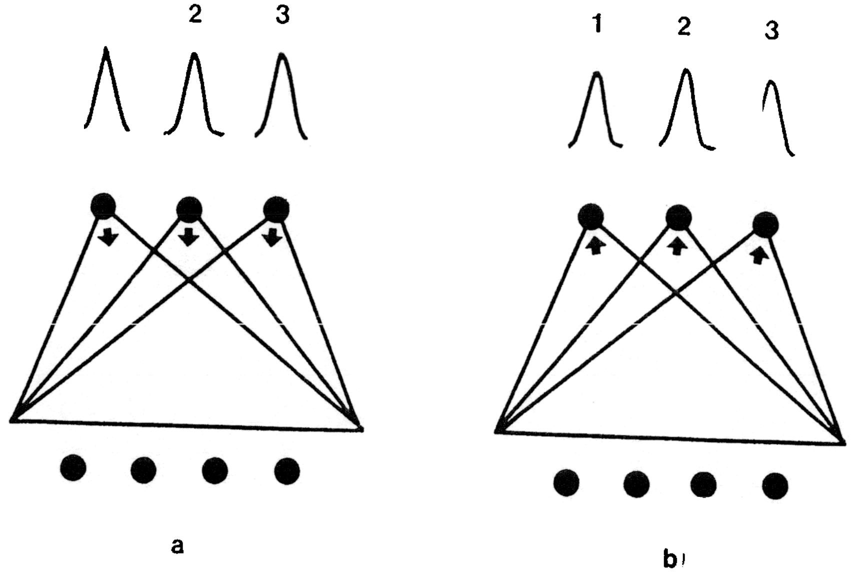

Figure 6. (a) In the simplest problem of space-time pattern learning, a mechanism is needed to excite discriminable sampling sources in a prescribed order $1, 2, 3, \ldots$ during learning trials, and to repeat the same order $1, 2, 3, \ldots$ of excitation during recall trials. (b) In the simplest problem of space-time code development, a space-time pattern at the sampled cells initially excites sampling sources in a prescribed order $1, 2, 3, \ldots$ due to the *a priori* biases in the filter from sampled cells to sampling cells. Whether this ordering approaches a stable configuration as development proceeds, or a temporally unstable sequence of coding representations is triggered, depends on details of network design.

and

$$\frac{d}{dt} z_{ij} = D_j z_{ij} + E_j x_i, \tag{66}$$

$i \in I, j \in J$. Although all the terms in (65) and (66) work together to achieve code development, perhaps the terms $F_j \equiv \sum_{k \in I} B_k z_{kj}$ should be singled out as the principal ones. We all learn in calculus or linear algebra that the terms F_j are dot products, or inner products, of the vector $B = (B_k : k \in I)$ with the vectors $z_j = (z_{kj} : k \in I)$; viz.,

$$F_j = B \cdot z_j. \tag{67}$$

If we define the vector $F = (F_j : j \in J)$ of dot products, then we can also recognize that the mapping $B \to F$ is a *linear filter*. By reversing arrows to go from pattern learning to code development, we hereby convert the property of independent read-out of probed pattern recall (Section 10) into the property of linear filtering—*without a change of model.*

An elementary formula about dot products underlies basic properties of code development. This formula is the law of cosines:

$$B \cdot z_j = \| B \| \, \| z_j \| \; \cos(B, z_j), \tag{68}$$

where $\| V \|$ is the Euclidean length of vector V. By (68), given a fixed signal pattern B, F_j is maximized among all z_j of fixed length by choosing z_j parallel to B. Thus in response to a fixed pattern B, the nodes v_j for which z_j is most parallel to B will be the ones most highly activated by B, other things being equal. If (66) causes the LTM vectors z_j of highly activated nodes v_j to become more parallel to B due to frequent past B presentations, then on future B presentations these nodes will respond ever more vigorously to B. Let us call alterations in F due to past experience with B *adaptive filtering*. Then we can say that an interaction between adaptive filtering from $\{v_i : i \in I\}$ to $\{v_j : j \in J\}$ and competition within $\{v_j : j \in J\}$ to select active nodes controls at least the most elementary features of code development.

23. Synaptic Conservation, Code Invariance, and Code Instability

These observations about adaptive filtering did not, perhaps surprisingly, appear in Malsburg's original paper. Other important properties also have not been sharply articulated by computer analysis. For example, presenting a given pattern several times can recode not only the nodes which respond most vigorously to this pattern but also the responses of inactive nodes to other patterns presented later on, because each node can encode infinite sets of patterns which it has never before experienced. This has the nice consequence that the code can learn to recognize certain invariant properties of pattern classes without having to experience all the patterns in each class (Fukushima, 1980; Grossberg, 1978a).

Another deeper consequence is not so pleasant. If there exist many more patterns than encoding nodes v_j, $j \in J$, there need not exist *any* temporally stable coding rule; that is, the nodes which respond most vigorously to a given pattern can continually change through time as the same list is repetitively practiced (Grossberg, 1976b). I was led to suspect that such a result might hold due to my prior theorems about absolute stability of parallel pattern learning, which suggested possible destabilizing effects of the STM decay terms A_j in (65). This important instability result has been missed by all the computer studies that I know because these studies typically present small numbers of patterns to small numbers of cells. Indeed, they usually present small numbers of patterns (e.g., 19) to much larger sets of cells (e.g., 169), as in the careful analysis of Amari and Takeuchi (1978).

An instability result forces one to ask which properties are essential to code development and which properties are adventitious. For example, Malsburg supplemented equations (3), (5), (6), and (7) with a synaptic conservation rule that requires the sum $\sum_{k \in I} z_{kj}$ of all the synaptic strengths which converge on a node v_j to be constant through time. Because I was aware of the duality between pattern learning and code development, I realized that the synaptic conservation rule is incompatible with the simplest example of associative learning, namely classical conditioning (Grossberg, 1976a). This is because the UCR pattern must be extinguished in memory as the CR pattern is learned if synaptic conservation holds. I was therefore reluctant to accept the synaptic conservation rule without an important physical or mathematical reason.

I could, however, see the need for a type of conservation, or normalization, that would regulate the temporal stability of the sampling process. By the time I read Malsburg's paper, I knew that long-range shunting competition, as opposed to the additive competition which Malsburg inherited from me, can automatically normalize the total suprathreshold STM activity of a network (Grossberg, 1973). The normalized STM activity can, in turn, normalize, or conserve, total synaptic strength across the network via feedback of E_j to z_{ij} in (66). This synaptic conservation mechanism is, moreover, compatible with classical conditioning. I therefore concluded that shunting competition, which can be absorbed into the STM decay terms A_j of (65), should formally replace synaptic conservation until more pressing reasons to the contrary are given. Some experimental tests of synaptic competition versus STM normalization are suggested in Grossberg (1981).

In their more recent contributions to retinotectal development, Malsburg and Willshaw have continued to use additive interactions, supplemented by the synaptic conservation rule and another rule for matching the similarity between retinal and tectal marker patterns (Malsburg and Willshaw, 1977, 1981; Willshaw and Malsburg, 1976). Since shunting networks automatically have matching properties as well as normalization properties, I take the need for these special assumptions as support for the idea that shunting operations subserve retinotectal development (Grossberg, 1976b, 1981). I have also argued that shunting interactions are operative in a variety of non-neural developmental examples, such as gastrulation in the sea urchin, slime mold aggregation, and regeneration in *Hydra* (Grossberg, 1978e). In all of these cases, I believe that alternative models have adapted analogies from chemical kinetics that do not incorporate mass action properties of cellular tissues. Notable differences between additive and shunting models occur in their explanations of the self-regulation mechanism that underlies the invariance of form when growth occurs (Gierer and Meinhardt, 1972; Grossberg, 1980b; Meinhardt and Gierer, 1974) and the contrast enhancement mechanism of categorical perception (Anderson, Silverstein, Ritz, and Jones, 1977; Grossberg, 1978f).

24. Critical Period Termination, the Stability-Plasticity Dilemma, and Adaptive Resonance

The fact that no temporally stable code need exist in response to a large family of input patterns, such as occurs in vision, made the problem of terminating those critical periods which are sensitive to behavioral experience seem more severe. This fact suggested that either the critical period is terminated by a chemical switch, but then there is a high likelihood that the code will incorporate adventitious statistical fluctuations of the most recent input sequences, or that the code is stabilized by a gradual process of dynamic buffering in response to network states that signify the behavioral relevance of the coded data. This dilemma led me to build my theory of *adaptive resonances* (Grossberg, 1976c, 1978a, 1980a, 1982b) which formalizes an answer to what I call the *stability-plasticity dilemma*.

The stability-plasticity dilemma asks how internal representations can maintain themselves in a stable fashion against the erosive effects of behaviorally irrelevant environmental fluctuations, yet can nonetheless adapt rapidly in response to environmental fluctuations that are crucial to survival. How does a network as a whole know the difference between behaviorally irrelevant and relevant events even though its individual cells do not possess this knowledge? How does a network transmute this knowledge into the difference between slow and fast rates of adaptation, respectively? Classical examples of the stability-plasticity balance are found in the work of Held and his colleagues on rapid visual adaptation in adults to discordant visuomotor data (Held, 1961, 1967; Held and Hein, 1963) and in the work of Wallach and his colleagues on rapid visual adaptation to discordant cues for the kinetic depth effect and cues for retinal disparity (Wallach and Karsh, 1963a, 1963b; Wallach, Moore, and Davidson, 1963). The stability-plasticity issue is raised on a pharmacological level by the experiments of Pettigrew and Kasamatsu (1978) which show that the visual plasticity of normal adult cats can be restored by selectively adding some noradrenaline to cortical tissues which already possess a functioning noradrenaline arousal system.

The adaptive resonance theory which I introduced in Grossberg (1976c) can explain the Pettigrew and Kasamatsu (1978) data; see Grossberg (1982b) for a review. Let me briefly indicate why the Bienenstock, Cooper, and Munro (1982) work cannot.

First note what happens when (60) is embedded in a system such as (65) wherein several sampling nodes can compete for activity. By the threshold rule of (60), a node v_j which has successfully won this competition in the past will acquire a progressively higher threshold due to persistent activation by its own input $\sum_k B_k z_{kj}$. By contrast, other nodes v_m which do not win the STM competition when $\sum_k B_k z_{kj}$ occurs, but which receive significant projections $\sum_k B_k z_{km}$, will maintain a low threshold. Thus, the tradeoff between input size and threshold can ultimately favor a new set of nodes. When this happens, the pattern will be recoded, and a temporally unstable coding cycle will be initiated. This instability does not require a large number of coding patterns to occur. It can occur only when one pattern is repeatedly presented to a network containing more than one encoding node. In fact, the last examples in Grossberg (1976b, p.132) consider history-dependent threshold changes, much like those in the Bienenstock *et al.* example. I note their instability in a competitive sampling milieu before introducing the adaptive resonance theory in Grossberg (1976c) as a possible way out.

One might object to the above criticism by claiming that the original winning node v_j acquries a high threshold so quickly that only the adaptively enhanced input $\sum_k B_k z_{kj}$ can exceed this threshold. In other words, the parameters may be carefully chosen to quickly shut off the critical period. But then one cannot understand how adding a little noradrenaline can turn it back on. In this example, either the critical period does not shut off, whence temporal instabilities in coding can occur, or its does shut off, whence critical period reversal by noradrenaline application cannot be explained. Of course, quickly raising the threshold might in any case trigger unstable coding by favoring new nodes.

25. Stable Coding of Pattern Sequences

I will end my remarks with two theorems about stable pattern coding (Grossberg, 1976b). These theorems do not even dent the surface of the mathematical challenges raised by the theory of adaptive resonances. The theorems consider the simplest case wherein:

(1) The patterns across nodes v_i, $i \in I$, are immediately and perfectly normalized. Thus input $C_i(t) = \theta_i C(t)$ generates activity $x_i(t) = \theta_i$.

(2) The signals B_k in (65) are linear functions of the activities x_k. Choose $B_k = \theta_k$ for definiteness.

(3) The competition among nodes v_j, $j \in J$, normalizes the total activity (to the value 1 for definiteness) and rapidly chooses the nodes v_j for STM storage which receive the largest input. In other words,

$$x_j = \begin{cases} 1 & \text{if } F_j > \max\{\epsilon, F_k : k \neq j\} \\ 0 & \text{if } F_j \leq \max\{\epsilon, F_k : k \neq j\} \end{cases} \tag{69}$$

where

$$F_j = \sum_{k \in I} \theta_k z_{kj} \tag{70}$$

and ϵ represents the quenching threshold of the competition (Grossberg, 1973).

(4) The LTM traces sample the pattern $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ only when their sampling cell is active. Thus

$$\frac{d}{dt} z_{ij} = (-z_{ij} + \theta_i) x_j. \tag{71}$$

Amari and Takeuchi (1978) study essentially identical equations and arrive at related results in the case of one encoding cell. They also study the response of the equations to

inputs which simulate experiments on monocular and alternate-monocular deprivation of the kitten visual cortex.

The first result shows that if a single pattern is practiced, it maximizes the input (inner product) to its encoding cell population v_j by making z_j become parallel to θ. Simultaneously, the length of z_j becomes normalized.

Theorem 5 (Single Pattern Code)

Given a pattern θ, suppose that there exists a unique $j \in J$ such that

$$F_j(0) > \max\{\epsilon, F_k(0) : k \neq j\}. \tag{72}$$

Let θ be practiced during a sequence of non-overlapping intervals $[U_k, V_k]$, $k = 1, 2, \ldots$. Then the angle between $z_j(t)$ and θ monotonically decreases, the signal $F_j(t)$ is monotonically attracted towards $\| \theta \|^2$, and $\| z_j(t) \|^2$ oscillates at most once as it tracks $F_j(t)$. In particular, if $\| z_j(0) \| \leq \| \theta \|$, then $F_j(t)$ is monotone increasing. Except in the trivial case that $F_j(0) = \| \theta \|^2$, the limiting relations

$$\lim_{t \to \infty} \| z_j(t) \|^2 = \lim_{t \to \infty} F_j(t) = \| \theta \|^2 \tag{73}$$

hold if and only if

$$\sum_{k=1}^{\infty} (V_k - U_k) = \infty. \tag{74}$$

The second result characterizes those sets of input patterns which can generate a temporally stable code, and shows that the classifying vectors $z_j(t)$ approach the convex hull of the patterns which they encode. The latter property shows that the nodes v_j ultimately receive the maximal possible inputs from the pattern sets which they encode.

To state the theorem, the following notion is convenient. A *partition* $\oplus_{k=1}^{K} P_k$ of a finite set P is a subdivision of P into non-overlapping and exhaustive subsets P_j. The *convex hull* $H(P)$ of P is the set of all convex combinations of elements in P. Given a set $Q \subset P$, let $R = P \backslash Q$ denote the elements in P that are not in Q. If the classifying vector $z_j(t)$ codes the set of patterns $P_j(t)$, let $P_j^*(t) = P_j(t) \cup \{z_j(t)\}$. The distance between a vector p and a set of vectors Q, denoted by $\| p - Q \|$, is defined by $\| p - Q \| = \inf\{\| p - Q \| : q \in Q\}$.

Theorem 6 (Stability of Sparse Pattern Codes)

Let the network practice any finite set $P = \{\theta^{(i)} : i = 1, 2, \ldots M\}$ of patterns for which there exists a partition $P = \oplus_{k=1}^{N} P_k(T)$ at some time $t = T$ such that

$$\min\{u \cdot v : u \in P_j(T), v \in P_j^*(T)\} > \max\{u \cdot v : u \in P_j(T), v \in P^*(T) \backslash P_j^*(T)\} \tag{75}$$

for all $j = 1, 2, \ldots N$. Then

$$P_j(t) = P_j(T) \quad \text{for} \quad t \geq T, \quad j = 1, 2, \ldots, N, \tag{76}$$

and the functions

$$D_j(t) = \| z_j(t) - H(P_j(t)) \| \tag{77}$$

are monotone decreasing for $t \geq T$, $j = 1, 2, \ldots, N$. If, moreover, the patterns $P_j(T)$ are practiced in the time intervals $[U_{jk}, V_{jk}]$, $k = 1, 2, \ldots$ such that

$$\sum_{k=1}^{\infty} (V_{jk} - U_{jk}) = \infty, \tag{78}$$

then

$$\lim_{t\to\infty} D_j(t) = 0. \tag{79}$$

Despite the fact that the code of a sparse pattern class is stable, it is easy to construct examples of pattern sequences which are densely distributed in pattern space for which no temporally stable code exists. To stabilize a behaviorally sensitive developing code in an arbitrary input environment, I have constructed the adaptive resonance theory, which uses the same feedback laws to stabilize infant code development as are needed to analyze data on adult attention. I have therefore elsewhere suggested that adult attention is a continuation on a developmental continuum of the mechanisms needed to solve the stability-plasticity dilemma in infants.

REFERENCES

Adams, J.A., **Human memory**. New York: McGraw-Hill, 1967.

Amari, S.-I., A method of statistical neurodynamics. *Kybernetik*, 1974, **14**, 201–215.

Amari, S.-I., A mathematical approach to neural systems. In J. Metzler (Ed.), **Systems neuroscience**. New York: Academic Press, 1977.

Amari, S.-I. and Takeuchi, A., Mathematical theory on formation of category detecting nerve cells. *Biological Cybernetics*, 1978, **29**, 127–136.

Anderson, J.R. and Bower, G.H., **Human associative memory**. Washington, DC: V.H. Winston and Sons, 1973.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., and Jones, R.S., Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 1977, **84**, 413–451.

Asch, S.E. and Ebenholtz, S.M., The principle of associative symmetry. *Proceedings of the American Philosophical Society*, 1962, **106**, 135–163.

Atkinson, R.C. and Shiffrin, R.M., Human memory: A proposed system and its control processes. In K.W. Spence and J.T. Spence (Eds.), **Advances in the psychology of learning and motivation research and theory** (Vol. 2). New York: Academic Press, 1968.

Bienenstock, E.L., Cooper, L.N., and Munro, P.W., Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. Preprint, 1982.

Bower, G.H. (Ed.), **Human memory: Basic processes**. New York: Academic Press, 1977.

Cohen, M.A. and Grossberg, S., Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. Submitted for publication, 1982.

Cole, K.S., **Membranes, ions, and impulses**. Berkeley, CA: University of California Press, 1968.

Collins, A.M. and Loftus, E.F., A spreading-activation theory of semantic memory. *Psychological Review*, 1975, **82**, 407–428.

Cooper, L.N., Lieberman, F., and Oja, E., A theory for the acquisition and loss of neuron specificity in visual cortex. *Biological Cybernetics*, 1979, **33**, 9.

Dixon, T.R. and Horton, D.L., **Verbal behavior and general behavior theory**. Englewood Cliffs, NJ: Prentice-Hall, 1968.

Feigenbaum, E.A. and Simon, H.A., A theory of the serial position effect. *British Journal of Psychology*, 1962, **53**, 307–320.

Fukushima, K., Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980, **36**, 193–202.

Geman, S., The law of large numbers in neural modelling. In S. Grossberg (Ed.), **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society, 1981.

Gierer, A. and Meinhardt, H., A theory of biological pattern formation. *Kybernetik*, 1972, **12**, 30–39.

Gottlieb, G. (Ed.), **Neural and behavioral specificity** (Vol. 3). New York: Academic Press, 1976.

Grossberg, S., Senior Fellowship thesis, Dartmouth College, 1961.

Grossberg, S., **The theory of embedding fields with applications to psychology and neurophysiology**. New York: Rockefeller Institute for Medical Research, 1964.

Grossberg, S., Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Sciences*, 1968, **60**, 758–765.

Grossberg, S., Embedding fields: A theory of learning with physiological implications. *Journal of Mathematical Psychology*, 1969, **6**, 209–239 (a).

Grossberg, S., On learning, information, lateral inhibition, and transmitters. *Mathematical Biosciences*, 1969, **4**, 255–310 (b).

Grossberg, S., On the serial learning of lists. *Mathematical Biosciences*, 1969, **4**, 201–253 (c).

Grossberg, S., On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statistical Physics*, 1969, **1**, 319–350 (d).

Grossberg, S., Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, I. *Journal of Mathematics and Mechanics*, 1969, **19**, 53–91 (e).

Grossberg, S., On the production and release of chemical transmitters and related topics in cellular control. *Journal of Theoretical Biology*, 1969, **22**, 325–364 (f).

Grossberg, S., Neural pattern discrimination. *Journal of Theoretical Biology*, 1970, **27**, 291–337 (a).

Grossberg, S., Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Studies in Applied Mathematics*, 1970, **49**, 135–166 (b).

Grossberg, S., Pavlovian pattern learning by nonlinear neural networks. *Proceedings of the National Academy of Sciences*, 1971, **68**, 828–831 (a).

Grossberg, S., On the dynamics of operant conditioning. *Journal of Theoretical Biology*, 1971, **33**, 225–255 (b).

Grossberg, S., Pattern learning by functional-differential neural networks with arbitrary path weights. In K. Schmitt (Ed.), **Delay and functional-differential equations and their applications**. New York: Academic Press, 1972 (a).

Grossberg, S., Neural expectation: Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik*, 1972, **10**, 49–57 (b).

Grossberg, S., A neural theory of punishment and avoidance, I: Qualitative theory. *Mathematical Biosciences*, 1972, **15**, 39–67 (c).

Grossberg, S., A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosciences*, 1972, **15**, 253–285 (d).

Grossberg, S., Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 1973, **52**, 217–257.

Grossberg, S., Classical and instrumental learning by neural networks. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**. New York: Academic Press, 1974.

Grossberg, S., On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biological Cybernetics*, 1976, **21**, 145–159 (a).

Grossberg, S., Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 1976, **23**, 121–134 (b).

Grossberg, S., Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, 1976, **23**, 187–202 (c).

Grossberg, S., A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**, Vol. 5. New York: Academic Press, 1978 (a).

Grossberg, S., Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, 1978, **3**, 199–219 (b).

Grossberg, S., Decisions, patterns, and oscillations in nonlinear competitive systems with applications to Volterra-Lotka systems. *Journal of Theoretical Biology*, 1978, **73**, 101–130 (c).

Grossberg, S., Competition, decision, and consensus. *Journal of Mathematical Analysis and Applications*, 1978, **66**, 470–493 (d).

Grossberg, S., Communication, memory, and development. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**, Vol. 5. New York: Academic Press, 1978 (e).

Grossberg, S., Do all neural models really look alike? A comment on Anderson, Silverstein, Ritz, and Jones. *Psychological Review*, 1978, **85**, 592–596 (f).

Grossberg, S., How does a brain build a cognitive code? *Psychological Review*, 1980, **1**, 1–51 (a).

Grossberg, S., Intracellular mechanisms of adaptation and self-regulation in self-organizing networks: The role of chemical transducers. *Bulletin of Mathematical Biology*, 1980, **42**, 365–396 (b).

Grossberg, S., Biological competition: Decision rules, pattern formation, and oscillations. *Proceedings of the National Academy of Sciences*, 1980, **77**, 2338–2342 (c).

Grossberg, S., (Ed.), Adaptive resonance in development, perception, and cognition. In **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society, 1981.

Grossberg, S., **Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control**. Boston: Reidel Press, 1982 (a).

Grossberg, S., Some psychophysiological and pharmacological correlates of a developmental, cognitive, and motivational theory. In J. Cohen, R. Karrer, and P. Tueting (Eds.), **Proceedings of the 6th evoked potential international conference**, June 21–26, 1981, Lake Forest, Illinois. New York: New York Academy of Sciences, 1982 (b). (Published as **Brain and information: Event related potentials**, **425**, 58–151, Annals of the New York Academy of Sciences, 1984.)

Grossberg, S. and Pepe, J., Schizophrenia: Possible dependence of associational span, bowing, and primacy versus recency on spiking threshold. *Behavioral Science*, 1970, **15**, 359–362.

Grossberg, S. and Pepe, J., Spiking threshold and overarousal effects in serial learning. *Journal of Statistical Physics*, 1971, **3**, 95–125.

Held, R., Exposure-history as a factor in maintaining stability of perception and coordination. *Journal of Nervous and Mental Diseases*, 1961, **132**, 26–32.

Held, R., Dissociation of visual functions by deprivation and rearrangement. *Psychologische Forschung*, 1967, **31**, 388–348.

Held, R. and Hein, A., Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 1963, **56**, 872–876.

Hirsch, M., Systems of differential equations which are competitive or cooperative, I: Limit sets. Preprint, 1982 (a).

Hirsch, M., Systems of differential equations which are competitive or cooperative, II: Convergence almost everywhere. Preprint, 1982 (b).

Hubel, D.H. and Wiesel, T.N., Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London (B)*, 1977, **198**, 1–59.

Hunt, R.K. and Jacobson, M., Specification of positional information in retinal ganglion cells of *Xenopus laevis*: Intraocular control of the time of specification. *Proceedings of the National Academy of Sciences*, 1974, **71**, 3616–3620.

Jung, J., **Verbal learning**. New York: Holt, Rinehart, and Winston, 1968.

Katz, B., **Nerve, muscle, and synapse**. New York: McGraw-Hill, 1966.

Khinchin, A.I., **Mathematical foundations of information theory**. New York: Dover Press, 1967.

Klatsky, R.L., **Human memory: Structures and processes**. San Francisco: W.H. Freeman, 1980.

Kuffler, S.W. and Nicholls, J.G., **From neuron to brain**. Sunderland, MA: Sinauer Press, 1976.

Loftus, G.R. and Loftus, E.F., **Human memory: The processing of information**. Hillsdale, NJ: Erlbaum, 1976.

Maher, B.A., **Contributions to the psychopathology of schizophrenia**. New York: Academic Press, 1977.

Malsburg, C. von der, Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 1973, **14**, 85–100.

Malsburg, C. von der and Willshaw, D.J., How to label nerve cells so that they can interconnect in an ordered fashion. *Proceedings of the National Academy of Sciences*, 1977, **74**, 5176–5178.

Malsburg, C. von der and Willshaw, D.J., Differential equations for the development of topological nerve fibre projections. In S. Grossberg (Ed.), **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society, 1981.

May, R.M. and Leonard, W.J., Nonlinear aspects of competition between three species. *SIAM Journal on Applied Mathematics*, 1975, **29**, 243–253.

McGeogh, J.A. and Irion, A.L., **The psychology of human learning**, Second Edition. New York: Longmans and Green, 1952.

Meinhardt, H. and Gierer, A., Applications of a theory of biological pattern formation based on lateral inhibition. *Journal of Cell Science*, 1974, **15**, 321–346.

Murdock, B.B., **Human memory: Theory and data**. Potomac, MD: Erlbaum, 1974.

Norman, D.A., **Memory and attention: An introduction to human information processing**. New York: Wiley and Sons, 1969.

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

Pettigrew, J.D. and Kasamatsu, T., Local perfusion of noradrenaline maintains visual cortical plasticity. *Nature*, 1978, **271**, 761–763.

Plonsey, R., **Bioelectric phenomena**. New York: McGraw-Hill, 1969.

Ruch, T.C., Patton, H.D., Woodbury, J.W., and Towe, A.L., **Neurophysiology**. Philadelphia: Saunders, 1961.

Schneider, W. and Shiffrin, R.M., Automatic and controlled information processing in vision. In D. LaBarge and S.J. Samuels (Eds.), **Basic processes in reading: Perception and comprehension**. Hillsdale, NJ: Erlbaum, 1976.

Stein, P.S.G., Intersegmental coordination of swimmeret and motoneuron activity in crayfish. *Journal of Neurophysiology*, 1971, **34**, 310–318.

Underwood, B.J., **Experimental psychology**, Second Edition. New York: Appleton-Century-Crofts, 1966.

Wallach, H. and Karsh, E.B., Why the modification of stereoscopic depth-perception is so rapid. *American Journal of Psychology*, 1963, **76**, 413–420 (a).

Wallach, H. and Karsh, E.B., The modification of stereoscopic depth-perception and the kinetic depth-effect. *American Journal of Psychology*, 1963, **76**, 429–435 (b).

Wallach, H., Moore, M.E., and Davidson, L., Modification of stereoscopic depth-perception. *American Journal of Psychology*, 1963, **76**, 191–204.

Willshaw, D.J. and Malsburg, C. von der, How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London (B)*, 1976, **194**, 431–445.

Wolpert, L., Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*, 1969, **25**, 1–47.

Young, R.K., Serial learning. In T.R. Dixon and D.L. Horton (Eds.), **Verbal behavior and general behavior theory**. Englewood Cliffs, NJ: Prentice-Hall, 1968.

THE ADAPTIVE BRAIN I

Cognition, Learning, Reinforcement, and Rhythm

Edited by

Stephen GROSSBERG
Center for Adaptive Systems
Boston University
Boston, Massachusetts
U.S.A.

1988

NORTH-HOLLAND
AMSTERDAM · NEW YORK · OXFORD · TOKYO