(1976)
S. Grossberg

Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors

Biological Cybernetics 23:121–134

Abstract. This paper analyses a model for the parallel development and adult coding of neural feature detectors. The model was introduced in Grossberg (1976). We show how experience can retune feature detectors to respond to a prescribed convex set of spatial patterns. In particular, the detectors automatically respond to average features chosen from the set even if the average features have never been experienced. Using this procedure, any set of arbitrary spatial patterns can be recoded, or transformed, into any other spatial patterns (universal recoding), if there are sufficiently many cells in the network's cortex. The network is built from short term memory (STM) and long term memory (LTM) mechanisms, including mechanisms of adaptation, filtering, contrast enhancement, tuning, and nonspecific arousal. These mechanisms capture some experimental properties of plasticity in the kitten visual cortex. The model also suggests a classification of adult feature detector properties in terms of a small number of functional principles. In particular, experiments on retinal dynamics, including amacrine cell function, are suggested.

1. Introduction

This paper analyses a model for the development of neural feature detectors during an animal's early experience with its environment. The model also suggests mechanisms of adult pattern discrimination that remain after development has been completed. The model evolved from earlier experimental and theoretical work. Various data showed that there is a critical period during which experimental manipulations can alter the patterns to which feature detectors in the visual cortex are tuned (e.g., Barlow and Pettigrew, 1971; Blakemore and Cooper, 1970; Blakemore and Mitchell, 1973; Hirsch and Spinelli, 1970, 1971; Hubel and Wiesel, 1970; Wiesel and Hubel, 1963, 1965). This work led Von der Malsburg (1973) and Pérez et al. (1974) to construct models of the cortical tuning process, which they analysed using computer methods. Their models are strikingly similar. Both use a mechanism of long term memory (LTM) to encode changes in tuning. This mechanism learns by classical, or Pavlovian, conditioning (Kimble, 1967) within a neural network. Such a concept was qualitatively described by Hebb (1949) and was rigorously analysed in its present form by Grossberg (e.g., 1967, 1970a, 1971, 1974). The LTM mechanism in a given interneuronal pathway is a plastic synaptic strength which has two crucial properties: (a) it is computed from a time average of the product of presynaptic signals and postsynaptic potentials; (b) it multiplicatively gates, or shunts, a presynaptic signal before it can perturb the postsynaptic cell.

Given this LTM mechanism, both models invoke various devices to regulate the retinocortical signals that drive the tuning process. On-center off-surround networks undergoing additive interactions, attenuation of small retinocortical signals at the cortex, and conservation of the total synaptic strength impinging on each cortical cell are used in both models. Grossberg (1976) realized that all of these mechanisms for distributing signals could be replaced by a minimal model for parallel processing of patterns in noise, which is realized by an on-center off-surround recurrent network whose interactions are of shunting type (Grossberg, 1973). Three crucial properties of this model are: (a) normalization, or adaptation, of total network activity; (b) contrast enhancement of input patterns; and (c) short term memory (STM) storage of the contrast-enhanced pattern. Using these properties, Grossberg (1976) eliminates the conservation of total synaptic strength—which is incompatible with

Supported in part by the Advanced Research Projects Agency under ONR Contract No. N00014-76-C-0185

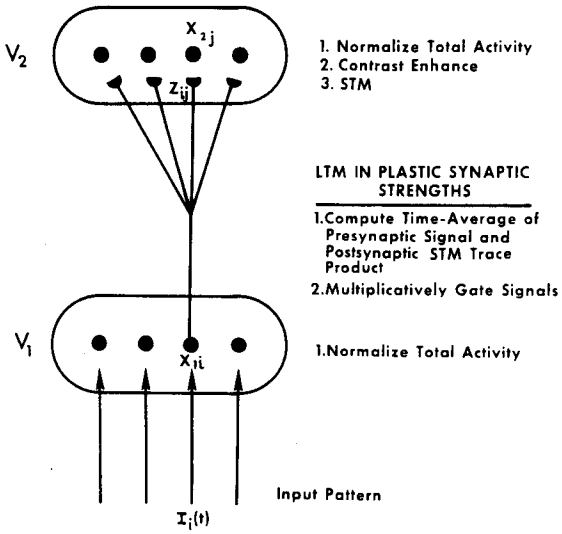

Fig. 1. Minimal model of developmental tuning using STM and LTM mechanisms

classical conditioning—and shows that the tuning process can be derived from *adult* STM and LTM principles. The model is schematized in Figure 1. It describes the interaction via plastic synaptic pathways of two network regions, V_1 and V_2, that are separately capable of normalizing patterns, but V_2 can also contrast enhance patterns and store them in STM. In the original models of Von der Malsburg and Pérez et al., V_1 was interpreted as a "retina" or "thalamus" and V_2 as "visual cortex". In Part II, an analogous anatomy for V_1 as "olfactory bulb" and V_2 as "prepyriform cortex" will be noted. In Section 5, a more microscopic analysis of the model leads to a discussion of V_1 as a composite of retinal receptors, horizontal cells, and bipolar cells, and of V_2 as a composite of amacrine cells and ganglion cells. Such varied interpretations are possible because the same functional principles seem to operate in various anatomies.

Using this abstract structure, it was suggested in Grossberg (1976) how hierarchies of cells capable of discriminating arbitrary spatial patterns can be synthesized. Also a striking analogy was described between the structure and properties of certain reaction-diffusion systems that have been used to model development (Gierer and Meinhardt, 1972; Meinhardt and Gierer, 1974) and of reverberating shunting networks. This paper continues this program by rigorously analysing mathematical properties of the model, which thereupon suggest other developmental and adult STM and LTM mechanisms that are related to it. The following sections will describe these connections with a minimum of mathematical detail. Mathematical proofs are contained in the Appendix.

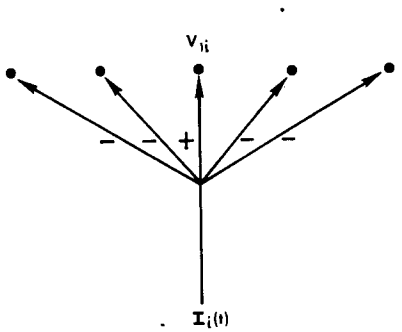

Fig. 2. Nonrecurrent, or feedforward, on-center off-surround network

2. The Tuning Process

This section reviews properties of the model that will be needed below. Suppose that V_1 consists of n states (or cells, or cell populations) v_{1i}, $i=1, 2, \ldots, n$, which receive inputs $I_i(t)$ whose intensity depends on the presence of a prescribed feature, or features, in an external pattern. Let the population response (or activity, or average potential) of v_{i1} be $x_{1i}(t)$. The relative input intensity $\Theta_i = I_i I^{-1}$, where $I = \sum_{k=1}^{n} I_k$, measures the relative importance of the feature coded by v_i in any given input pattern. If the Θ_i's are constant during a given time interval, the inputs are said to form a *spatial pattern*. How can the laws governing the $x_{1i}(t)$ be determined so that $x_{1i}(t)$ is capable of accurately registering Θ_i? Grossberg (1973) showed that a bounded, linear law for x_{1i}, in which x_{1i} returns to equilibrium after inputs cease, and in which neither input pathways nor populations v_{1i} interact, does not suffice; cf., Grossberg and Levine (1975) for a review. The problem is that as the total input I increases, given *fixed* Θ_i values, each x_{1i} saturates at its maximal value. This does not happen if off-surround interactions also occur. For example, let the inputs I_i be distributed via a nonrecurrent, or feedforward, on-center off-surround anatomy undergoing shunting (or mass action, or passive membrane) interactions, as in Figure 2. Then

$$\dot{x}_{1i} = -Ax_{1i} + (B - x_{1i})I_i - x_{1i} \sum_{k \neq i} I_k \tag{1}$$

with $0 \leqq x_{1i}(0) \leqq B$. At equilibrium (namely, $\dot{x}_{1i} = 0$),

$$x_{1i} = \Theta_i \frac{BI}{A + I}, \tag{2}$$

which is proportional to Θ_i no matter how large I becomes. Since also $BI(A+I)^{-1} \leqq B$, the total activity $x_1 \equiv \sum_{k=1}^{n} x_{1k}$ never exceeds B; it is normalized, or adapts, due to automatic gain control by the in-

hibitory inputs. The normalization property in (2) shows that x_{1i} codes Θ_i rather than instantaneous fluctuations in I.

To store patterns in STM, recurrent or feedback pathways are needed to keep signals active after the inputs cease. Again the problem of saturation must be dealt with, so that some type of recurrent on-center off-surround anatomy is suggested. The minimal solution is to let V_2 be governed by a system of the form

$$\dot{x}_{2j} = -Ax_{2j} + (B - x_{2j})[f(x_{2j}) + I_{2j}] - x_{2j} \sum_{k \neq j} f(x_{2k}), \tag{3}$$

where $f(w)$ is the average feedback signal produced by an average activity level w, and I_{2j} is the total excitatory input to v_{2j} (Fig. 3a). In particular, v_{2j} excites itself via the term $(B - x_{2j})f(x_{2j})$, and v_{2k} inhibits v_{2j} via the term $-x_{2j}f(x_{2k})$, for every $k \neq j$. The choice of $f(w)$ dramatically influences how recurrent interactions within V_2 transform the input pattern $I^{(2)} = (I_{21}, I_{22}, \ldots, I_{2N})$ through time. Grossberg (1973) shows that a sigmoid, or S-shaped, $f(w)$ can reverberate important inputs in STM after contrast-enhancing them, yet can also suppress noise.

Various generalizations of recurrent networks have been studied, such as

$$\dot{x}_{2j} = -Ax_{2j} + (B - x_{2j})\left[\sum_{k=1}^{N} f(x_{2k})C_{kj} + I_{2j}\right] - (x_{2j} + D)\sum_{k=1}^{N} f(x_{2k})E_{kj}, \tag{4}$$

$D \geqq 0$, where the excitatory coefficients C_{kj} ("on-center") decrease with the distance between populations v_{2k} and v_{2j} more rapidly than do the inhibitory coefficients E_{kj} ("off-surround"). Levine and Grossberg (1976) show that, in such cases, the inhibitory off-surround signals $\sum_{k=1}^{N} f(x_{2k})E_{kj}$ to v_{2j} can be chosen strong enough to offset the saturating effects of inputs I_{2j} plus excitatory on-center signals $\sum_{k=1}^{N} f(x_{2k})C_{kj}$. Ellias and Grossberg (1975) study generalizations of (4) in which inhibitory interneurons interact with their excitatory counterparts.

Below we will consider networks in which the excitatory signals I_{2j} to V_2 are sums of signals from many populations in V_1. Moreover, the synaptic strengths of these signals can be trained. This fact suggests another reason for making V_2 recurrent. A recurrent anatomy is needed within V_2 to prevent saturation in response to trainable signals. To see this, note in the nonrecurrent network (1) that each excitatory input to v_{1i} is replicated as an inhibitory input to all v_{1k}, $k \neq i$. The size of a trainable signal to v_{2j} depends on the activity at v_{2j}. This signal therefore cannot be replicated at populations v_{2k}, $k \neq j$, unless recurrent interactions within V_2 exist. Moreover, whether or not signals are trainable, whenever I_{2j} is a sum of signals from many populations, recurrent signals within V_2 prevent saturation at a large saving of extra signal pathways to the populations v_{2k}, $k \neq j$.

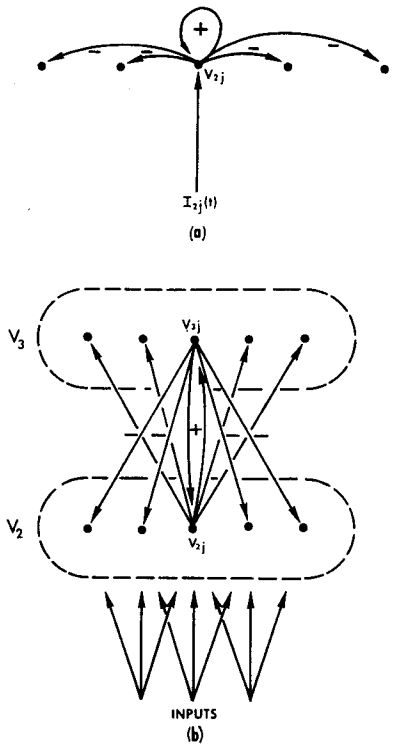

Fig. 3. Some recurrent, or feedback, on-center off-surround networks

A related scheme for marrying sums of (trainable) signals with pattern normalization is illustrated in Figure 3b. Here a sum of signals I_{2j} from V_1 perturbs each v_{2j}. Population v_{2j} thereupon excites an on-center of cells near v_{3j}, and inhibits a broad off-surround of populations centered at v_{3j}. Thus, when a pattern $I^{(2)}$ arrives at V_2, it is normalized at V_3 before saturation can take place across V_2. Then feedback signals from V_3 to V_2 prevent saturation at V_2 from setting in as follows. Each population v_{3j} that receives a large net excitatory signal from V_2 excites its on-center of cells near v_{2j}, and inhibits a broad off-surround of populations centered at v_{2j}. This feedback inhibition prevents the pattern $I^{(2)}$ from saturating V_2, much as recurrent inhibition in Equation (4) works. Figure 3b can also be expanded to explicitly include inhibitory interneurons, as in Ellias and Grossberg (1975).

Normalization in V_1 by (1) occurs gradually in time, as each x_{1i} adjusts to its new equilibrium value, but

it will be assumed below to occur instantaneously with x_{1i} approaching Θ_i rather than $\Theta_i BI(A+I)^{-1}$. These simplifications yield theorems about the tuning process that avoid unimportant details. The assumption that normalization occurs instantaneously is tenable because the normalized pattern at V_1 drives slow changes in the strength of connections from V_1 to V_2. Instantaneous normalization means that the pattern at V_1 normalizes itself before the connection strengths have a chance to substantially change.

Let the synaptic strength of the pathway from v_{1i} to the j^{th} population v_{2j} in V_2 be denoted by $z_{ij}(t)$ (see Fig. 1). Let the total signal to v_{2j} due to the normalized pattern $\Theta=(\Theta_1, \Theta_2, \ldots, \Theta_n)$ at V_1 and the vector $z^{(j)}(t)=(z_{1j}(t), z_{2j}(t), \ldots, z_{nj}(t))$ of synaptic strengths be

$$S_j(t) \equiv \Theta \cdot z^{(j)}(t) \equiv \sum_{k=1}^{n} \Theta_k z_{kj}(t); \tag{5}$$

that is, each $z_{kj}(t)$ *gates* the signal Θ_k from v_{1k} on its way to v_{2j}, and these gated signals combine additively at v_{2j} (cf., Grossberg, 1967, 1970a, 1971, 1974). Since $z^{(j)}(t)$ determines the size of the input to v_{2j}, given any pattern Θ, it is called the *classifying vector* of v_{2j} at time t. *Every* v_{2j}, $j=1, 2, \ldots, N$, in V_2 receives such a signal when Θ is active at V_1. In this way, Θ creates a pattern of activity across V_2.

Given any activity pattern across V_2, it can be transformed in several ways as time goes on. Two main questions about this process are: (a) will the *total* activity of V_2 be suppressed, or will some of its activities be stored in STM? and (b) which of the *relative* activities across V_2 will be preserved, suppressed, or enhanced? Several papers (Ellias and Grossberg, 1975; Grossberg, 1973; Grossberg and Levine, 1975) analyse how the parameters of a reverberating shunting on-center off-surround network determine the answers to these questions. Below some of these facts are cited as they are needed. In particular, if all the activities are sufficiently small, then they will not be stored in STM. If they are sufficiently large, then they will be contrast enhanced, normalized, and stored in STM. Figure 4 schematizes two storage possibilities. Figure 4a depicts a pattern of activity across V_2 before it is transformed by V_2. Given suitable parameters, if some of the initial activities exceed a quenching threshold (QT), then V_2 will *choose* the population having maximal initial activity for storage in STM, as in Figure 4b. Under other circumstances, all initial activities below the QT are suppressed, whereas *all* initial activities above QT are contrast enhanced, normalized, and stored in STM (Fig. 4c); that is, *partial* contrast in STM is possible. Grossberg (1973) shows that partial contrast can occur if the signals between populations in a recurrent shunting on-center off-surround network are sigmoid (*S*-shaped) functions of

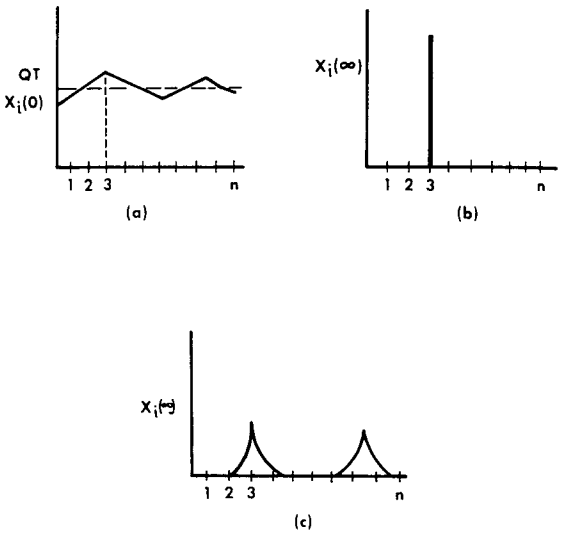

Fig. 4. Contrast enhancement and STM by recurrent network: (a) initial pattern; (b) choice; (c) partial contrast

their activity levels. Ellias and Grossberg (1975) show that partial contrast can occur if the self-excitatory signals of populations in V_2 are stronger than their self-inhibitory signals, and moreover if the excitatory signals between populations in V_2 decrease with inter-population distance faster than the inhibitory signals.

The enhancement and STM storage processes also occur much faster than the slow changes in connection strengths z_{ij}; hence, it is assumed below that these processes occur instantaneously in order to focus on the slow changes in z_{ij}.

The slow changes in z_{ij} are assumed to be determined by a time averaged product of the signal from v_{1i} to v_{2j} with the cortical response at v_{2j}; thus

$$\dot{z}_{ij} = -C_{ij} z_{ij} + D_{ij} x_{2j},$$

where C_{ij} is the decay rate (possibly variable) of z_{ij}, and D_{ij} is the signal from v_{1i} to v_{2j}. For example, if $C_{ij}=1$, the V_1 and V_2 patterns are normalized, and V_2 chooses only the population v_{2j} whose initial activity is maximal for storage in STM (Fig. 4b), then while v_{2j} is active,

$$\dot{z}_{ij} = -z_{ij} + \Theta_i, \quad \text{for all } i=1, 2, \ldots, n.$$

It remains to determine how these z_{ij} and all other z_{ik}, $k \neq j$, change under other circumstances. To eliminate conceptual and mathematical difficulties that arise if z_{ij} can decay even when V_1 and V_2 are inactive, we let *all* changes in each z_{ij} be determined by which populations in V_2 have their activities chosen for storage in STM. In other words, all changes in z_{ij} are driven by the *feedback* within the excitatory re-

current loops of V_2 that establish STM storage. Then

$$\dot{z}_{ij}=(-z_{ij}+\Theta_i)x_{2j} \tag{6}$$

where $\sum_{k=1}^{N} x_{2k}(t)=1$ if STM in V_2 is active at time t, whereas $\sum_{k=1}^{N} x_{2k}(t)=0$ if STM in V_2 is inactive at time t.

If V_2 *chooses* a population for storage in STM, as in Figure 4b, then

$$x_{2j}=\begin{cases}1 \text{ if } S_j>max\{\varepsilon, S_k:k\neq j\}\\ 0 \text{ if } S_j<\max\{\varepsilon, S_k:k\neq j\},\end{cases} \tag{7}$$

where as in (5), $S_j=\Theta\cdot z^{(j)}$ with $\Theta_i=I_i\left(\sum_{k=1}^{n} I_k\right)^{-1}$

Equation (7) omits the cases where two or more signals S_j are equal, and are larger than all other signals and ε. In these cases, the x_{2j}'s of such S_j's are equal and add up to 1. Such a normalization rule for equal maximal signals will be tacitly assumed in all the cases below, but will otherwise be ignored to avoid tedious details. Equation (6) shows that z_{ij} can change only if $x_{2j}>0$. Equation (7) shows that V_2 chooses the maximal activity for storage in STM. This activity is normalized ($x_{2j}=0$ or 1), and it corresponds to the population with largest initial signal ($S_j>\max\{S_k:k\neq j\}$). No changes in z_{ij} occur if all signals S_j are too small to be stored in STM (all $S_j\leqq\varepsilon$).

If partial contrast in STM holds, as in Figure 4c, then the dynamics of a reverberating shunting network can be approximated by a rule of the form

$$x_{2j}=\begin{cases}f(S_j)\left[\sum_{S_k>\varepsilon} f(S_k)\right]^{-1} & \text{if } S_j>\varepsilon\\ 0 & \text{if } S_j<\varepsilon\end{cases} \tag{8}$$

where $f(w)$ is an increasing nonnegative function of w such that $w=0$; e.g., $f(w)=w^2$. In (8), the positive constant ε represents the QT; the function $f(w)$ controls how suprathreshold signals S_j will be contrast enhanced; and the ratio of $f(S_j)$ to $\sum\{f(S_k):S_k>\varepsilon\}$ expresses the normalization of STM.

3. Ritualistic Pattern Classification

After developmental tuning has taken place, the above mechanisms describe a model of pattern classification in the "adult" network. These mechanisms will be described first as interesting in themselves, and as a helpful prelude to understanding the tuning process. They are capable of classifying arbitrarily complicated spatial patterns into mutually nonoverlapping, or partially overlapping, sets depending on whether (7) or (8) holds. These mechanisms realize basic principles of pattern discrimination using shunting interactions. An alternative scheme of pattern discrimination using a mixture of shunting and additive mechanisms has already been given (Grossberg, 1970b, 1972). Together these schemes suggest numerous anatomical and physiological variations that embody the same small class of functional principles. Since particular anatomies imply that particular physiological rules should be operative, intriguing questions about the dynamics of various neural structures, such as retina, neocortex, hippocampus, and cerebellum, are suggested.

First consider what happens if V_2 chooses a population for storage in STM. After learning ceases (that is, $\dot{z}_{ij}\equiv 0$), all classifying vectors $z^{(j)}$ are constant in time, and Equations (6) and (7) reduce to the statement that population v_{2j} is stored in STM if

$$S_j>\max\{\varepsilon, S_k:k\neq j\}. \tag{9}$$

In other words, v_{2j} *codes* all patterns Θ such that (9) holds; alternatively stated, v_{2j} is a *feature detector* in the sense that all patterns

$$P_j=\{\Theta:\Theta\cdot z^{(j)}>\max(\varepsilon, \Theta\cdot z^{(k)}:k\neq j)\} \tag{10}$$

are classified by v_{2j}. The set P_j defines a *convex cone* C_j in the space of nonnegative input vectors $J=(I_1, I_2, \ldots, I_n)$, since if two such vectors $J^{(1)}$ and $J^{(2)}$ are in C_j, then so are all the vectors $\alpha J^{(1)}$, $\beta J^{(2)}$, and $\gamma J^{(1)}+(1-\gamma)J^{(2)}$, where $\alpha>0$, $\beta>0$, and $0<\gamma<1$. The convex cone C_j defines the *feature* coded by v_{2j}.

The classification rule in (10) has an informative geometrical interpretation in n-dimensional Euclidean space. The signal $S_j=\Theta\cdot z^{(j)}$ is the inner product of Θ and $z^{(j)}$ (Greenspan and Benney, 1973). Letting $\|\xi\|=\sqrt{\sum_{k=1}^{n}\xi_k^2}$ denote the Euclidean length of any real vector $\xi=(\xi_1, \xi_2, \ldots, \xi_n)$, and $\cos(\eta, \omega)$ denote the cosine between two vectors η and ω, it is elementary that

$$S_j=\|\Theta\|\,\|z^{(j)}\|\cos(\Theta, z^{(j)}).$$

In other words, the signal S_j is the length of the projection of the normalized pattern Θ on the classifying vector $z^{(j)}$ times the length of $z^{(j)}$. Thus if all $z^{(j)}$, $j=1, 2, \ldots, N$, have equal length, then among all patterns with the same length, (10) classifies all patterns Θ in P_j whose angle with $z^{(j)}$ is smaller than the angles between Θ and any $z^{(k)}$, $k\neq j$, and is small enough to satisfy the ε-condition. In particular, patterns Θ that are *parallel* to $z^{(j)}$ are classified in P_j. The choice of classifying vectors $z^{(j)}$ hereby determines how the patterns Θ will be divided up. Section 8 will show that the tuning mechanism (6)–(7) makes the $z^{(j)}$ vectors more parallel to prescribed patterns Θ, and thereupon changes the classifying sets P_j. In summary

(i) the number of populations in V_2 determines the maximum number N of pattern classes P_j;

(ii) the choice of classifying vectors $z^{(j)}$ determines

how different these classes can be; for example, choosing all vectors $z^{(j)}$ equal will generate one class that is redundantly represented by all v_{2j}; and

(iii) the size of ε determines how similar patterns must be to be classified by the same v_{2j}.

If the choice rule (7) is replaced by the partial contrast rule (8), then an important new possibility occurs, which can be described either by studying STM responses to all Θ at fixed v_{2j}, or to a fixed Θ at all v_{2j}. In the former case, each v_{2j} has a *tuning curve*, or *generalization gradient*; namely, a maximal response to certain patterns, and submaximal responses to other patterns. In the latter case, each pattern Θ is *filtered* by V_2 in a way that shows how close Θ lies to *each* of the classifying vectors $z^{(j)}$. The pattern will only be classified by v_{2j}—that is, stored in STM—if it lies sufficiently close to $z^{(j)}$ for its signal S_j to exceed the quenching threshold of V_2.

For example, suppose that some of the classifying vectors $z^{(j)}$ are chosen to create large signals at V_2 when vertical lines perturb V_1, and that other $z^{(j)}$ create large signals at V_2 when horizontal lines perturb V_1. If a pattern containing both horizontal and vertical lines perturbs V_1, then the population activities in V_2 corresponding to both types of lines can be stored in STM, unless competition between their populations drives all activity below the QT. Now let V_3 be another "cortex" that receives signals from V_2, in the same fashion that V_2 receives signals from V_1. Given an appropriate choice of classifying vectors for V_3, there can exist cells in V_3 that fire in STM only if horizontal *and* vertical lines perturb a prescribed region of V_1; e.g., hypercomplex cells. The existence of tuning curves in a given cortex V_i hereby increases the discriminative capabilities of the next cortex V_{i+1} in a hierarchy; cf., Grossberg (1976).

The above mechanisms will now be discussed as cases of a general scheme of pattern classification. This is done with two goals in mind: firstly, to emphasize that these mechanisms might well exist in other than "retinocortical" analogs; and secondly, to generate explicit experimental directives in a variety of neural structures. One such directive will be described in Section 5.

4. Shunts vs. Additive Interactions as Mechanisms of Pattern Classification

The processing stages utilized in Section 3 are the following:

A) Normalization

Input patterns are normalized in V_1 by an on-center off-surround anatomy undergoing shunting interactions.

B) Partial Filtering by Signals

The signals S_j generated at V_2 by a normalized pattern on V_1 create the data base on which later computations are determined. The signal generating rule (5), for example, has the following important property. Suppose that an input $I_i(t)=\Theta_i I(t)$ is normalized to x_{1i}, as in (2), rather than to the approximate value Θ_i. The signal from V_1 to v_{2j} becomes

$$\tilde{S}_j = BI(A+I)^{-1}S_j$$

and (9) is replaced by the analogous rule

$$\tilde{S}_j > \max\{\varepsilon, \tilde{S}_k : k \neq j\}.$$

Then V_2 will classify a given pattern into the same class P_j no matter how large I is chosen. In other words, the signal generating rule is invariant under suprathreshold variations of the total activity at V_1. If I_i is the transduced receptor response to an external input J_i—that is, $I_i = g(J_i)$—then the signal-generating rule is invariant, given *any* $z^{(j)}$'s, if $g(w)=w^p$ for some $p>0$.

C) Contrast Enhancement of Signals

The signals S_j are contrast enhanced by the recurrent on-center off-surround anatomy within V_2, and either a choice (Fig. 4b) or a tuning curve (Fig. 4c) results.

Two successive stages of lateral inhibition are needed in this model. The first stage normalizes input patterns. The second stage sharpens the filtering of signals.

Additive mechanisms can also achieve classification of arbitrarily complicated spatial patterns. These mechanisms also employ three successive stages A)–C) of pattern processing, with stage A) normalizing input patterns, stages A) and C) using inhibitory interactions, and stage C) completing the pattern classification that is begun by the signal generating rules of stage B). The additive model can differ in several respects from the shunting model:

(i) its anatomy can be feedforward; that is, there need not be a recurrent network in stage C);

(ii) threshold rules replace the inner product signal-generating rule (5) to determine partial filtering of signals; and

(iii) the responses in time of stages A)–C) to a sustained pattern at V_1 are not the same in the additive model. For example, sustained responses in the shunting model can be replaced by responses to the onset and offset of the pattern in the additive model (Grossberg, 1970b).

Mixtures of additive and shunting mechanisms are also possible. The additive mechanisms will now be summarized to illustrate the basic stages A)–C).

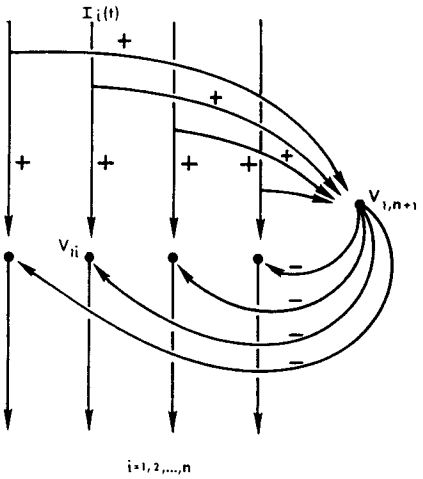

Fig. 5. Normalization and low-band filtering by subtractive nonspecific interneuron and signal threshold rules

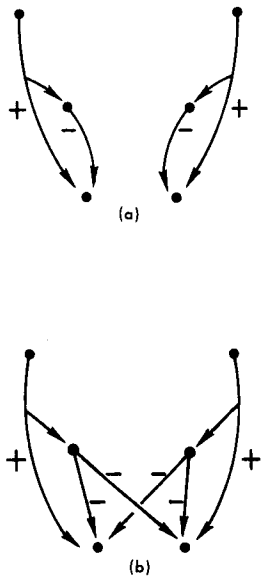

Fig. 6. (a) Specific subtractive inhibitory interneurons: (b) Nonspecific inhibitory interneurons

An additive nonspecific inhibitory interneuron normalizes patterns at V_1 (Fig. 5). Many variations on this theme exist (Grossberg, 1970b) in which such parameters as the lateral spread of inhibition, the number of cell layers, and the rates of excitatory and inhibitory decay can be varied. The idea in its simplest form is this. The excitatory input I_i excites a bifurcating pathway. One branch of the pathway is specific, and the other branch is nonspecific. The lateral inhibitory interneuron $v_{1,n+1}$ lies in the nonspecific branch. It sums the excitatory inputs I_i, and generates a nonspecific signal back to all the specific pathways if a signal threshold Γ is exceeded. Each input I_i also generates a specific signal from v_{1i} that is a linear function of I_i above a signal threshold. Each pathway from v_{1i} in V_1 to v_{2j} in V_2 has its own signal threshold Γ_{ij}. The net signal from v_{1i} to v_{2j} is

$$K_{ij}=[I_i-\Gamma_{ij}]^+-\left[\sum_{k=1}^{n} I_k-\Gamma\right]^+,$$

where the notation $[u]^+=\max(u,0)$ defines the threshold rule. Define $\Theta_{ij}=\Gamma_{ij}\Gamma^{-1}$ and let the spatial pattern $I_i=\Theta_i I$ perturb V_1. Then

$$K_{ij}=[\Theta_i I-\Theta_{ij}\Gamma]^+-[I-\Gamma]^+. \tag{11}$$

The net signal K_{ij} has the following properties:

(i) $K_{ij}\leqq 0$ for all values of $I>0$ if $\Theta_i\leqq\Theta_{ij}$;

(ii) $K_i>0$ for $I>\Theta_{ij}\Theta_i^{-1}$ if $\Theta_i>\Theta_{ij}$; and

(iii) $K_{ij}\leqq(\Theta_i-\Theta_{ij})\Gamma$ for all $I>0$.

In other words, by (i), no signal is emitted from v_{1i} to v_{2j} if $\Theta_i<\Theta_{ij}$; by (ii), if $\Theta_i>\Theta_{ij}$, a signal is emitted from v_{1i} if I exceeds a threshold depending on Θ_i and Θ_{ij}; and by (iii), the total activity in the cells v_{1i} is normalized. Partial filtering of signals is thus achieved by the choice of threshold pattern $\Theta^{(j)}=(\Theta_{1j},\Theta_{2j},\ldots,\Theta_{nj})$ rather than by the choice of classifying vector $z^{(j)}=(z_{1j},z_{2j},\ldots,z_{nj})$.

Stage C) is needed because the total signal to v_{2j} can be maximized by patterns Θ which are very different from the threshold pattern $\Theta^{(j)}$. This problem arises because the signals K_{ij} continue to grow linearly as a function of I after the threshold value $\Theta_{ij}\Theta_i^{-1}$ is exceeded. Grossberg (1970b) shows that the problem can be avoided by inhibiting each signal K_{ij} if it gets too large. For example, let the net signal from v_{1i} to v_{2j} be

$$S_{ij}^*=K_{ij}-\alpha[K_{ij}-\beta]^+, \tag{12}$$

where $\alpha>1$ and $0<\beta\ll 1$. This mechanism inhibits the signal from v_{1i} to v_{2j} if it represents a Θ_i which is too much larger than Θ_{ij}. Equation (12) can be realized by any of the several inhibitory mechanisms: a specific subtractive inhibitory interneuron (Fig. 6a), a switchover from net excitation to net inhibition when the spiking frequency in the pathway from v_{1i} to v_{2j} becomes too large (Bennett, 1971; Blackenship et al., 1971; Wachtel and Kandel, 1971), or postsynaptic blockade of the v_{2j} cell membrane at sufficiently high spiking frequencies. Signal S_{ij}^* is positive only if Θ_i is sufficiently close to Θ_{ij} in value. Stage C) is completed by choosing the signal threshold of v_{2j} so high that v_{2j} only fires if *all* signals S_{ij}^*, $i=1,2,\ldots,n$, are positive; that is, only if the input pattern Θ is close to the threshold pattern $\Theta^{(j)}$. The second stage of

inhibition hereby completes the partial filtering process by choosing a population v_{2j} in V_2 to code $\Theta^{(j)}$, as in Figure 4b. If the specific inhibitory interneurons if Figure 6a are replaced by a lateral spread of inhibition, as in Figure 6b, then a tuning curve is generated, as in Figure 4c.

5. What Do Retinal Amacrine Cells Do?

This section illustrates how the principles A)–C) can generate interesting questions about particular neural processes. Grossberg (1970b, 1972) introduces a retinal model in which shunting and additive interactions both occur. In this model, retinal amacrine cells are examples of the inhibitory interaction in stage C). We will note that amacrine cells have *opposite* effects on signals if they realize a shunting rather than an additive model. In the retinal model of Grossberg (1972), normalization is accomplished by an on-center off-surround anatomy undergoing shunting interactions. Analogously, in vivo receptors excite bipolar cells (on-center) as well as horizontal cells, and the horizontal cells inhibit bipolar cells via their lateral interactions (off-surround). Partial filtering of the normalized inputs is accomplished by signal thresholds; for example, using the normalized x_{1i} activities in (2), the simplest signal function from v_{1i} to v_{2j} is $K_{ij}=[x_{1i}-\Gamma_{ij}]^+$. Stage C) is then accomplished by a mechanism such as (12), by which large signals are inhibited. Whether a choice (Fig. 4b) or a tuning curve (Fig. 4c) is generated depends, in part, on how broadly these lateral inhibitory signals that complete stage C) are distributed. This second stage of inhibition is identified with the inhibition that amacrine cells, fed by bipolar cell activity, generate at ganglion cells. Grossberg (1972) notes data that support the idea that stage C) is realized by an additive mechanism such as (12). In particular, amacrine cells often respond when an input pattern is turned on, or off, or both. Two questions about amacrine cells now suggest themselves.

(i) If this interpretation of amacrine cells is true, then they will shut off signals from the bipolar cells to the ganglion cells when these signals become too *large*; that is, they act as high-band filters. By contrast, inhibition in stage C) of the shunting model shuts off signals if they become too *small*. Opposite effects due to the second inhibitory stage can hereby create a similar functional transformation of the input pattern! If a shunting role for amacrine cells is sought, then the following types of anatomy would be anticipated: inhibitory bipolar-to-amacrine-to-bipolar cell feedback that contrast enhances the receptor-to-bipolar signals, or inhibitory ganglion-to-amacrine-to-ganglion cell feedback that contrast enhances the bipolar-to- ganglion cell signals, or some functionally similar feedback loop. To decide between these two possible roles for amacrine cells, one must test whether amacrine cells suppress large signals or small ones; in either case, if the model is applicable, contrast enhancement of the normalized and filtered retinal pattern is the result, so that this property cannot be used as a criterion.

(ii) Does the spatial extent of lateral amacrine interaction determine the amount of contrast, or the breadth of the tuning curves, in ganglion cell responses, as in Figures 4b and 4c? For example, there exist narrow field diffuse amacrine cells, wide field diffuse amacrine cells, stratified diffuse amacrine cells, and unstratified amacrine cells (Boycott and Dowling, 1969). Do these specializations guarantee particular tuning characteristics in the corresponding ganglion cells?

Grossberg (1972) also suggests a cerebellar analog based on the same principles. Thus at least formal aspects of various neural structures seem to be emerging as manifestations of common principles. These results suggest a program of classifying seemingly different anatomical and physiological data according to whether they realize similar functional transformations of patterned neural activity, such as total activity normalization, partial filtering by signals, and contrast enhancement of the signal pattern. Below are described certain properties of the shunting mechanism that will be needed when development is discussed.

6. Arousal as a Tuning Mechanism

The recurrent networks in V_2 all have a quenching threshold (QT); namely, a criterion activity level that must be exceeded before a population's activity can reverberate in STM. Changing the QT or, equivalently, changing the size of signals to V_2, can retune the responsiveness of populations in V_2 to prescribed patterns at V_1. For example, suppose that an unexpected, or novel, event triggers a nonspecific arousal input to V_2, which magnifies all the signals from V_1 to V_2 (see Part II). Then certain signals, which could not otherwise be stored in STM, will exceed the QT and be stored. For example, if V_2 is capable of partial contrast in STM and also receives a nonspecific arousal input, then (8) can be replaced by

$$x_{2j}=\begin{cases} f(\phi S_j)\left[\sum_{\phi S_k>\varepsilon} f(\phi S_k)\right]^{-1} & \text{if} \quad \phi S_j>\varepsilon \\ 0 & \text{if} \quad \phi S_j<\varepsilon \end{cases} \tag{13}$$

where ϕ is an increasing function of the arousal level. Note that an increase in ϕ allows more V_2 populations to reverberate in STM; cf., Grossberg (1973) for mathematical proofs. In a similar fashion, if an unexpected event triggers nonspecific shunting inhibition of the

inhibitory interneurons in the off-surrounds of V_2, then the QT will decrease (Grossberg, 1973; Ellias and Grossberg, 1975), yielding an equivalent effect. Equation (8) can then be changed to

$$x_{2j}=\begin{cases} f(S_j)\left[\sum_{S_k>\phi^*\varepsilon} f(S_k)\right]^{-1} & \text{if } S_j>\phi^*\varepsilon \\ 0 & \text{if } S_j<\phi^*\varepsilon \end{cases} \tag{14}$$

where ϕ^* is a decreasing function of the arousal level.

Reductions in arousal level have the opposite effect. For example, if (13) holds, and arousal is lowered until only one population in V_2 exceeds the QT, then a choice will be made in STM, as in Figure 4b. Thus a choice in STM can be due either to *structural* properties of the network, such as the rules for generating signals between populations in V_2 [cf., the faster-than-linear signal function in Grossberg (1973)], or to an arousal level that is not high enough to create a tuning curve. Similarly, if arousal is too small, then all functions x_{2j} in (13) will always equal zero, and no STM storage will occur.

Changes in arousal can have a profound influence on the time course of LTM, as in (6), because they change the STM patterns that drive the learning process. For example, if during development arousal level is chosen to produce a choice in STM, then the tuning of classifying vectors $z^{(j)}$ will be sharper than if the arousal level were chosen to generate partial contrast in STM.

The influence of arousal on tuning of STM patterns can also be expressed in another way, which suggests a mechanism that will be needed in Part II when universal recoding is discussed.

7. Arousal as a Search Mechanism

Suppose that arousal level is fixed during learning trials, and that a given pattern Θ at V_1 does not create any STM storage at V_2 because all the inner products $\Theta \cdot z^{(j)}$ are too small. If arousal level is then increased in (13) until some $x_{2j}>0$, STM storage will occur. In other words, changing the arousal level can facilitate *search* for a suitable classifying population in V_2.

Why does arousal level increase if no STM storage occurs at V_2? This is a property of the expectation mechanism that is developed in Part II. Also in Part II a pattern Θ at V_1 that is not classified by V_2 will use this mechanism to release a subliminal search routine that terminates when an admissible classification occurs.

8. Development of an STM Code

System (6)–(7) will be analysed mathematically because it illustrates properties of the model in a particularly simple and lucid way. The first result describes how this system responds to a single pattern that is iteratively presented through time.

Theorem 1 (One Pattern)

Given a pattern Θ, suppose that there exists a unique j such that

$$S_j(0)>\max\{\varepsilon, S_k(0): k\neq j\}. \tag{15}$$

Let Θ be practiced during a sequence of nonoverlapping intervals $[U_k, V_k]$, $k=1, 2, \ldots$. Then the angle between $z^{(j)}(t)$ and Θ monotonically decreases, the signal $S_j(t)$ is monotonically attracted towards $\|\Theta\|^2$ and $\|z^{(j)}\|^2$ oscillates at most once as it pursues $S_j(t)$. In particular, if $\|z^{(j)}(0)\| \leqq \|\Theta\|$, then $S_j(t)$ is monotone increasing. Except in the trivial case that $S_j(0)=\|\Theta\|^2$, the limiting relations

$$\lim_{t\to\infty}\|z^{(j)}(t)\|^2=\lim_{t\to\infty} S_j(t)=\|\Theta\|^2 \tag{16}$$

hold if and only if

$$\sum_{k=1}^{\infty}(V_k-U_k)=\infty. \tag{17}$$

Remark. If $z^{(j)}(0)$ is small, in the sense that $\|z^{(j)}(0)\| \leqq \|\Theta\|$, then by Theorem 1, as time goes on, the learning process maximizes the inner product signal $S_j(t)=\Theta \cdot z^{(j)}(t)$ over all possible choices of $z^{(j)}$ such that $\|z^{(j)}\| \leqq \|\Theta\|$. This follows from the obvious fact that

$$\sup\{\Theta \cdot \psi: \|\psi\| \leqq \|\Theta\|\}=\|\Theta\|^2.$$

Otherwise expressed, learning makes $z^{(j)}$ parallel to Θ, and normalizes the length of $z^{(j)}$.

What happens if several different spatial patterns $\Theta^{(k)}=(\Theta_1^{(k)}, \Theta_2^{(k)}, \ldots, \Theta_n^{(k)})$, $k=1, 2, \ldots, M$, all perturb V_1 at different times? How are changes in the z_{ij}'s due to one pattern prevented from contradicting changes in the z_{ij}'s due to a different pattern? The choice-making property of V_2 does this for us; it acts as a sampling device that prevents contradictions from occurring. A heuristic argument will now be given to suggest how sampling works. This argument will then be refined and made rigorous. For definiteness, suppose that M spatial patterns $\Theta^{(k)}$ are chosen, $M \leqq N$, such that their signals at time $t=0$ satisfy

$$\Theta^{(k)} \cdot z^{(k)}(0)>\max\{\varepsilon, \Theta^{(k)} \cdot z^{(j)}(0): j\neq k\} \tag{18}$$

for all $k=1, 2, \ldots, M$. In other words, at time $t=0$, $\Theta^{(k)}$ is coded by v_{2k}. Let $\Theta^{(1)}$ be the first pattern to perturb V_1. By (18), population v_{21} receives the largest signal from V_1. All other populations v_{2j}, $j\neq 1$, are thereupon inhibited by the off-surround of v_{21}, whereas v_{21} reverberates in STM. By (6), none of the synaptic strengths $z^{(j)}(t)$, $j\neq 1$, can learn while $\Theta^{(1)}$ is presented. As in Theorem 1, presenting $\Theta^{(1)}$ makes $z^{(1)}(t)$ more parallel

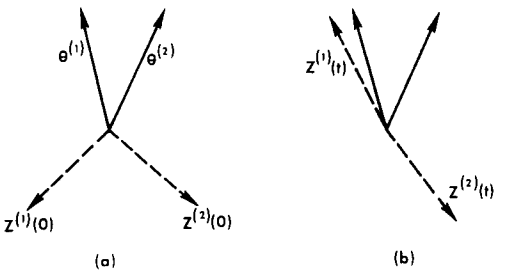

Fig. 7. Practicing $\Theta^{(1)}$ brings $z^{(1)}(t)$ closer to $\Theta^{(1)}$ and $\Theta^{(2)}$ than $z^{(2)}(0)$

to $\Theta^{(1)}$ as t increases. Consequently, if a different pattern, say $\Theta^{(2)}$, perturbs V_1 on the next learning trial, then it will excite v_{22} more than any other v_{2j}, $j \neq 2$: it cannot excite v_{21} because the coefficients $z^{(1)}(t)$ are more parallel to $\Theta^{(1)}$ than before; and it cannot excite any v_{2j}, $j \neq 1, 2$, because the v_{2j} coefficients $z^{(j)}(t)$ still equal $z^{(j)}(0)$. In response to $\Theta^{(2)}$, v_{22} inhibits all other v_{2j}, $j \neq 2$. Consequently none of the v_{2j} coefficients $z^{(j)}(t)$ can learn, $j \neq 2$: learning makes the coefficients $z^{(2)}(t)$ become more parallel to $\Theta^{(2)}$ as t increases. The same occurs on all learning trials. By inhibiting the post-synaptic part of the learning mechanism in all but the chosen V_2 population, the on-center off-surround network in V_2 samples one vector $z^{(j)}(t)$ of trainable coefficients at any time. In this way, V_2 can learn to distinguish as many as N patterns if it contains N populations.

This argument is almost correct. It fails, in general, because by making (say) $z^{(1)}(t)$ more parallel to $\Theta^{(1)}$, it is also possible to make $z^{(1)}(t)$ more parallel to $\Theta^{(2)}$ than $z^{(2)}(0)$ is. Thus when $\Theta^{(2)}$ is presented, it will be coded by v_{21} rather than v_{22}. In other words, practicing one pattern can recode other patterns. A typical example of this property is illustrated in Figure 7. Figure 7a depicts the two dimensional patterns $\Theta^{(1)}$ and $\Theta^{(2)}$ as solid vectors, and the two classifying vectors $z^{(1)}(0)$ and $z^{(2)}(0)$ as dotted vectors. Clearly (18) holds for $j = 1, 2$. As a result of practicing $\Theta^{(1)}$ during a fixed interval, Figure 7b is produced. Note that $\Theta^{(2)} \cdot z^{(1)}(t) > \Theta^{(2)} \cdot z^{(2)}(t)$ after the practice interval terminates. Consequently, v_{21}, rather than v_{22}, codes $\Theta^{(2)}$ when $\Theta^{(2)}$ is practiced. This property can be iterated to show how systematic trends in the sequence of practiced patterns can produce systematic drifts in recoding. Consider Figure 8. Again two dimensional patterns are denoted by solid vectors and classifying vectors are denoted by dotted vectors. Let the patterns be practiced in the order $\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(M)}$, where $M \gg N$. By successively practicing $\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(r-1)}$, the vector $z^{(1)}(t)$ is dragged along clockwise until it almost reaches $\Theta^{(r-1)}$. Then $\Theta^{(r)}$ is practiced, and since $\Theta^{(r)}$ is coded by v_{22}, $z^{(1)}(t)$ stops moving and $z^{(2)}(t)$ begins to move clockwise; $z^{(2)}(t)$ continues to move clockwise while $\Theta^{(r+1)}, \Theta^{(r+2)}, \ldots, \Theta^{(2r-1)}$ are practiced. Then $z^{(3)}(t)$ begins to move clockwise, and so on. The clockwise drift in the practice schedule hereby shifts each $z^{(j)}(t)$, $j = 1, 2, \ldots, M-1$, to a position that is close to the one $z^{(j-1)}(0)$ occupied. In other words, essentially *all* vectors in V_2 are reclassified. If the same practice schedule $\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(M)}$ is repeated on a second learning trial, then essentially all v_{2i} are recoded by $v_{2,i+2}$, and so on. Each learning trial recodes V_2 until all the N populations in V_2 code one of the N most clockwise vectors $\Theta^{(k)}$. This asymptotic coding of V_2 is stable, except for a wild oscillation in the coding of v_{21} on each learning trial, if the same practice schedule is always repeated. If, however, a counter-clockwise drift in practiced patterns is then imposed, all of V_2 will be recoded until the N most counter-clockwise vectors $\Theta^{(k)}$ are coded. In general, if there are many patterns relative to the number of populations in V_2, and if the statistical structure of the practice sequences continually changes, then there need not exist a stable coding rule in V_2. This is quite unsatisfactory.

By contrast, if there are few, or sparse, patterns relative to the number of populations in V_2, then a stable coding rule does exist, and the STM choice rule in V_2 does provide an effective sampling technique. Such a situation is approximated, for example, when the network is exposed to a "visually deprived" environment, in imitation of experiments on young animals. A theorem concerning this case will now be stated, if only to suggest what auxiliary mechanisms will be needed to establish a stable coding rule in the general case. This theorem shows how populations learn to code convex regions of features. In particular, if v_{2j} learns to code a certain set of features, then it automatically codes *average* features derived from this set.

The following nomenclature will be needed to state the theorem. A *partition* $\bigoplus_{k=1}^{K} \mathscr{P}_k$ of a finite set $\mathscr{P}$ is a subdivision of $\mathscr{P}$ into nonoverlapping and exhaustive subsets $\mathscr{P}_j$. The *convex hull* $\mathscr{H}(\mathscr{P})$ of a finite set $\mathscr{P}$ is the set of all convex combinations of elements in $\mathscr{P}$: for example, if $\mathscr{P} = \{\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(M)}\}$, then

$$\mathscr{H}(\mathscr{P}) = \left\{ \sum_{k=1}^{M} \lambda_k \Theta^{(k)} : \text{each } \lambda_k \geqq 0 \text{ and } \sum_{k=1}^{M} \lambda_k = 1 \right\}.$$

Given a set $\mathscr{P}$ with subset $\mathscr{Q}$, let $\mathscr{R} = \mathscr{P} \setminus \mathscr{Q}$ denote the set of elements in $\mathscr{P}$ that are not in $\mathscr{Q}$. If the classifying vector $z^{(j)}(t)$ codes the set of patterns $\mathscr{P}_j(t)$, let $\mathscr{P}_j^*(t) = \mathscr{P}_j(t) \cup \{z^{(j)}(t)\}$. The *distance* between a vector P and a set of vectors $\mathscr{Q}$, denoted by $\|P - \mathscr{Q}\|$, is defined by

$$\|P - \mathscr{Q}\| = \inf\{\|P - Q\| : Q \in \mathscr{Q}\}.$$

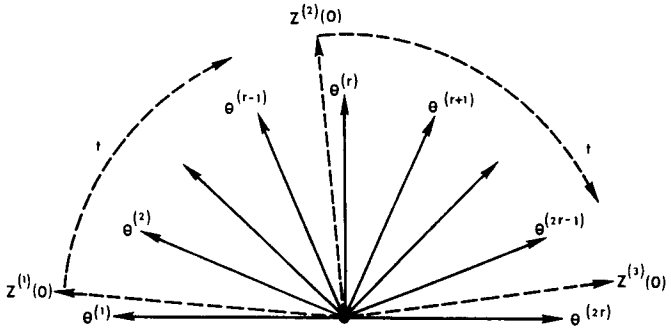

Fig. 8. Practicing a sequence of spatial patterns can recode all the populations

Theorem 2 (Sparse Patterns)

Let the network practice any set $\mathscr{P} = \{\Theta^{(i)}: i = 1, 2, \ldots, M\}$ of patterns for which there exists a partition $\mathscr{P} = \bigoplus_{k=1}^{N} \mathscr{P}_k(0)$ such that

$$\min\{u \cdot v : u \in \mathscr{P}_j(0), v \in \mathscr{P}_j^*(0)\} > \max\{u \cdot v : u \in \mathscr{P}_j(0), v \in \mathscr{P}^*(0) \backslash \mathscr{P}_j^*(0)\} \tag{19}$$

for all $j = 1, 2, \ldots, N$. Then $\mathscr{P}_j(t) = \mathscr{P}_j(0)$ and the functions

$$D_j(t) = \|z^{(j)}(t) - \mathscr{K}(\mathscr{P}^{(j)}(t))\| \tag{20}$$

are monotone decreasing for $t \geqq 0$ and $j = 1, 2, \ldots, N$. If moreover the patterns in $\mathscr{P}^{(j)}(0)$ are practiced in intervals $[U_{jm}, V_{jm}]$, $m = 1, 2, \ldots$, such that

$$\sum_{m=1}^{\infty} (V_{jm} - U_{jm}) = \infty \tag{21}$$

then

$$\lim_{t \to \infty} D_j(t) = 0. \tag{22}$$

Remarks. In other words, if the classifying vectors initially code the patterns into sparse classes, in the sense of (19), then this code persists through time, and the classifying vectors approach a convex combination of their coded patterns. As (20) and (22) show, learning permits each v_{2j} to respond as vigorously as possible to its class of coded patterns.

The above results indicate that, given a fixed number of patterns, it becomes easier to establish a stable code for them as the number of populations in V_2 increases. Once V_2 is constructed, however, it is not possible to increase its number of populations at will. Moreover, *in vivo*, an enormous variety of patterns typically barrages the visual system. How can a stable code be guaranteed no matter how many patterns perturb V_1?

One way is to assume that a biochemically determined *critical period* exists during which the z_{ij}'s are capable of learning; once the critical period terminates, some chemical factor is removed and the z_{ij}'s remain fixed in the last code to be established. The existence of a critical period has been reported (Hubel and Wiesel, 1970), but whether it is due to a chemical factor, or *merely* to a chemical factor, is as yet unknown. From a formal point of view, such a mechanism suffers from several significant related disadvantages. The most obvious one is that all the coded information that is learned throughout the critical period can be obliterated if its last phase exhibits an unlikely statistical trend. In addition, a repetitive statistical trend can prevent many patterns from being coded at all. For example, in Figure 8, once the classifying vectors code the N most clockwise patterns, many of the other $M-N$ patterns might be too far away from $z^{(1)}$ to satisfy the ε-condition in (7); they will then not be coded by any population. Yet each of these $M-N$ patterns has been presented as frequently as the N patterns that are coded. More generally, because populations which are already coded can be recoded so easily, it is hard to search for as yet uncommitted populations to code as yet uncoded patterns. This problem prevents a universal recoding from being achieved (see Part II).

These negative remarks can be supplemented by intriguing positive observations. Stabilizing the code seems to require the same formal machinery that is needed in models of adult attention and discrimination learning (Grossberg, 1975). This machinery, in turn, is highly evokative of data concerning attentional modulation of olfactory patterns by the prepyriform cortex of cats (Freeman, 1974). Auxiliary mechanisms for stabilizing the code will therefore be motivated below. It is understood that a biochemically triggered critical period can coexist with these mechanisms, or indeed can preempt them in sufficiently primitive organisms.

Various mechanisms can be contemplated which partially stabilize the code, but which are not sufficient. A satiation mechanism will be sketched below to

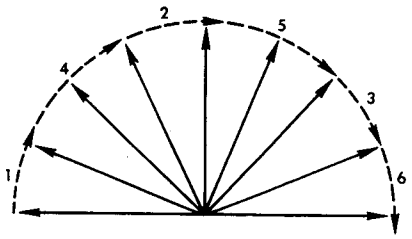

Fig. 9. Practicing in the order 1, 2, 3, 4, 5, 6 can recode all the populations even if satiation exists

clarify what is needed. Consider (6) with

$$x_{2j}(t)=\begin{cases}G_j(t) & \text{if } S_j(t)G_j(t)>\max\{\varepsilon, S_k(t)G_k(t):k\neq j\}\\ 0 & \text{if } S_j(t)G_j(t)<\max\{\varepsilon, S_k(t)G_k(t):k\neq j\}\end{cases} \tag{23}$$

where

$$G_j(t)=g\left(1-\int_0^t x_{2j}(v)K(t-v)dv\right). \tag{24}$$

In (24), $g(w)$ is a monotone increasing function such that $g(0)=0$ and $g(1)=1$. $K(w)$ is a monotone decreasing function such that $K(0)=1$ and $K(\infty)=0$; for example, $K(w)=e^{-w}$. Equation (23) says that persistent activation of v_{2j} causes its STM response to satiate, or adapt; if v_{2j} is active during a sufficiently long interval, its activity approaches zero. Correspondingly, $z^{(j)}$'s fluctuations are damped within a time interval of fixed length. Such a mechanism is inadequate if the training schedule allows v_{2j} to recover its maximal strength. Figure 9 shows, for example, an ordering of patterns that permits recoding of essentially all populations in V_2.

This problem is only made worse by replacing the choice rule in (23) by a partial contrast rule such as

$$x_{2j}=\begin{cases}\dfrac{f(S_jG_j)}{\sum\limits_{S_kG_k>e} f(S_kG_k)} & \text{if } S_jG_j>\varepsilon\\ 0 & \text{if } S_jG_j<\varepsilon.\end{cases}$$

Here, if a prescribed pattern Θ causes a maximal STM response at v_{2j}, then the activity x_{2j} is suppressed by G_j more rapidly than the activities of other Θ-activated populations. There can consequently be a shift in the locus of maximal responsiveness even to a single pattern—that is, recoding—in addition to the difficulty cited in Figure 9.

Such examples clarify what is essential:

(A) *Before* $z^{(j)}(t)$ learns a pattern, or class of related patterns, it must be able to fluctuate freely in response to pattern inputs in search of a classification.

(B) *After* $z^{(j)}(t)$ learns a pattern, it must be prevented from coding very different patterns, no matter what the training schedule is. In particular, satiating $z^{(j)}$'s ability to change through time does not suffice, since a very different pattern can still be coded by $z^{(j)}$ if this pattern elicits a larger signal at v_{2j}, say due to the size of $\|z^{(j)}\|$ rather than the direction of vector $z^{(j)}$, than at any of the uncommited populations.

Requirements (A) and (B) constrain the interaction of STM and LTM mechanisms, given that (6) holds. For example, by (6), if a pattern Θ creates signals while v_{2j} is active in STM, then $z^{(j)}(t)$ will change. Suppose that a sequence $\Theta^{(1)}$, $\Theta^{(2)}$ of two very different patterns is successively presented to V_1, and that $z^{(1)}(t)$ codes $\Theta^{(1)}$. In response to $\Theta^{(1)}$, v_{21} is activated, but $z^{(1)}(t)$ does not substantially change because it already codes $\Theta^{(1)}$. Now let $\Theta^{(2)}$ perturb V_1. By requirement (B), $z^{(1)}(t)$ must not be allowed to change. By (6), $z^{(1)}(t)$ will change unless either no signal is emitted from V_1 when v_{21} is active, or a signal is emitted from V_1 only after v_{21} is inactivated. These two cases will be separately considered in the next two paragraphs.

In the former case, some type of feedback to V_1 must suppress the V_1-to-V_2 signals that would otherwise be generated by $\Theta^{(2)}$. This feedback somehow tells V_1 that $\Theta^{(2)}$ is very different from the pattern $\Theta^{(1)}$ that is presently coded in STM. By (A), however, $\Theta^{(2)}$ can generate V_1-to-V_2 signals at *some* time, either to search for a classifying vector, or to activate its already learned STM representation. Thus after V_1-to-V_2 signals are suppressed long enough for STM activity in v_{21} to also be suppressed, then V_1-to-V_2 signals are reactivated.

In the latter case, changing $\Theta^{(1)}$ to $\Theta^{(2)}$ somehow suppresses the STM activity that codes $\Theta^{(1)}$; in particular, somehow the network can tell when the spatial patterns that perturb V_1 are changed. In both cases, the same general issue is raised: how does the network process a temporal succession $\Theta^{(1)}$, $\Theta^{(2)}$, ..., $\Theta^{(k)}$, ... of spatial patterns $\Theta^{(k)}=(\Theta_1^{(k)}, \Theta_2^{(k)}, \ldots, \Theta_n^{(k)})$; that is, a *space-time pattern*. Space-time patterns are the typical inputs to a receptive field *in vivo*. The problem of stabilizing the STM code forces us to consider their processing in some detail. Part II of this paper considers this problem.

Appendix

Proof of Theorem 1. Consider the case in which

$$|\Theta|^2>S_j(0)>\max\{\varepsilon, S_k(0):k\neq j\}. \tag{A1}$$

The case in which $S_j(0)\geqq|\Theta|^2$ can be treated similarly. First it will be shown that if the inequalities

$$|\Theta|^2>S_j(t)>\max\{\varepsilon, S_k(t):k\neq j\} \tag{A2}$$

hold at any time $t=T\in\bigcup_{m=1}^{r}[U_m, V_m]$, then they hold at all times $t\in[T, \infty)\cap\bigcup_{m=1}^{r}[U_m, V_m]$. By (A2), $x_{2j}(T)=1$ and $x_{2k}(T)=0$, $k\neq j$.

Consequently, by (6),

$$\dot{z}_{ij}(T) = -z_{ij}(T) + \Theta_i \tag{A3}$$

and

$$\dot{z}_{ik}(T) = 0 \tag{A4}$$

for $k \neq j$ and $i = 1, 2, \ldots, n$. By (A2)–(A4),

$$\dot{S}_j(T) = -S_j(T) + |\Theta|^2 > 0 = \dot{S}_k(T), \tag{A5}$$

$k \neq j$. Thus (A2) holds for all $t \in [T, \infty) \cap \bigcup_{m=1}^{\infty} [U_m, V_m]$. By (A2) and (A5), for all $t \in \bigcup_{m=1}^{\infty} [U_m, V_m]$, $S_j(t)$ increases monotonically towards $|\Theta|^2$ and (16) holds if and only if (17) holds. For $t \notin \bigcup_{m=1}^{\infty} [U_m, V_m]$, all $\dot{S}_k(t) = 0$, $k = 1, 2, \ldots, n$.

Letting $N_j = |z^{(j)}|^2$ and $C_j = \cos(z^{(j)}, \Theta) \equiv S_j N_j^{-1/2} |\Theta|^{-1}$, it readily follows from (A5) that for all $t \in \bigcup_{m=1}^{\infty} [U_m, V_m]$,

$$\dot{N}_j = 2(-N_j + S_j) \tag{A6}$$

and

$$\dot{C}_j = |\Theta| N_j^{-1/2} (1 - C_j^2). \tag{A7}$$

Equation (A7) shows that the angle between $z^{(j)}(t)$ and Θ closes monotonically as Θ is practiced. Since $S_j(t)$ is a monotonic function, (A6) shows that $N_j(t)$ oscillates at most once.

In particular, suppose $\|z^{(j)}(0)\| \leqq \|\Theta\|$. Then $S_j(0) \leqq \|\Theta\|^2$, since otherwise

$$\Theta \cdot z^{(j)}(0) > \Theta \cdot \Theta \geqq z^{(j)}(0) \cdot z^{(j)}(0)$$

which implies

$$1 \geqq C_j(0) > \|\Theta\| \, \|z^{(j)}(0)\|^{-1} \geqq \|z^{(j)}(0)\| \, \|\Theta\|^{-1},$$

and thus

$$\|z^{(j)}(0)\| > \|\Theta\| > \|z^{(j)}(0)\|,$$

which is a contradiction. By (A5), therefore $\|z^{(j)}(0)\| \leqq \|\Theta\|$ implies that $S_j(t)$ is monotone increasing

Proof of Theorem 2. Inequality (19) is based on the fact that, if a fixed set of patterns $\Theta^{(j_1)}, \Theta^{(j_2)}, \ldots, \Theta^{(j_k)}$ is classified by $z^{(j)}(t)$ for all $t \geqq 0$, then

$$z^{(j)}(t) \in \mathscr{H}(\Theta^{(j_1)}, \Theta^{(j_2)}, \ldots, \Theta^{(j_k)}, z^{(j)}(0)), \tag{A8}$$

for all $t \geqq 0$. For example, suppose that the patterns are practiced in the order $\Theta^{(j_1)}, \Theta^{(j_2)}, \ldots, \Theta^{(j_k)}$ during the nonoverlapping intervals $[U_1, V_1], [U_2, V_2], \ldots, [U_k, V_k]$. Except during these intervals, $z^{(j)} = 0$. Thus for $t \in [U_1, V_1]$,

$$\dot{z}^{(j)} = -z^{(j)} + \Theta^{(j_1)},$$

or

$$z^{(j)}(t) = z^{(j)}(0) e^{-(t - U_1)} + \Theta^{(j_1)}(1 - e^{-(t - U_1)}),$$

so that

$$z^{(j)}(t) \in \mathscr{H}(\Theta^{(j_1)}, z^{(j)}(0)) \subset \mathscr{H}(\Theta^{(j_1)}, \ldots, \Theta^{(j_k)}, z^{(j)}(0)).$$

For $t \in [U_2, V_2]$,

$$z^{(j)}(t) = [z^{(j)}(0) e^{-(V_1 - U_1)} + \Theta^{(j_1)}(1 - e^{-(V_1 - U_1)})] e^{-(t - U_2)} + \Theta^{(j_2)}(1 - e^{-(t - U_2)}). \tag{A9}$$

Hence

$$z^{(j)}(t) \in \mathscr{H}(\Theta^{(j_1)}, \Theta^{(j_2)}, z^{(j)}(0)) \subset \mathscr{H}(\Theta^{(j_1)}, \ldots, \Theta^{(j_k)}, z^{(j)}(0)),$$

and so on.

Condition (19) is then applied using the fact that, for any $U \in P_j(0)$, $V \in \mathscr{H}(P_j^*(0))$, and $W \in \mathscr{H}(P^*(0) \backslash P_j^*(0))$,

$$U \cdot V > \max\{\varepsilon, U \cdot W\} \tag{A10}$$

because

$$U \cdot V \geqq \min\{u \cdot v : u \in P_j(0), v \in P_j^*(0)\}$$

and

$$\max\{u \cdot v : u \in P_j(0), v \in P^*(0) \backslash P_j^*(0)\} \geqq U \cdot W.$$

Until a pattern is reclassified, however, (A8) shows that $z^{(j)}(t) \in \mathscr{H}(P_j^*(0))$ and that $z^{(k)}(t) \in \mathscr{H}(P^*(0) \backslash P_j^*(0))$ for any $k \neq j$. But then, by (A10), reclassification is impossible.

That $D_j(t)$ in (20) is monotone decreasing follows from iterations of (A9). That (21) implies (22) follows just as in the proof of Theorem 1.

References

Barlow, H. B., Pettigrew, J. D.: Lack of specificity of neurones in the visual cortex of young kittens. J. Physiol. (Lond.) **218**, 98–100 (1971)

Bennett, M. V. L.: Analysis of parallel excitatory and inhibitory synaptic channels. J. Neurophysiol. **34**, 69–75 (1971)

Blackenship, J. E., Wachtel, H., Kandel, E. R.: Ionic mechanisms of excitatory, inhibitory, and dual synaptic actions mediated by an identified interneuron in abdominal ganglion of Aplysia. J. Neurophysiol. **34**, 76–92 (1971)

Blakemore, C., Cooper, G. F.: Development of the brain depends on the visual environment. Nature (Lond.) **228**, 477–478 (1970)

Blakemore, C., Mitchell, D. E.: Environmental modification of the visual cortex and the neural basis of learning and memory. Nature (Lond.) New Biol. **241**, 467–468 (1973)

Boycott, B. B., Dowling, J. E.: Organization of the primate retina: light microscopy. Phil. Trans. roy. Soc. B. **255**, 109–184 (1969)

Ellias, S. A., Grossberg, S.: Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks. Biol. Cybernetics **20**, 69–98 (1975)

Freeman, W. J.: Neural coding through mass action in the olfactory system. Proceeding IEEE Conference on biologically motivated automata theory 1974

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

Greenspan, H. P., Benney, D. J.: Calculus. New York: McGraw-Hill 1973

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

Grossberg, S.: Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. Stud. appl. Math. **49**, 135–166 (1970a)

Grossberg, S.: Neural pattern discrimination. J. theor. Biol. **27**, 291–337 (1970b)

Grossberg, S.: Pavlovian pattern learning by nonlinear neural networks. Proc. nat. Acad. Sci. (Wash.) **68**, 828–831 (1971)

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

Grossberg, S.: Contour enhancement, short term memory, and constancies in reverberating neural networks. Stud. appl. Math. **52**, 213–257 (1973)

Grossberg, S.: Classical and instrumental learning by neural networks. In: Rosen, R. and Snell, F. (Eds.): Progress in Theoretical Biology, pp. 51–141. New York: Academic Press 1974

Grossberg, S.: A neural model of attention, reinforcement, and discrimination learning. Int. Rev. Neurobiol. **18**, 263–327 (1975)

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

Grossberg, S., Levine, D.S.: Some developmental and attentional biases in the contrast enhancement and short term memory of recurrent neural networks. J. theor. Biol. **53**, 341–380 (1975)

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

Hirsch, H.V.B., Spinelli, D.N.: Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. Science **168**, 869–871 (1970)

Hirsch, H.V.B., Spinelli, D.N.: Modification of the distribution of receptive field orientation in cats by selective visual exposure during development. Exp. Brain Res. **12**, 509–527 (1971)

Hubel, D.H., Wiesel, T.N.: The period of susceptibility to the physiological effects of unilateral eye closure in kittens. J. Physiol. (Lond.) **206**, 419–436 (1970)

Kimble, G.A.: Foundations of conditioning and learning. New York: Appleton-Century-Crofts 1967

Levine, D.S., Grossberg, S.: Visual illusions in neural networks: line neutralization, tilt aftereffect, and angle expansion. J. theor. Biol., in press (1976)

Meinhardt, H., Gierer, A.: Applications of a theory of biological pattern formation based on lateral inhibition. J. Cell. Sci. **15**, 321–346

Pérez, R., Glass, L., Shlaer, R.: Development of specificity in the cat visual cortex. J. math. Biol. (1974)

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

Wachtel, H., Kandel, E.R.: Conversion of synaptic excitation to inhibition at a dual chemical synapse. J. Neurophysiol. **34**, 56–00 (1971)

Wiesel, T.N., Hubel, D.H.: Single-cell responses in striate cortex of kittens deprived of vision in one eye. J. Neurophysiol. **26**, 1003–1017 (1963)

Wiesel, T.N., Hubel, D.H.: Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. J. Neurophysiol. **28**, 1029–1040 (1965)