Self-Organization in a Perceptual Network

Ralph Linsker

IBM Research

A young animal or child perceives and identifies features in its environment in an apparently effortless way. No presently known algorithms even approach this flexible, general-purpose perceptual capability. Discovering the principles that may underlie perceptual processing is important both for neuroscience and for the development of synthetic perceptual systems.

Two important aspects of the mystery of perception are

(1) What processing functions does the neural "machinery" perform on perceptual input, and what is the circuitry that implements these functions?

(2) How does this "machinery" come to be?

Unlike conventional computer hardware, neural circuitry is not hard-wired or specified as an explicit set of point-to-point connections. Instead it develops under the influence of a genetic specification and epigenetic factors, such as electrical activity, both before and after birth. How this happens is in large part unknown.

Biological development processes are far too complex to hope that a relatively complete understanding of how a perceptual system develops and functions will soon emerge. But we are familiar with complex synthetic systems, such as computers, whose principles of organization can be understood without one's knowing in detail how the components work. Furthermore, the same principles can be used to build computers in any of several different technologies. Might there be organizing principles

(1) that explain some essential aspects of how a perceptual system develops and functions;

(2) that we can attempt to infer without waiting for far more detailed experimental information; and

(3) that can lead to profitable experimental programs, testable predictions, and applications to synthetic perception as well as neuroscientific understanding?

How can a perceptual system develop to recognize specific features of its environment, without being told which features it should analyze, or even whether its identifications are correct?

I believe the answer is yes, and that the use of theoretical neural networks that embody biologically-motivated rules and constraints is a powerful tool in this study.

This optimism is encouraged by recent work[1] in which I have found that a multilayered network, developing according to simple yet biologically plausible "Hebb-type" rules,[2] self-organizes to produce feature-analyzing "cells." These "cells" have response properties that are qualitatively similar to those cells of the first few processing stages of the mammalian visual system.[3] These properties include sensitivity to light-dark contrast and sensitivity to the orientation of an edge or bar. These properties develop before birth in certain animals, hence before structured visual experience, and in the theoretical network the corresponding properties develop even in the absence of structured input, using only random signaling activity in the input layer of the network.

Why does a feature-analyzing function emerge from these development rules? Is it a mere accident or curiosity? Or are the development rules perhaps acting to optimize some quantity that is important to the information processing function of a perceptual system?

0018-9162/88/0300-00105$01.00 © 1988 IEEE

In this article, I briefly summarize the network ideas from an earlier publication[1] and review some of the main results. This sets the stage for exploring why a feature-analyzing function emerges. I then show that even a single developing cell of a layered network exhibits a remarkable set of optimization properties. These properties are closely related to issues in statistics, theoretical physics, adaptive signal processing, the formation of knowledge representations in artificial intelligence, and information theory.

Next, I use these results to infer an information-theoretic principle that can be applied to the network as a whole, rather than a single cell. The organizing principle I propose is that the network connections develop in such a way as to maximize the amount of information that is preserved when signals are transformed at each processing stage, subject to certain constraints.

I illustrate how this principle works for some very simple cases. Much more work will be needed to apply the principle to practical computations of biologically important cases, but the approach appears very promising. I conclude with some speculative comments on why this principle, or some variant of it, may be important for the emergence of perceptual function in biological and synthetic systems.

A layered self-adaptive network

The visual system is the best studied perceptual system in mammals. Visual information is processed in stages. Simple aspects of form, such as contrast and edge orientation, are analyzed in the earlier stages; more complex features are analyzed later. Other aspects of visual processing, such as color and motion analysis, proceed in parallel with the analysis of form.

Both the retina and cortex are organized into layers of cells with interconnections within and between layers. Within an anatomical layer, at least for the early processing stages, there is a population of cells each of which performs approximately the same processing function on its inputs. This population of cells can be thought of as an array of filters. Each cell processes input from a limited region of visual space, called the "receptive field" of that cell. More than one population of cells can share an anatomical layer.

Many cells respond to input activity by firing an electrical pulse, or *action potential*, that travels down the output fiber, or axon. These pulses cause a chemical neurotransmitter substance to be released at synapses, or regions of near-contact with other cells. The latter cells receive and process these chemical input signals. Some cells, for example in the retina, do not produce action potentials, but instead exhibit more graded electrochemical phenomena that can be used for signaling.

Although a cell's response function is in general nonlinear, visual neurophysiologists have found that for many cells, a linear summation approximation is appropriate. In this approximation, the cell's output response varies monotonically with some linear combination of the cell's input signal values. For cells that produce action potentials, the output response can be defined as the firing rate at which the cell generates action potential pulses in response to its input signals.

Specification of the network. Will a simple self-adaptive network develop feature-analyzing cells without our specifying which features are to be analyzed? If it does, are these cell types related to those observed in biological systems? To address these questions, we first study a network that embodies some of the important biological properties described above, but omits many complicating factors. This approach is useful both because many of the details are unknown, and because our goal is to understand what principles are most important for the development of perceptual functions. For example, if we want to know how nonlinearity of response may be important for development, it is valuable to see first whether a linear response system exhibits the main feature-analyzing properties that are biologically observed. Also, feedback connections from later to earlier processing stages are known to exist, but it is not known how these connections might relate to the development of feature-analyzing functions. (There are many other functions that feedback may serve, such as control of dynamic range, attentional mechanisms, and so on.) We choose to analyze networks without feedback, to understand their developmental properties first.

The interconnections within the retina are known to be more complicated than a simple feedforward arrangement. Also, mechanisms that are not dependent on neural activity appear to be involved in the development of some feature-analyzing properties. The main purpose of our simulations is to explore what types of simple yet biologically plausible development rules *suffice* to generate feature-analyzing cell assemblies, rather than to rule out other ways of generating them. From the results of our simple model, we will infer a potential organizing principle that can encompass nonlinear cell response, more complex connectivity, and a variety of ways of forming and modifying connections.

Our network is shown in Figure 1. The cells are organized into two-dimensional layers A, B, C, and so on, with feedforward connections to each cell from an overlying neighborhood of cells of the previous layer. Layer A receives input from the visual world (if there is any such input). We focus especially on the case in which there is no input, but instead only random activity of the cells of layer A, with no correlation of activity from one cell to the next. This activity resembles random noise or snow on a TV screen. We consider this case in order to understand how certain feature-analyzing cells may emerge even before birth, as has been observed in certain primates.

The positions of the connections to each cell need not be regular as in Figure 1, but can be chosen randomly according to a density distribution, such as a Gaussian, that favors connections from nearby cells of the previous layer. For simplicity, these positions are fixed for the duration of the development process. Each cell, at each time, has some signaling activity which we denote by a real number. Each cell exhibits a simple linear response, that is, the output is a linear combination of the inputs, with each input being weighted by a *connection strength* that will develop in a certain way. Each model cell thus acts as a linear filter.

Two points should be noted:

(1) Defining the output response as a nonlinear, for example sigmoid, function of the weighted sum of the inputs would more closely approximate some properties of the firing rates of biological neurons. These are always nonnegative and saturate at some maximum rate. However, we will see that even a linear response rule can lead to the formation of feature-analyzing cells, and we will explore what properties of linear adaptive filters are responsible for this formation. Some of the insights gained will be applicable to the nonlinear response case as well.

(2) Any transformation implemented by a feedforward sequence of layers of lin-

ear filters is a linear transformation, and hence could be implemented by a single layer of connections with properly chosen, or in this case hardwired, connection strengths. However, our purpose is not to implement a particular transformation, but rather to study what transformations are learned by a network without supervision. This multistage learning process depends upon the presence of multiple network layers.

A Hebb rule. For the development process, we use a version of an idea proposed by the neuropsychologist Donald Hebb in 1949. This idea has been central to much work on synthetic neural networks over the years, as well as to the thinking of neuroscientists about how the development of synaptic connections may relate to memory and learning phenomena. Hebb's idea was that if cell 1 is one of the cells providing input to cell 2, and if cell 1's activity tends to be "high" whenever cell 2's activity is "high", then the future contribution that the firing of cell 1 makes to the firing of cell 2 should increase.

In the language of neural networks, the connection strength is increased, or made more positive. A mathematical formulation needs to be more precise than this, and state under what conditions the strength may decrease. We use a form in which the change in strength contains a term proportional to the product of input and output activities at that connection. The Hebbian idea of modifying connection strengths according to the degree of correlated activity between input and output is central to what follows.

For an analogy to a Hebbian rule, consider a group of people whose collective opinion on a question is by definition the weighted average of the opinions of its members. If, over time, a member's opinion tends to agree with the group's opinion, then the analog of the Hebb rule states that the individual member's vote on future issues is to be weighted more strongly. The member's vote is given less weight, or even negative weight, if he consistently disagrees with the group's opinion. This type of positive-feedback control of weighting factors tends to lead to consensus within the group. As we shall see, it has other surprising consequences for the properties of the group, or output cell, response.

Mathematical formulation. This subsection and the next summarize simulations that are described in detail in my previous work.[1]

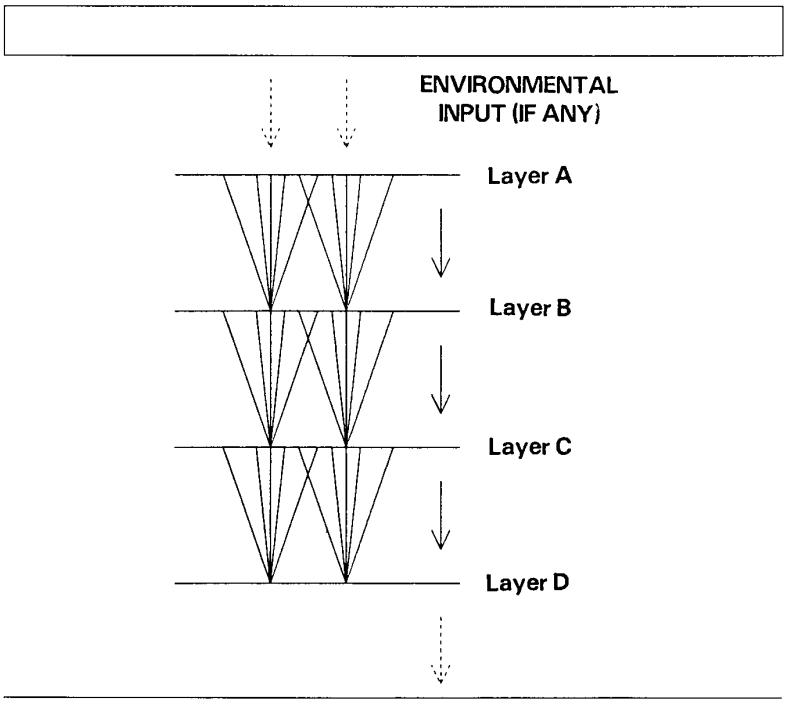

Figure 1. A layered self-adaptive network with local feedforward connections. Each two-dimensional layer contains many cells. Five input connections to each of two cells in layers B, C, and D are shown. Several hundred inputs to each cell are used in simulations. Each cell also provides input to many cells of the following layer. Lateral connections within a layer, as discussed in the text, are not indicated here.

Consider a cell M and the cells $L_1, L_2, \ldots, L_N$ that provide input to M. For simplicity, we avoid treating effects that depend upon the time sequence of signal activity values. Instead, we think of the activity history of a layer as a set of "snapshots," in which the ordering of the snapshots plays no role. That is, a set of activity values, denoted by $(L_1^\pi, L_2^\pi, \ldots, L_N^\pi)$, is presented as input to the M cell, the M cell generates an output activity value M^π, and a new set of input activities is then presented. The superscript π indexes the presentation of inputs, that is, the particular snapshot, and the corresponding output. Then the linear response rule is

$$M^\pi = a_1 + \Sigma_j L_j^\pi c_j \qquad (1)$$

where c_j is the strength of the jth input connection to the M cell. Our Hebb-type rule is

$$(\Delta c_i)^\pi = a_2 L_i^\pi M^\pi + a_3 L_i^\pi + a_4 M^\pi + a_5 \qquad (2)$$

where the a's are arbitrary constants ($a_2 > 0$). We assume that the c values change slowly from one presentation to the next. Then we can average Equation 2 over an ensemble of many presentations, and use Equation 1 to express M^π in terms of the $\{L_j^\pi\}$ to obtain the rate of change of each c value. Some algebraic manipulation[1] gives

$$\dot{c}_i = \Sigma_j Q_{ij} c_j + [k_1 + (k_2/N)\Sigma c_j] \qquad (3)$$

where $k_{1,2}$ are particular combinations of the constants $a_{1\text{-}5}$. Apart from the determined values of $k_{1,2}$, the constants $a_{1\text{-}5}$ play no further role in what follows. Here

$$Q_{ij} \equiv <(L_i^\pi - \bar{L}) \times (L_j^\pi - \bar{L})> \qquad (4)$$

is the covariance of the activities of input cells i and j, where $< \ldots >$ and the overbar both denote the ensemble average. (For our purposes, $\bar{L}$, the ensemble average of the input activity at a synapse, can

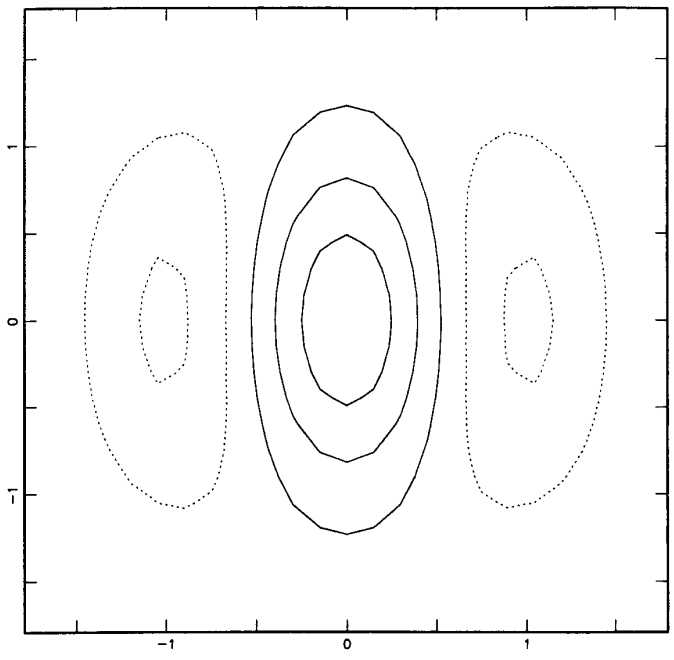

Figure 2. Receptive field map of a computed orientation-selective cell. A point of illumination at any position in the plane evokes an output response from the model cell that is proportional to the contour value at that position. Positive contour values (solid curves) denote an excitatory output response; negative values (dotted curves) denote an inhibitory response. Contour values range from -0.45 to $+0.75$ in steps of 0.30. The peak response (at the receptive field center) is normalized to unity. The parameter values that generated this particular orientation-selective cell, and the units (r_G) of distance along the axes, are given in reference 1. (See Figure 1a, p. 8780). Axes denote distance of illumination point from receptive field center.

be taken to be the same for all synapses i, j.) The appearance of the input covariance matrix Q does not mean that there is any direct interaction between synapses i and j. Q appears simply because the Hebb rule causes $\dot{c}_i$ to depend upon the product $<L_i^n M^n>$, and M^n in turn depends upon all the $\{L_j^n\}$ values (via Equation 1). The Q matrix will play an important role in what follows.

To prevent c values from becoming infinite during the development process, a saturation constraint is imposed. Each c value is constrained to lie between two values c_- and c_+. In a more biologically realistic case, there are excitatory synapses that have $0 \leq c \leq c_+$ and inhibitory synapses that have $c_- \leq c \leq 0$. The analysis of this case gives the same result.

First the connections from layer A to B mature, or develop to their final values. That is, the initial c values are chosen at random, the set of differential equations given by Equation 3 (for $i = 1, 2, \ldots, N$) is solved, using the Q_{ij} function that applies to layer A activity. (For random snow activity in layer A, Q_{ij} is 1 when i and j are the same A cell, and 0 otherwise.) Knowing the mature c values for the A-to-B connections, as well as the Q_{ij} function for layer A, then allows us to compute the Q_{ij} function for the mature layer B. Then the development of the B-to-C connections is computed, using the Q_{ij} function appropriate to layer B. By repeating the process, we compute in turn the connection strengths for successive layers of connections.

Simulation results. A few parameters for each layer of cells determine the mature c values of the cells in that layer. These parameters include k_1 and k_2 and the breadth of the region in the previous layer that provides input to a cell of the developing layer. (See Figure 1.) As we shall see, the choice of the $k_{1,2}$ values determines the mature value of the total connection strength Σc_j of the inputs to the M cell.

When we explore the parameter space, we find that there are a limited number of ways each layer can develop. Briefly, we find that a sequence of feature-analyzing cell types emerges as one layer after another matures.

The first cell type emerges in layer B. There is a parameter regime in which each c value reaches its excitatory limit c_+. In this case, each B cell, once it has matured, computes the local average of the activity in the overlying region of layer A from which it receives input.

Once the B cells have matured in this way, nearby B cells have correlated activity. Each activity pattern in layer B is a blurred image of random snow. If one B cell's activity happens to be "high" at a given time, its neighbors' activities are likely to be "high" also. As a result of this activity correlation, a new cell type emerges in layer C. This *center-surround* cell type[1] acts as a contrast-sensitive filter—it responds maximally to a bright circular spot centered on the cell's receptive field, against a dark background. Center-surround cells having the reverse property—they respond maximally to a dark spot on a bright background—also emerge.

The Q function for pairs of center-surround cells in layer C determines the developmental possibilities for the C-to-D connections, and so on. We find that the next new type of feature-analyzing cell to emerge as we pass to succeeding layers is an *orientation-selective* cell. This cell responds maximally to a bright edge or bar against a dark background, or the reverse, when the edge or bar has a particular orientation. The receptive field map for such a computed cell is shown in Figure 2. This map is a contour plot showing the response of the cell to point illumination, as a function of the position of the illumination in visual space.

Each orientation-selective cell will develop to favor an arbitrary orientation if the network contains only feedforward connections as in Figure 1. However, if lateral connections between nearby cells of the orientation-selective cell layer are

included in the simulation, then the orientation preferences of the cells in the layer can become organized in certain arrangements. Cells having similar orientation preferences develop to occupy irregular band-shaped regions. (See reference 1 and the front cover, right side, of this issue.)

Discussion of the simulations. Center-surround cells are a prominent feature of mammalian retina. Orientation-selective cells emerge in cat and monkey visual cortex.[3,4] Irregular band-shaped regions of cells of similar orientation—called *orientation columns*—are a prominent feature in the orientation-selective cell layers.[3,4] (Once again, see the front cover, left side.) The role that lateral connections in cortex play in the formation of orientation selectivity is at present experimentally unsettled. As we noted, certain primates exhibit well-formed orientation selectivity at birth, in the absence of any structured visual experience.

Our point is not to suggest that feature-analyzing cells—particularly the center-surround cells—arise in animals in the same way they do in this synthetic network. As noted previously, the anatomy of inter-layer connections in the retina is more complex than a simple feedforward arrangement. Furthermore, center-surround cells can be constructed by a simple non-adaptive model in which excitatory inputs from some narrow region, and inhibitory inputs from a broader region, both converge on a cell. In our simulations we assumed that the breadth of the input region to a cell was the same for excitatory and inhibitory synapses, in order to avoid biasing the solution toward the formation of a center-surround cell type.

Our point is rather that a set of progressively more complex feature-analyzing cell types develops in the layered network, and that these cell types, and their organization, qualitatively exhibit some of the most salient features found in the first few stages of mammalian visual processing. The results suggest that some properties whose origin has been mysterious—such as orientation selectivity — may have a natural explanation in terms of the functioning of a Hebb-type development process in a layered network.

Two simple examples of how *structured* input to layer A would affect the simulation results are worth noting:

(1) If nearby pixels have correlated intensity values, and this is the only important input correlation present, then Q in layer A would resemble the Gaussian Q that we found in layer B. The subsequent development of the model would proceed in a way similar to that which we described, except that the appearance of each feature-analyzing cell type could be advanced one layer.

(2) If layer A is shown an ensemble of patterns, each consisting of sinusoidal stripes with arbitrary phase and orientation, then orientation selectivity can develop as early as layer B.[1]

We have assumed, for simplicity, that the statistical properties of the ensemble of presentations, that is, the covariances Q_{ij}, are unchanged or stationary during development. If the ensemble statistics change, cells that had reached their apparently final mature c values may change these c values in accordance with the new ensemble characteristics. Thus, although we always speak of cell *development*, the present approach is equally applicable to studying questions of cell *plasticity* during the life of the animal.

Hebb rules and optimization properties

We have seen that even a simple layered network with local feedforward connections obeying a Hebb-type rule develops a sequence of progressively more sophisticated feature-analyzing properties as we pass from one layer to the next. We will now examine some remarkable optimization properties of a Hebb-type rule.

Maximization of output activity variance. Consider a cell M that receives input from cells $L_1, L_2, \ldots, L_N$. Here and later, "input" means local input to cell M, not the environmental input to the network as a whole. Similarly, "output" refers to the M cell's activity value, not to the output from the network as a whole. Let the M cell's development be described as in Equations 1-4, with a saturation constraint on the range of each c value. We assume that the ensemble statistical properties of the L-cell activities, that is, the Q_{ij} function for the L cells as in Equation 4, are unaffected by the choice of c values. This is true if there is no feedback from M, or the cells it influences, to the L cells. It should be a satisfactory approximation if the feedback is present but is sufficiently weak, although this has not been studied quantitatively.

Define the function

$$E \equiv E_Q + E_k \tag{5}$$

where

$$E_Q \equiv -(1/2)<(M^\pi - \overline{M})^2> = -(1/2)\Sigma_i \Sigma_j Q_{ij} c_i c_j \tag{6}$$

and

$$E_k \equiv -k_1 \Sigma c_j - (k_2/2N)(\Sigma c_j)^2 \tag{7}$$

I have constructed the function E to have the property that $-\partial E/\partial c_i = \dot{c}_i$ for each i. This means that, as the Hebb rule causes each of the c values to change with time, the value of E, as a function of the c's, decreases along a path of locally steepest, or gradient, descent. (If $\dot{c}_i > 0$, then $\partial E/\partial c_i < 0$, so c_i increases and E decreases with time. If $\dot{c}_i < 0$, then $\partial E/\partial c_i > 0$, so c_i decreases and E again decreases with time.)

The value of E thus achieves a local minimum at cell maturity. Moreover, for the cases of interest here—including those that lead to the center-surround and orientation-selective cell types—this minimum is a *global* near-minimum as well.[1] We therefore will focus on the case in which the development process does not get stuck in high-lying local minima. This appears to be the typical case for a perceptual network exposed to a large ensemble of presentations, although it is an empirical finding and I have not established the limits of its validity.

What is the meaning of E achieving a global, or absolute, minimum value? For any given value of total connection strength Σc_j, E is minimized when $<(M^\pi - \overline{M})^2>$—the statistical variance of M—is maximized. Changing the values of the parameters $k_{1,2}$ adjusts, or tunes, the mature value of Σc_j. The E_k term, which is a function of Σc_j and $k_{1,2}$ only, plays a role similar to a Lagrange multiplier term, although E_k is parabolic rather than linear in Σc_j.

Therefore, the development rule of Equation 3 causes a cell to develop so as to maximize the variance of its output activity, subject to the constraint that the total connection strength have a given, parameter-determined, value and subject to the saturation bounds for each c value. Let us see intuitively what variance maximization means to a perceptual system.

Consider first a hypothetical M cell whose c values are such that the cell's output variance is zero. That is, regardless of the input values $(L_1^\pi, L_2^\pi, \ldots, L_N^\pi)$ chosen from the ensemble of presentations,

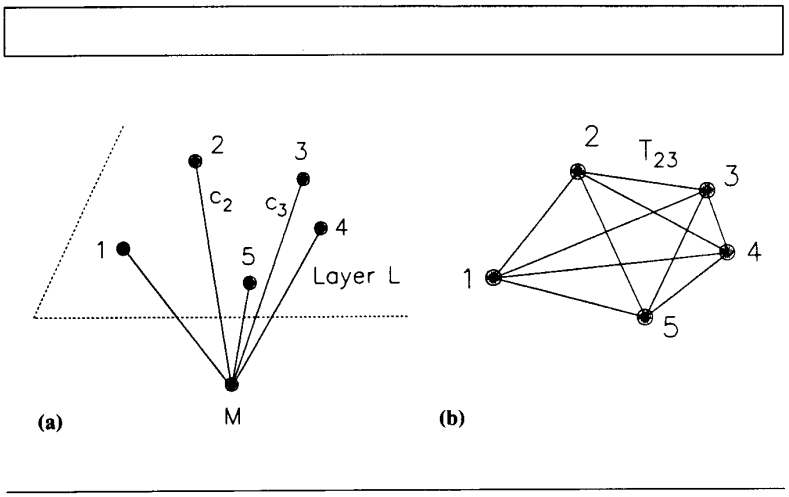

Figure 3. Relationship between networks: (a) a single M cell (with N inputs) of a layered self-adaptive network; (b) a Hopfield network with N cells and $N(N-1)/2$ connections, where $N=5$.

the output is always the same. This cell would be useless for conveying any information about the environment to later parts of the perceptual system.

On the other hand, if the c values are chosen in a different and special way, then the M cell's output value exhibits the largest possible spread or variance, consistent with the constraints on the c's, as the set of input values ranges over its ensemble. We have shown that a Hebb-type rule tends to generate c values satisfying this special condition. In an informal sense, provided certain conditions are met, the Hebb rule acting on our described M cell tends to produce an M cell whose output activity optimally preserves the information contained in the set of input activities. Later, we will make this statement more precise, by applying some concepts from information theory, and we will modify it to accommodate the situation in which multiple M cells interact with one another.

Optimization in another type of neural network. Hopfield[5] emphasized that the dynamics of a neural network can be described in some cases by the local minimization of a function. An interesting mathematical relationship exists between the E function defined in Equations 5-7 and Hopfield's energy function—although the network structure and behavior that each describes are very different.

Once again, our E function is $E \equiv E_Q + E_k$ where E_Q is shown in Equation 6. The development rule causes E_Q to be minimized subject to the constraint that Σc_j have a specified value and subject to the saturation constraints on each c value. The arrangement described by Equation 6 consists of one M cell with N inputs from cells $L_1, L_2, \ldots, L_N$ and is shown in Figure 3a. The $N \times N$ matrix of elements Q_{ij} is the covariance matrix of the input cell activities. The c's are the connection strengths from each input cell to the output cell. The minimization of E describes the development of the c's under the influence of the ensemble of inputs characterized by the covariance matrix Q.

In Hopfield's case,[5] as illustrated in Figure 3b, there are N cells and the activity state of the ith cell is called V_i. Each pair of cells is connected with fixed connection strength T_{ij}, so the number of connections is of order $N^2/2$, and the energy function is

$$E' \equiv -(1/2)\Sigma_i\Sigma_j T_{ij} V_i V_j \qquad (8)$$

The activities V_i change with time according to a linear summation rule with a threshold: V_i increases, unless it is already at its upper limit, if $\Sigma T_{ij} V_j > 0$, and decreases if $\Sigma T_{ij} V_j < 0$. If the T_{ij} matrix is symmetric, then the V_i's change so as to decrease the value of E' to a local minimum. Connection strengths are fixed; there is no learning or network development. The dynamical process described by Equation 8 is the change in the activities $\{V_i\}$ from some initial state to a final state of locally minimum E'. If we want to use the network for memory retrieval, a suitable choice of T_{ij} is given by an expression that is essentially the covariance of V_i^k and V_j^k over the ensemble of memories, indexed by k, to be stored.

Note that E' has the identical structure as our E_Q, if we identify V_i with c_i and T_{ij} with Q_{ij}. When T is a covariance matrix, Hopfield's network computes a local minimum of E' using N cells and order $N^2/2$ connections, explicitly embodying the T values. The state for which E' is minimal is the set of final activities $(V_1, V_2, \ldots, V_N)$.

One cell of our network computes a local minimum of the same function, our E_Q, using N connections. The Q function, which corresponds to T, is nowhere explicitly represented in the network. The Hebb rule implicitly responds to the covariance matrix, Q, as the ensemble of input patterns is presented to the M cell. The state for which E_Q is minimal is not a set of activities, but a set of mature connection strengths $(c_1, c_2, \ldots, c_N)$.

Thus, for T matrices that are covariance matrices, one cell of our network can locally optimize the same function as a fully connected Hopfield-type network. In our network, this optimization process consists of developing a final set of c values, starting with some initial set of values, under the influence of a statistically stationary ensemble of input patterns having covariance matrix T. In the Hopfield-type network case, the process consists of seeking a final set of cell activity values starting with some initial set of values, in a network whose connection strengths are fixed and prespecified to be the T values themselves.

These considerations lead to an interesting connection, only briefly outlined here, between memory retrieval and perception in a network model.

Memory retrieval and perception in a network model. If there are sufficiently few memory patterns to be stored, relative to N, then E' or E_Q will tend to have minima at the $\{V_i\}$ or $\{c_i\}$ values, respectively, corresponding to those memories. Depending upon the initial choice of the V's or c's, one or another of these memory states will be activated or selected. In the case of Hopfield's network, "activated" means that the final activity state

will match one of the stored memories. In the case of a cell in a layered self-adaptive network, "selected" means that the final set of c values will cause the M cell to be a matched filter for one of these memories. That is, the mature M cell will respond most strongly when presented with the set of input activities corresponding to that memory.

If the number of patterns in the ensemble is large, then the E_Q function will no longer capture details of any one of the patterns. The structure of the E_Q function may become simpler. The global minimum of E_Q will lie at the $(c_1, c_2, \ldots)$ value for which the M cell's variance is maximized. The mature M cell will function as a feature-analyzing cell, rather than as a matched filter to a particular memory. The particular feature or pattern element to which the mature cell will optimally respond, such as an oriented edge, need not even appear in any of the presented patterns.

Principal component analysis. There is a special case in which variance maximization corresponds to an important, and widely-used, statistical method for feature extraction. This is the case in which the output variance is maximized subject to the constraint that $\Sigma c_i^2 = 1$. Oja[6] showed that this maximization can be achieved by using a particular form of the Hebb rule, equivalent to

$$\dot{c}_i \propto < M^\pi (L_i^\pi - M^\pi c_i) > \qquad (9)$$

For this expression, we put $M^\pi \equiv \Sigma L_i^\pi c_i$ and define the activities, subtracting nonzero mean values if necessary, so that $< L_i^\pi > = 0$ for all i. The additional term in the Hebb-type rule, proportional to c_i, causes Σc_i^2 to be close to 1, and no explicit constraint needs to be imposed.

In statistics, principal component analysis, or PCA, is a standard method, reviewed in Huber,[7] for identifying "interesting" but unanticipated structure, such as clustering, in high-dimensional data sets. For example, an economist confronted with 1000 dimensions of data, such as the prices of different commodities, may want to know which several features of the data, for example, which several linear combinations of the 1000 quantities, are most salient.

PCA works as follows. Consider a set of data points indexed by π, each point $\mathbf{L}^\pi$ having coordinates $(L_1^\pi, L_2^\pi, \ldots, L_N^\pi)$. For PCA, we compute a vector $\mathbf{c}$ for which the projection of the set of data points onto the axis parallel to $\mathbf{c}$ has maximum variance. The projection of $\mathbf{L}^\pi$ onto $\mathbf{c}$, when $\Sigma c_i^2 = 1$, is just $M^\pi = \Sigma_i L_i^\pi c_i$, and the variance of the projected distribution is identical to the variance of M^π.

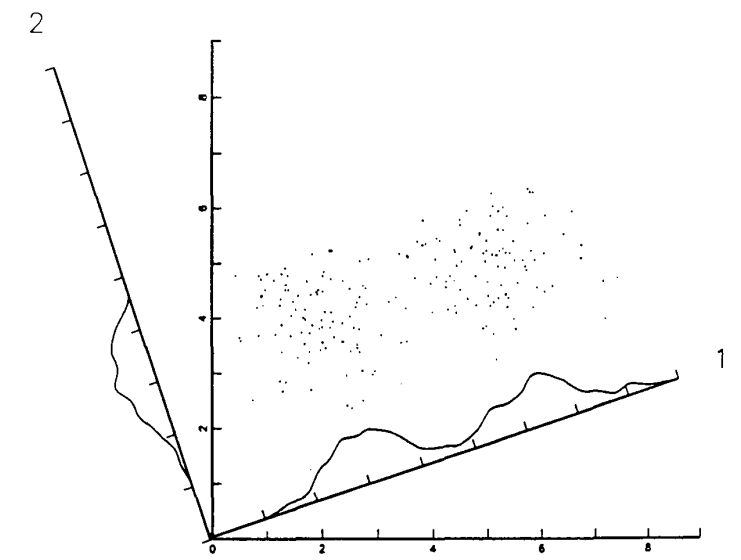

Figure 4. Illustration of principal component analysis. A cloud of data points is shown in two dimensions, and the density plots formed by projecting this cloud onto each of two axes 1 and 2 are indicated. The projection onto axis 1 has maximum variance, and clearly shows the bimodal, or clustered, character of the data.

An example of PCA is illustrated in Figure 4. Projecting the cloud of data points onto line 1 captures the salient feature of the data—that there are two clusters. The variance, or spread, of the data points along this axis is greater than for any other projection axis. Projecting the cloud onto line 2 would obscure the cluster structure. While the cluster structure is evident in the raw data of the two-dimensional plot shown here, such structure is often totally concealed in high-dimensional data sets, until an analysis method such as PCA is applied.

Since the PCA method corresponds to choosing $\mathbf{c}$ so as to maximize the variance of M^π subject to $\Sigma c_i^2 = 1$, it follows that the mature M cell generated by Oja's version of the Hebb rule performs PCA on its set of inputs.[6]

Optimal inference. Consider an arbitrary M cell characterized by a set of c values and having the linear response rule $M^\pi = \Sigma_i L_i^\pi c_i$ with $< L_i^\pi > = 0$ for all i. Suppose we know the c values, and are told a particular value of the output, M^π. We are asked to estimate the input activities $(L_1^\pi, L_2^\pi, \ldots, L_N^\pi)$ for that presentation. Let us score any such estimate by

(1) computing the difference between the estimate L_i^π (est) and the true value of L_i^π
(2) squaring this difference, and
(3) summing this squared error over i.

Averaging this score over an ensemble of presentations gives the mean square error

$$\text{MSE} \equiv \Sigma_i < [L_i^\pi - L_i^\pi(\text{est})]^2 > \qquad (10)$$

What estimation rule will give the best, meaning the minimum, MSE? For a linear estimation rule of the form $L_i^\pi(\text{est}) \equiv g_i M^\pi$, where we want to know what g values to use, the answer is found by

minimizing MSE with respect to each of the g_i's. This is easily done by differentiating MSE. It is also a simple case of the Gauss-Markoff theorem,[8] which applies more generally to the optimal estimation of a set of inputs given a set of outputs, rather than just one output. The result is

$$L_i^{\pi}(\text{opt est}) = M^{\pi} \times (\Sigma_j Q_{ij} c_j)/(\Sigma_i \Sigma_j c_i Q_{ij} c_j) \quad (11)$$

The MSE corresponding to this optimal estimate is then

$$\text{MSE(opt)} \equiv \Sigma_i < [L_i^{\pi} - L_i^{\pi}(\text{opt est})]^2 > = \Sigma_i < (L_i^{\pi})^2 > - H \quad (12)$$

where

$$H \equiv [\Sigma_i (\Sigma_j Q_{ij} c_j)^2]/(\Sigma_i \Sigma_j c_i Q_{ij} c_j) \quad (13)$$

Expressed in matrix form, with $\mathbf{c}$ denoting the column vector $(c_1, c_2, \ldots)$ and $\mathbf{Q}$ denoting the matrix (Q_{ij}), we have $H \equiv (\mathbf{c^T Q Q c})/(\mathbf{c^T Q c})$, where the superscript $\mathbf{T}$ denotes the matrix transpose.

The calculation so far involves a standard use of optimal estimation theory.[8] The linear filter, represented here by the set of c values, is specified. The result of a measurement using the filter—that is, the output value—is given. The task is to reconstruct the input values with minimum error, using a simple mean squared error criterion.

We now go beyond this simple framework to ask[9]: For what linear filter—what set of c values—is this minimum-error reconstruction the most accurate? That is, what choice of c's minimizes MSE(opt) of Equation 12?

Since the first term on the right-hand side of Equation 12 is independent of the c's, minimizing MSE(opt) is accomplished by maximizing H. The mathematical condition for this to occur is that the vector $\mathbf{c}$ be an eigenvector of $\mathbf{Q}$ having maximal eigenvalue. This is identical to the condition that $\mathbf{c}$ needs to satisfy in order for the M cell to perform PCA on its input values.

Therefore the PCA condition and the principle of optimal inference—namely, that MSE(opt) be minimized, or H be maximized—lead to the same set of c values. A Hebb rule of the form of Equation 9 generates an M cell that satisfies both conditions. In the presence of other constraints, or additional cost terms, there is no guarantee that PCA and H-maximization are equivalent, since the PCA principle maximizes the quantity $(\mathbf{c^T Q c})/(\mathbf{c^T c})$ which is not identical to the expression for H in Equation 13.

Optimization in the presence of processing noise and constraints on output variance. We have identified several optimization properties related to the cell's output variance. Suppose, however, that for some reason the variance is itself constrained. For example, the output activity may be confined to lie within some operating range. This is a biologically plausible situation. In this case, what is optimized by a suitable Hebb-type rule? We will discuss this case for a particular processing model, giving only the main results and omitting the details.

Suppose the signal L_j^{π} on the jth input line or connection is corrupted by noise, v_j^{π}, where v_j^{π} has a mean of zero and a variance B, and is uncorrelated both with the noise on other input lines and with any of the input signals L_j^{π}. The cell computes the weighted sum $x \equiv \Sigma_j (L_j^{\pi} + v_j^{\pi}) c_j$. The variance of x is the sum of two terms: the variance due to the signal in the absence of noise, $\Sigma_{ij} Q_{ij} c_i c_j$; and the variance due to the noise, $B \Sigma c_j^2$. Consider a suitable synaptic modification rule in which $\dot{c}_i$ contains a term of the form $< L_i x >$. This rule causes the model cell to develop such that the variance of x due to the signal is maximized relative to the variance due to the noise. This type of signal-to-noise optimization property can also emerge when the cell's output M is a monotonic nonlinear function of x, such as a sigmoid function, if the synaptic modification rule is of the form described.

Adaptive signal processing. Returning to the case of a linear-response model neuron, suppose we wish to train a linear cell to respond to each of a set of prescribed input vectors by generating an output that best matches a prescribed desired output. An input vector is denoted $\mathbf{L}^{\pi} \equiv (L_1^{\pi}, L_2^{\pi}, \ldots, L_N^{\pi})$ and each desired output is a scalar number M_{des}^{π}. The actual output is $M^{\pi} = \Sigma_j L_j^{\pi} c_j$ where each $< L_i^{\pi} > = 0$ and the optimal c values are to be determined by a learning process. A mean square measure of error is used:

$$\text{MSE}' = < (M^{\pi} - M_{\text{des}}^{\pi})^2 > \quad (14)$$

where $< \ldots >$ again indicates the ensemble average. MSE′ is a minimum when the c values are chosen to satisfy $< M_{\text{des}}^{\pi} L_i^{\pi} > = \Sigma_j Q_{ij} c_j$ for all i. (Recall that $Q_{ij} \equiv < L_i^{\pi} L_j^{\pi} >$.) The least mean square, or LMS, algorithm of Widrow and Hoff[10] uses an estimate of the gradient of MSE′ and in effect performs gradient descent to compute the optimal c values. An ensemble-averaged form of the algorithm can be written as

$$\dot{c}_i \propto < L_i (M_{\text{des}}^{\pi} - \Sigma_j L_j^{\pi} c_j) > \quad (15)$$

Equations 14 and 15 give an objective function to be minimized and an algorithm for a *supervised* learning process. Both the inputs and the desired outputs are presented to the cell, and the error term $(M_{\text{des}}^{\pi} - M^{\pi})$—the amount by which the actual output differs from the desired output—is fed back to change the c values until the mean square error is minimized.

Our optimal inference criterion, namely, the minimization of the objective function of Equation 12, and a Hebb-type rule that implements it (Equation 9) are formally similar to Equations 14 and 15. But the optimal inference criterion provides a method for *unsupervised* learning. The criterion does not make any use of a desired output; it simply states that the M cell should have the property that knowing its output activity value allows one to infer the input activities with greatest possible accuracy.

Information theory and the principle of maximum information preservation

For a single M cell receiving inputs from a given set of L cells, we have seen that, for a particular Hebb rule given in Equation 9, knowledge of the output activity value allows inference of the input values with greatest accuracy, in the sense of minimum mean squared error. For more general Hebb-type rules, we found that the variance of the output activity was maximized subject to various constraints. This result led us to suggest that, at least in an intuitive sense, a Hebb rule may act to generate an M cell whose output activity *preserves maximum information* about the input activities, subject to constraints.

We will now make this notion of maximum information preservation more precise, and will extend it to the case of an entire layer of M cells, by introducing some concepts from information theory. The goal is to see what this principle implies for the development of each layer of a perceptual system. That is, given the statistical properties of the ensemble of

input patterns at layer L, and certain constraints, what particular processing functions do the connections from layer L to layer M, and within layer M, develop to implement?

Shannon information. We will regard each presentation of real-valued inputs $\mathbf{L} = (L_1, L_2, \ldots, L_N)$ as a message, where L_i denotes the activity of the ith L cell in the layer. We omit the π superscripts for clarity. Strictly speaking, even one real number carries an infinite amount of information. To avoid encountering expressions of the form, $\infty - \infty$, and because infinite precision is physically and biologically meaningless, we will think of the N-dimensional space of the **L** vectors as being divided into small boxes. Each box is labeled by its location **L**. Two messages are regarded as identical if they lie in the same box. In the end, we will pass to the continuum limit, and the sums will become integrals.

Given an ensemble of messages, let $P(\mathbf{L})$ be the probability that a randomly chosen message lies in box **L**. Shannon[11] showed, in a classic paper, that the information conveyed by sending a message that lies in box **L** is $I(\mathbf{L}) = [-\ln P(\mathbf{L})]$. The average information conveyed per message is $< [-\ln P(\mathbf{L})] > = -\Sigma_{\mathbf{L}} P(\mathbf{L}) \ln P(\mathbf{L})$, where $< \ldots >$ is the usual ensemble average. If the base-2, rather than natural, logarithm were used here, the information would be measured in bits.

Now suppose each input presentation **L** generates a set of output values, denoted by the vector **M**, via some known computation. Suppose that we are told the value of **M**—or, more strictly, which discrete box **M** lies in. (In general, **M** will not be uniquely determined by **L** because noise may be introduced in the computation of **M**.) How much additional information would we need to reconstruct the input message **L** that gave rise to **M**? (Shannon calls the ensemble average of this amount of additional information the equivocation.)

The answer is $I_{\mathbf{M}}(\mathbf{L}) = [-\ln P(\mathbf{L}|\mathbf{M})]$, where $P(\mathbf{L}|\mathbf{M})$ is the conditional probability of the input message lying in box **L** given that the output lies in box **M**. Therefore the amount of information that knowing **M** conveys about **L** is the difference, $I(\mathbf{L}) - I_{\mathbf{M}}(\mathbf{L}) = \ln[P(\mathbf{L}|\mathbf{M})/P(\mathbf{L})]$. The ensemble average of this quantity is the rate R, per message, of transmission of information from the cell's inputs to its output. This is the average amount of information that knowing **M** conveys about **L**. We have

$$R = <\ln[P(\mathbf{L}|\mathbf{M})/P(\mathbf{L})]> \tag{16}$$

We have a standard identity $P(\mathbf{L}|\mathbf{M})P(\mathbf{M}) = P(\mathbf{L},\mathbf{M}) = P(\mathbf{M}|\mathbf{L})P(\mathbf{L})$, where $P(\mathbf{L},\mathbf{M})$ is the joint probability that the input lies in box **L** and the output lies in box **M**. Using this gives

$$\begin{aligned} R &= <\ln[P(\mathbf{M}|\mathbf{L})/P(\mathbf{M})]> \\ &= -<\ln P(\mathbf{M})> + <\ln P(\mathbf{M}|\mathbf{L})> \\ &\equiv <I(\mathbf{M})> - <I_{\mathbf{L}}(\mathbf{M})> \end{aligned} \tag{17}$$

The right-hand side is the ensemble average of the total information conveyed by **M**, minus the information that **M** conveys to one who already knows **L**. This second term is the "information" that **M** conveys about the processing noise, rather than about the signal **L**.

Maximum information preservation. Let us now state the proposed principle of maximum information preservation for each layer, or processing stage, of a perceptual network: Given a layer L of cells, and the stationary ensemble statistical properties of the signal activity values in the layer, and given that layer L is to provide input to another cell layer M, the transformation of activity values from L to M is to be chosen such that the rate R of information transmission from L to M is maximized, subject to constraints and/or additional cost terms. These constraints or costs may reflect, for example, biochemical and anatomical limitations on the formation of connections, or on the character of the allowed transformations.

The formulation of this principle arose from studying Hebb-type rules and recognizing certain optimization properties to which they lead for single M cells. Once formulated, however, the principle is independent of any particular local algorithm, whether Hebb-related or otherwise, that may be found to implement it. Let us explore

(1) the consequences of the principle for some simple cases;
(2) how the principle might be implemented; and
(3) how it may fit within a broader view of neural development.

A single M cell. Under certain conditions, maximizing the output activity variance of the M cell maximizes the Shannon information rate R. We illustrate this for a particularly simple but instructive case. The argument can be made somewhat more general than this, but it is not true that maximum information rate and maximum activity variance coincide when the probability distribution of signal values is arbitrary.

Suppose the M cell receives inputs from a set of L cells $L_1, L_2, \ldots, L_N$, and that the M cell's output in the presence of processing noise has the form

$$M^{\pi} = (\Sigma_i L_i^{\pi} c_i) + \nu^{\pi} \tag{18}$$

Here π indexes the particular set of input and output values, so that if **L** is repeated but the output M is different, owing to noise, this counts as a different set of input-output values. The quantity ν^{π} is the noise, a random variable differing from one presentation to the next. Suppose that

(1) M has a Gaussian distribution, with variance denoted by V;
(2) ν has a Gaussian distribution with a mean of zero and variance denoted by B; and
(3) ν is uncorrelated with any of the input components; that is, $<\nu L_i> = 0$ for all i.

Then, omitting the details, we find that the information rate is

$$R = (1/2)\ \ln(V/B) \tag{19}$$

For a given noise variance, B, this rate is maximized by maximizing the output variance V of the M cell. Note that V/B is essentially a signal-to-noise ratio.

Suppose that the noise model consists instead of independent Gaussian noise, ν_i, being introduced on each input line i, where each ν_i has variance B. Then $M^{\pi} = \Sigma_i (L_i^{\pi} + \nu_i^{\pi}) c_i$, and the information rate is found to be $R = (1/2)\ \ln[V/(B\Sigma c^2{}_i)]$. In this case, R is maximized for fixed B when $(V/\Sigma c_i^2)$ is maximized—that is, when the connection strengths are chosen so as to perform principal component analysis on the cell's inputs.

Redundancy and diversity. Suppose there is an arbitrary number of L cells but just two coupled linear M cells. Each M cell's output is some linear combination of the L cell's activities:

$$M_1^{\pi} = (\Sigma_i t_{1i} L_i^{\pi}) + \nu_1^{\pi} \tag{20}$$

$$M_2^{\pi} = (\Sigma_i t_{2i} L_i^{\pi}) + \nu_2^{\pi} \tag{21}$$

Each noise term is Gaussian and of variance B, the noise terms for the two M cells are uncorrelated with each other, and each noise term is uncorrelated with any of the

L cell activities. We treat the case in which M_1 and M_2 have Gaussian distributions, with $<M_1^n> = <M_2^n> = 0$. Our task is to determine what values of the t_{ni}'s lead to the maximum information being preserved during the processing of L-cell activities to give M-cell output activities.

Note that the t_{ni}'s do not in general stand for the strengths of particular connections. There may be both feedforward and lateral (M-to-M) connections whose joint effect, possibly over several time steps, is to produce the M-cell outputs of Equations 20 and 21. Our concern here is not with the particular connection strengths, nor with the development rule that may implement them, such as a Hebb-type rule, but rather with understanding what cell response properties—what t_{ni} values—are induced by the principle of maximum information preservation.

Omitting details of the proof, the resulting information rate for this case is

$$R = (1/2)\ \ln(\text{Det}\ Q^M) - \ln B \qquad (22)$$

where the elements of the 2×2 covariance matrix Q^M are $Q^M_{nm} \equiv <M_n^n M_m^n>$ and "Det" denotes the determinant. We find

$$\text{Det}\ Q^M = B^2 + B(W_1 + W_2) + W_1 W_2 (1 - \varrho_{12}^2) \qquad (23)$$

where W_n is the output variance of cell M_n in the absence of noise, and ϱ_{12} is the correlation coefficient of the activities of M cells 1 and 2, also in the absence of noise.

To maximize R, given B, we must maximize Det Q^M. When B is large, the third term on the right-hand side of Equation 23, which is independent of B, is small compared with the second term, which is of order B. In that case, maximizing Det Q^M means maximizing $(W_1 + W_2)$. If no constraint prevents us, we can achieve this maximization by maximizing W_1 and W_2 separately. But this means constructing each M cell so that its output variance, which is W_n in the absence of noise, or $W_n + B$ in the presence of noise, is maximized. This is exactly what we found to be the optimum solution when there is only one M cell. (See Equation 19.)

If the noise B is smaller, then the third term becomes relatively more important. The rate R is then maximized by making an optimal tradeoff between keeping W_1 and W_2 large, and making the responses of the two M cells uncorrelated.

We have thus found that, depending upon the noise level, there is competition between the value of having redundant M cell responses, which mitigate the information-destroying effects of noise, and the informational value of having different cells extract different linear combinations of the input. A high noise level favors redundancy. In this case, both M cells compute the same linear combination of inputs, if there is only one such combination that yields maximum output activity variance. A lower noise level favors diversity of response. In this case, the M cells compute different linear combinations of the L cell activities, even though each M cell's output variance may be reduced as a result of this choice.

To make this more concrete, consider a simple example. There are two L cells, and the Q matrix for L cell activity has $Q_{11} = Q_{22} = 1$ and $Q_{12} = Q_{21} = q$ with $0 < q < 1$. We arbitrarily impose the constraint that $t_{n1}^2 + t_{n2}^2 = 1$ for each M cell $(n = 1, 2)$.

The solution that maximizes the preservation of information then has $t_{11} = t_{22}$ and $t_{12} = t_{21}$, and the values of t_{11} and t_{12} are given in Figure 5 as a function of B and q. For large B, both M cells receive the same linear combination of inputs: $(L_1 + L_2)/\sqrt{2}$. For smaller B, the cells measure different linear combinations of L_1 and L_2. In the limit as B approaches zero, one M cell receives input only from cell L_1 and the other only from L_2.

A layer of M cells with nonlinearity and lateral connections. What does the principle of maximum information preservation, which we shall call the *infomax principle*, imply qualitatively in this more general case? Maximizing R means that we attempt to (1) maximize the total information conveyed by the output message **M**, and (2) minimize the information that **M** conveys to one who already knows the input message **L**. These criteria are related, but not equivalent, to the property of encoding signals so as to reduce redundancies present among the inputs to the perceptual system. The general idea that information theory can be useful for understanding perception is an old one. Significant contributions were made by Attneave in 1954, Barlow in the 1950s and 1960s, and Marr in 1970. Much of this work has focused on the role of redundancy reduction. This property is one, but only one, aspect of the infomax principle. For example, we have seen that infomax also leads to the introduction of redundancy when this is useful in countering the effects of noise.

I have analyzed the qualitative consequences of the infomax principle in some very simple models.[12] The results show that the principle can, under certain conditions, lead to L-to-M transformations with the following properties:

- Topographic mapping from layer L to layer M, when the spatial extent of lateral connections within layer M is assumed to be limited. That is, near-neighbors in L tend to map to near-neighbors in M.
- Map distortions, in which a greater number of M cells tend to represent the types of layer-L excitation patterns that occur more often.
- The infomax principle selects which features of the input signals are represented in layer M. Features having relatively high signal-to-noise ratios are favored. This is the extension of our previous redundancy-diversity result to the full-layer case.
- Orientation-selective cells, and the arrangement of such cells in orientation columns, can emerge for some very simple types of model input.
- When time-delayed information is made available to the layer, the infomax principle can cause M cells to extract and encode temporal correlations, in a manner similar to the extraction of spatial correlations.

I must emphasize that much work is required to determine the consequences of the infomax principle for cases involving more biologically realistic patterns of activity.

Discussion

From a simplified set of assumptions—a linear summation response, a simple Hebb-type rule having a covariance form, and feedforward connections only—we derived an optimization principle for the development of a single cell. This principle states that the mature M cell is such that its output activity variance is maximized subject to constraints. More generally, we can have cost terms instead of, or in addition to, constraints. Then the function maximized involves both the variance and the additional cost function.

This led us to infer a proposed principle of maximum information preservation, subject to constraints. It is equivalent to variance maximization in some simple cases, but it has a much broader scope. For example, it can be applied to cases in which a layer of L cells provides input to an entire

layer of M cells, with lateral as well as feedforward connections. It can likewise apply to cases in which the response function is not necessarily linear.

The consequences of this proposed principle are only beginning to be explored. One set of issues that needs clarification is the choice of biologically appropriate constraints and cost terms. A second, related, issue involves the choice of algorithms, whether of Hebb type or otherwise, that control the development of feedforward and lateral connections so as to implement the optimization principle. While much work needs to be done, I suggest that this principle, or something like it, may play an important role in determining the character of perceptual processing at least in its early stages, where there is a chance that feedback influences may not affect the development of feature-analyzing function in an essential way. Possibly the principle may play some role even in the presence of significant feedback, but it is not clear at this time how best to analyze this case.

What might we expect to be the character of a layer of cells developing according to the principle of maximum information preservation, for cases of biological interest? Although the necessary calculations for sufficiently realistic cases have not yet been carried out, we can speculate on the outcome.

Suppose there is a constraint on the distance within layer M over which the activity of one M cell can affect another. There might be, for example, a constraint on the length of lateral connections. Suppose also that each region of layer M "sees", or receives input from, only a limited region of layer L, and that nearby regions of M "see" nearby regions of L. Then, if the noise variance B is large, and there are not many M cells that "see" the same set of L cells, we might find that each M cell develops so as to maximize its activity variance, and performs processing that is redundant with that of many of its neighbors.

On the other hand, if B is smaller, or there are a large number of M cells that "see" the same L region, we may expect that the M cells in a region do not all perform the same processing function on the inputs from the L layer. Instead, they might span a range of feature-analyzing properties, each of which has a moderately high variance.

In the visual system of cats and monkeys, there are multiple layers of center-surround cells, followed by layers of orientation-selective cells. The orientation-selective cells begin at a different layer in cats than in monkeys. It is possible that, in response to the ensemble of inputs seen by a particular layer, the layer can develop either center-surround or orientation-selective cells, as occurred in our previous model simulations.[1] Perhaps a parameter such as the noise level B "tunes" for redundancy or diversity of response. Redundancy could favor center-surround cell formation, with many cells performing substantially the same processing function. Diversity could favor the formation of orientation-selective cells spanning the entire range of orientation preferences within each region of the layer. (Of a group of cells comprising all orientation preferences, only a small fraction will fire when presented with an oriented edge of illumination.) Hubel and Wiesel discovered[3,4] that orientation-selective cells are arranged, within a cortical layer, so that each small region of cortex ($\approx 1 \times 1$ millimeter) contains the "machinery" for analyzing substantially all edge orientations seen by either eye within a small region of visual space. Perhaps the principle of maximum information preservation, combined with limits on lateral interaction distance, can account for this efficient organization.

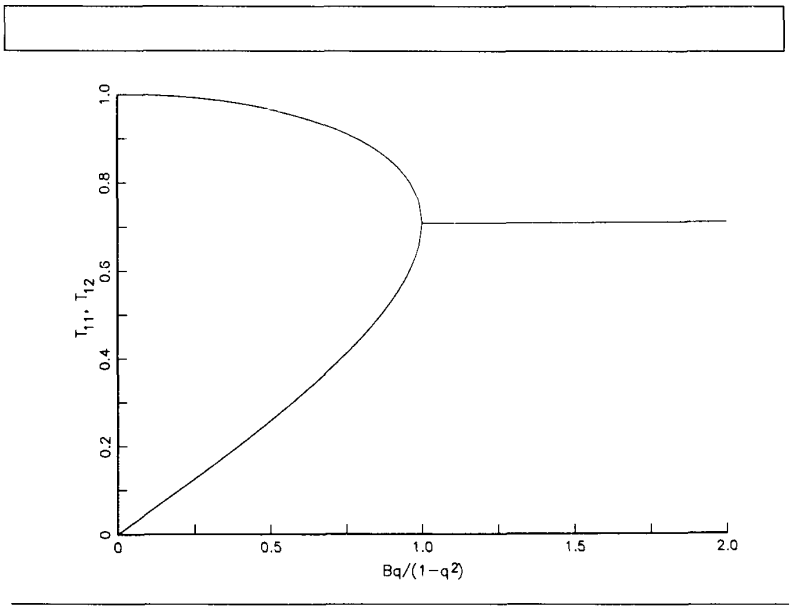

Figure 5. Values of the coefficients t_{ni} that maximize the preservation of information from layer L to M, for a simple case with two L and two M cells. Each M cell output (see Equations 20 and 21) includes random noise having variance B, and q is the correlation, or covariance, of the activities of the two L cells. For $x \equiv Bq/(1-q^2) \geq 1$, both M cells redundantly compute the same linear combination of the L cell activities (all $t_{ni} = 1/\sqrt{2}$). For $x < 1$, the optimal t values satisfy $t_{11} = t_{22}$ and $t_{12} = t_{21}$, where the upper curve gives t_{11} and the lower curve gives t_{12}, or the reverse. The curves for $x < 1$ are given by $y = (1/2)[(1+x)^{1/2} \pm (1-x)^{1/2}]$; this is derived by maximizing Det Q^M. (See Equation 23.)

Local algorithms. The infomax principle is stated in terms of maximizing a complicated expression (see Equation 16). Is there an algorithm or process that deals with much simpler quantities and computations—local to each cell or pair of connected cells in a network—and yet implements the infomax principle, at least approximately?

I have found[12] that, for some simple cases, a Hebb-related algorithm developed by Kohonen[13] implements some of the qualitative features required by the infomax principle. This algorithm was developed to show how lateral connections can

induce topographic order in a simple model, and makes no reference to noise or information content. These results suggest that it may be possible to devise a local algorithm that more fully embodies the requirements of the infomax principle.

The relationship between the principle and such an algorithm would be complementary. The principle would suggest what the function of the algorithm, and the lateral connections it describes, might be—that is, what role the processes and connections might serve in the construction of a perceptual system. The algorithm would show how a complex optimization principle could be implemented by a network of cells that individually have little computational power.

Although I have focused on algorithms that perform activity-dependent modification of connections, other types of mechanisms may be used to implement a given optimization principle. Biochemical cell-cell adhesion markers, chemical or other gradients that may help to establish topographic maps, particular cell types that implement complex types of connectivity (as in the retina), and other mechanisms may all play a role. An organizing principle by itself does not determine the many design details that a particular system—biological or synthetic—may use to implement it.

Infomax and perceptual data. Why might it be important for a perceptual system to maximize the amount of information preserved from one layer to the next?

Presumably, one goal of a perceptual system is to provide the brain with the means of discriminating different environmental situations that may demand different responses by the animal.

For a very simple network with only a couple of layers of processing from environmental input to motor output, we could imagine using some sort of supervised learning mechanism. The mechanism would pair inputs with the desired output responses and adjust the connection strengths accordingly. Such a process involving more than a few layers, however, appears biologically implausible, and its performance may scale poorly as the number of layers is increased.

In a complex network, or in an animal's brain, it is totally unclear how a component layer is to "decide" what transformation its connections should perform—*if* we assume that the layer needs to "know" what environmental features are important for the animal to respond to. This is the classic artificial intelligence credit assignment problem: if the final output from a complex system is correct, which connections should be rewarded or strengthened?

The approach we propose avoids this problem. Instead of requiring that a connection or layer "know about" the ultimate goals of the animal, we use only local information. The information that reaches a layer is processed so that the maximum amount of information is preserved. We have seen that this does not in general lead to a trivial one-to-one identity mapping, in which each M cell receives input from only one L cell. In general, the identity mapping is not a solution that maximally preserves information, owing to the role of noise in our model. Instead, each M cell tends to respond to features that are statistically and information-theoretically most significant, in a sense similar to that of principal component analysis. Applying the principle of maximum information preservation to each layer of processing in turn, results in the emergence of a sequence of feature-analyzing functions.

The following analogy may help you to see intuitively how the process works. Imagine a person in an organization, whose job is to make the most informative possible summary of the data that he receives each week. The type of data he receives depends upon the environment external to the organization, the structure of the organization (what "layer" he is part of), and various constraints. Over time, he finds that a particular representation of information—for example, graphical plots involving various variables—serves him best in preparing his summary. If he is allowed to interact with others in his "layer", the criterion can be broadened (as we did for the cells) to state that the composite output of his layer should be as informative as possible.

Note that some set of processing functions will end up being provided by this person's "layer", without the workers needing to know either what the goals of the entire organization are, or what information is deemed most important by their superiors in later "layers".

In both the organizational analogy and the real network, there is no need for any higher layer to attempt to reconstruct the raw data from the summary. The point is rather to enable the higher layers to use environmental information to discriminate the relative value of different actions. If the needed information has been lost at intermediate stages, it cannot be used. If a local optimization principle is to be used—one that does not attempt to take account of remote high-level goals—then we do not know what particular information is going to be needed at high levels. Since we don't know what information we can afford to discard, it is reasonable to preserve as much information as possible within the imposed constraints. The principle of maximum information preservation thus appears to be an extremely natural and attractive one to use in the construction of a layered perceptual system.

Evolution and infomax. The infomax principle may determine what transformation each layer of a given network will implement. However, it does not specify the "gross architecture" of the network; that is, which layers provide input to which other layers. Nor does it specify the various parameters that may affect layer development, such as noise level, the allowed range of lateral connections, and so on. These aspects of the design may be determined by biological evolution, or by other principles not yet identified.

For an analogy, think of an electronic circuit designer who is not free to modify the properties of the components he or she uses, but who can connect them to form a variety of circuits. In the case of our proposed principle, each "component" is an entire cell layer, and the infomax principle determines that layer's behavior given a particular gross architecture or "circuit design". Thus evolution can "close the loop" on the design process, favoring the survival of organisms whose perceptual systems are well-adapted to their environment.

There is a separate and important evolutionary function that a *generic* principle for the development of a perceptual network layer—whether it be infomax or some other principle—can serve. Suppose that an evolutionary mutation produces a modified eye, or merges auditory signals into the visual pathway at some new point. If there were no generic principle for layer development, we might imagine that mutations would have to occur simultaneously in the processing function of several layers, for those layers to be able to use the novel input properly. But if there is such a generic principle—one that applies to each layer regardless of what type of input reaches it—then the novel input will automatically be processed in accordance with that principle. This suggests that the existence of a generic principle may greatly

increase the likelihood of a mutation being adaptive.

A broader context. Other complex systems, besides neural networks, pose challenges similar to those we have discussed. How might complex structures and behaviors that may appear goal-oriented emerge from relatively simple local rules? We have seen that a local dynamical rule of Hebb type, acting at synapses, leads to an optimization principle—variance maximization—at the level of the whole cell. This suggested an optimization principle—maximum information preservation—that may apply at the level of an entire layer. From the standpoint of information theory, we may find that the immune response system and biological evolution, among other complex systems, have certain abstract similarities to the process of neural development and plasticity, although the dynamical rules and the substrates upon which they act are quite different.

A great deal of work remains to be done, if we are to take this or some other proposed organizing principle, extract testable predictions from it, and determine its scope and limitations. We need to identify and test such principles, in order to complement and help to focus the enormous amount of detail being revealed by progress in experimental neuroscience. The study of such principles may also provide the understanding needed to develop synthetic perceptual systems that require no explicit programming.□

References

1. R. Linsker, "From Basic Network Principles to Neural Architecture" (series), *Proc. Nat'l Academy of Sciences USA*, Vol. 83, Oct.-Nov. 1986, pp. 7508-7512, 8390-8394, 8779-8783.
2. D.O. Hebb, *The Organization of Behavior*, Wiley, New York, 1949.
3. D.H. Hubel and T.N. Wiesel, "Brain Mechanisms of Vision," *Scientific American*, Vol. 241, Sept. 1979, pp. 150-162.
4. D.H. Hubel and T.N. Wiesel, "Functional Architecture of Macaque Monkey Visual Cortex" (Ferrier lecture), *Proc. Royal Society London*, Vol. B198, 1977, pp. 1-59.
5. J.J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proc. Nat'l. Acad. Sci. USA*, Vol. 79, April 1982, pp. 2554-2558.
6. E. Oja, "A Simplified Neuron Model as a Principal Component Analyzer," *J. Math. Biology*, Vol. 15, 1982, pp. 267-273.
7. P. Huber, "Projection Pursuit," *Ann. Statistics*, Vol. 13, No. 2, June 1985, pp. 435-475.
8. P.B. Liebelt, *An Introduction to Optimal Estimation*, Addison-Wesley, Reading, Mass., 1967.
9. R. Linsker, "Development of Feature-Analyzing Cells and their Columnar Organization in a Layered Self-Adaptive Network," in R. Cotterill, ed., *Computer Simulation in Brain Science*, Cambridge Univ. Press, pp. 416-431, in press.
10. B. Widrow and S.D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, N.J., 1985.
11. C.E. Shannon, "A Mathematical Theory of Communication," *Bell Systems Tech. J.*, Vol. 27, 1948, pp. 623-656.
12. R. Linsker, "Towards an Organizing Principle for a Layered Perceptual Network," in D. Anderson, ed., *Neural Information Processing Systems—Natural and Synthetic*, Amer. Inst. of Physics (NY), to appear.
13. T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, 1984.

Ralph Linsker is on the research staff of the IBM T.J. Watson Research Center. His chief research interests are in the fields of neuroscience and machine perception and learning. Since coming to IBM in 1981 his interests have also included computer design and design automation and medical laser applications. He holds five patents and received IBM's Outstanding Innovation Award for a printed-circuit interconnection design system that has been in production use at IBM since 1983. In the field of medical lasers, he and his colleagues found that pulsed far-ultraviolet radiation can ablate blood-vessel lesions without thermal damage.

Linsker received the BA (summa cum laude) and PhD (in 1972) in theoretical physics from Columbia University, and the MD (in 1976) from Cornell University Medical College. He has done work in theoretical physics at the Princeton Plasma Physics Laboratory, and medical work at New York Hospital.

Linsker's address is Rm. 35-110, IBM T.J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598.

March 1988