and other learning problems of biological relevance.

#### *4.1.1 Building Receptive Fields With Multiplication*

Thus, a lookup table has a natural biological implementation given the ability to construct, place, and weight multi-dimensional receptive fields in the input space. A simple fact that has often been noted in the literature (e.g. Ballard, 1987; Mel, 1988; Poggio & Girosi, 1990) is that higher-dimensional receptive fields can be constructed by multiplying together the outputs of units with lower-dimensional receptive fields. For example, figure 5(a) shows that a 2-dimensional Gaussian results by multiplying two 1-dimensional Gaussians. In a sensory context, this is to say that a neuron selective simultaneously for stimulus parameters along two dimensions, such as the amplitude and frequency of an auditory stimulus (Suga, 1982), could be derived by multiplying the outputs of two afferent fibers, each selective along only a single stimulus dimension. As shown in fig. 5(b), the multiplication operation can also be approximated by a sum followed by a threshold.

It may therefore be conjectured that a biological implementation of a lookup table should be associated with i) an abundance of input neurons with ready-made low-dimensional receptive fields, and ii) multiplication of small groups of input fibers as an elemental operation of each output cell. In support of i), perhaps the most consistent feature of cortical representation at the single-unit level is the almost universal presence of place-coding (Ballard, 1986), i.e. the dedication of a set of units to each relevant input dimension(s), whose bell-shaped receptive fields overlap and whose peak sensitivities move systematically across the input space. For example, a visual neuron of area V1 may be described as having a 4-dimensional receptive field in the space of retinal x and y, orientation, and binocular disparity (Ballard et al., 1983). A multiplicative combination of two such units at a common site of projection would give rise to an eight-dimensional receptive field in the space of images. Multi-dimensional receptive fields are not limited to the visual cells, but occur in all modalities, such as the multi-joint neurons of the posterior parietal cortex (Hyvarinen, 1982; Mountcastle, 1975), and the combination sensitive neurons of the bat auditory cortex (Suga, 1982). Evidence in support of ii), i.e. the possibility of multiplicative nonlinearities in cortical dendrites, is taken up in detail in section 6.2.3.

### 4.2 Learning with Higher-Order Features

What role may multiplication play in the learning of pseudo-random associative maps over high-dimensional spaces? For difficult association learning problems in high-dimensional input spaces, such as the space of images, it may become impractical during learning for each output neuron to explicitly "record", in their entirety, each of a large number of complex images with which it was coactivated. With reference to the lookup-table algo-

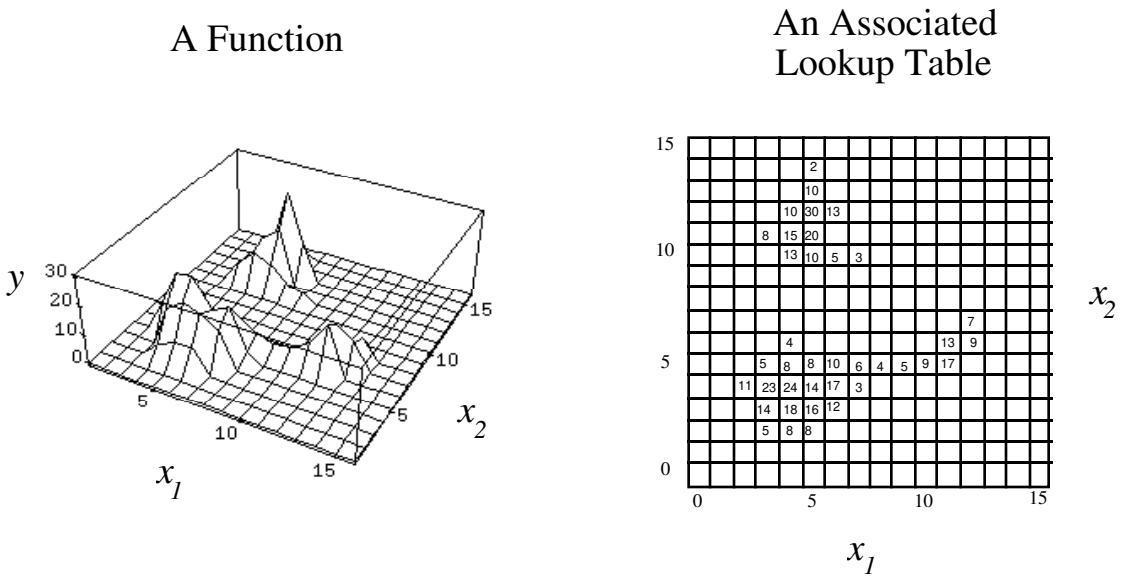


Figure 3: Graph of a function of two variables, and a corresponding lookup table. Table entries are assumed to have value zero by default. From a neural perspective, table entries at right may be viewed as nonoverlapping receptive fields.

Receptive fields are distributed over the input space only where the function has non-zero value, and are weighted according to the average height of the function within their respective receptive fields.

The classical disadvantages of tabular learning derive from the exponential time/space requirements associated with the construction of a multidimensional lookup table[6]. However, several effects can mitigate the problem of exponential blowup. First, trivially, we have defined the functions of interest, including a large class of biologically-relevant functions, to be of relatively low input dimension, i.e. $d_{\mathcal{X}} < 10$. For this case, the lookup table can be of tractable size. Second, the functions to be learned may be sparsely active, i.e. mostly 0-valued, such that table space may be substantially conserved using standard representational techniques; in the neural analogy of fig. 4, we simply omit receptive fields everywhere the function is 0-valued. Third, the functions to be learned may be locally smooth (even though nonlinear), such that fewer table entries (or receptive fields) are needed to achieve a given accuracy of local approximation (Poggio & Girosi, 1990). Fourth, large portions of the multi-dimensional input space may be of little or no use in practice. These regions may therefore be considered as "don't cares", further reducing the space/time requirements of learning. At least some, and possibly all of these mitigating factors are likely to apply in a wide range of sensory-motor

[6]The worst case storage requirements occur for functions that vary erratically across the entire input space, since this case requires that an exponentially big lookup table be completely filled (storage requirements $O(g^k)$, where $g$ is the number of table gradations per input dimension (defining the "grain" of the approximation), and $k$ is the number of input dimensions.)

nents of the input vector $\mathbf{x}$. Given an $(\mathbf{x}, y)$ pair to be learned, the $\mathbf{x}$ vector is used to index the appropriate table location (a $k$-dimensional box in the input space), and the associated $y$ value is stored there either outright, or combined (e.g. averaged) with a previously stored value. As shown in fig. 3, this process of associating function values with localized regions of the input space is a means for approximating the graph of the function being learned. Learning with Radial Basis Functions (RBF's) is a closely related method (Powell, 1985), recently reviewed and linked to regularization theory by Poggio & Girosi (1990). Rather than operate directly on the input space, input vectors are first "decoded" by a population of "receptive field" units with centers $\xi_i$ that each represents a local, often radially-symmetric, region in the input space. Thus, an output unit computes its activation level as a weighted sum over a set of units with appropriately placed receptive fields, i.e. $y = \sum_i w_i g(x - \xi_i)$, where $g$ defines a "radial basis function", commonly a Gaussian, and $w_i$ is its weight (fig. 4). The learning problem can then be characterized as one of finding weights $\mathbf{w} = \{w_1, \ldots, w_n\}$ that minimize the squared difference between the actual function value and the one produced by the output unit in training, averaged over the set of training examples. Learning schemes of this type lend themselves directly to simple Hebb-type rules for synaptic modification since the initially nonlinear learning problem, i.e. learning the mapping $f$ from $X$ to $Y$, is transformed into a linear one in the unknown parameters $\mathbf{w}$ (see Broomhead & Lowe, 1988). Atkeson (1990) reviews a number of other table-based approaches to learning, and Omohundro (1987) describes a range of efficient computer algorithms and data structures for their implementation, such as "k-d" trees and more recently "ball" trees (Omohundro, 1990). A handful of workers have explored connectionist versions of the locally-smoothing function lookup-table idea, including Albus (1971) and others (Broomhead & Lowe, 1988; Feldman & Ballard, 1982; Ballard, 1987; Lapedes & Farber, 1988; Mel, 1988, 1990; Miller et al., 1988; Moody & Darken, 1989; Walters, 1987).

When may tabular learning be well suited to brains? The classical advantages of tabular learning schemes, neural or otherwise, are learning speed, algorithmic simplicity, and lack of strong representational bias (i.e. *a priori* preference for a specific class of input-output functions). This latter trait is particularly important since a tabular learning system "doesn't care" in whose brain it sits, and to what type of learning problem it is applied. At root, all that is needed for a tabular approach to learning is i) a means for detecting specific regions of the input space, and ii) a means for associating and/or retrieving a value for each such region. In a neural context, these operations may be naturally interpreted as i) establishing a set of appropriately placed, partially overlapping receptive fields in the input space, and ii) attaching a weight to each receptive field given by the desired function value to be associated with that region. Figure 4 depicts a neural "lookup table", encoding a nonlinear function over a two-dimensional input space.

Since these functions are of relatively low underlying dimension, it is possible to gain experience over large portions of the input space through extended practice[5]. Since the functions are smooth, generalization to novel inputs is straightforward (Poggio & Girosi, 1990). The second, very different class of functions that is likely to consume significant cortical resources are the high-dimensional pseudo-random associations that define much of human experience and learning. A retinotopic representation of the visual field has a very large number of underying degrees of freedom (i.e. $d_{\mathcal{X}} \gg 10$)—perhaps thousands to millions. One behaviorally-relevant associative map based on retinal input would be the relation between an image of a face and its auditory name. Since functions of this type are unpredictable in general and of potentially high input dimension, the relevance of smooth generalization to novel inputs (i.e. predicting the name of a new face) is greatly reduced (although it must still be possible to generalize "locally", i.e. to a new face that is similar to a previously learned exemplar). In this type of learning problem, the ability to *discriminate* among large numbers of complex input patterns is of increased behavioral relevance. Our scheme must work well with both of these types of associative learning problems. On the assumption that they are of little biological significance, we do not concern ourselves here with the class of Boolean functions, such as $n$-bit parity, which are highly discontinuous yet require nontrivial global generalization. Functions of this kind have often been used as test cases for connectionist learning algorithms, precisely because they are extremely difficult to learn (Hinton, 1988).

# 4. Two Styles of Network Learning Algorithm: Motivating "Dendritic" Multiplication

We have observed that the problem of intermodal cortical association can be cast, at the single-neuron level, as one of real-valued function learning from examples. We review in the following sections two network learning algorithms relevant to the two types of associative learning problems discussed above, i.e. over low-dimensional continuous input spaces vs. high-dimensional discrete input spaces. For each, the utility of "dendritic" multiplication is highlighted.

## 4.1 Learning Functions with Lookup Tables

We consider first the problem of learning smooth nonlinear functions of low dimension. Perhaps the simplest way to learn a real-valued map $f(\mathbf{x}) : \mathbf{x} \mapsto y$ over a $k$-dimensional input space is to build a lookup table, i.e. to store function values $y$ into a k-dimensional table of numbers indexed by the compo-

[5]This would entail, for the above examples, exercising the arm or the speech articulators through tens or hundreds of thousands of trials or more in order to gain sufficient experience for their intelligent (i.e. predictive) use. Developmental evidence suggests that the learning of voice and limb control do in fact consume many months of practice in human infants (White, 1964).

give rise to "reasonable" $\mathbf{y}$ (for discussion of this issue see Rumelhart et al., 1986 Hinton, 1988; Poggio & Girosi, 1990). In general, $X$ may represent not just one, but several pre-synaptic populations of different modalities that project via modifiable pathways to $Y$. In addition, the teacher input that drives $Y$ may also derive from more than one neural source.

The learning problem may be illustrated from the perspective of a single post-synaptic neuron $y_0$, which, from the above example, we take to be a visual neuron. In this scenario, we assume an animal is flailing its forelimb in its visual field, and that $y_0$ is stimulated by its visual afferent $t_0$ (acting as an unconditioned "teacher" stimulus) whenever the limb happens to move through $y_0$'s visual receptive field. If the intensity of the visual input $t_0$ is plotted over the state space of $X$, i.e. over the space of all forelimb configurations, the problem faced by $y_0$ (and each other visual neuron) can be seen to be one of real-valued function learning from examples. In other words, $y_0$ must learn to predict the intensity of its visual input $t_0$ solely on the basis of the association vector from $X$ which encodes the positions of the joints. The empirical nature of single-unit responses in cerebral cortex suggests that typical unconditioned activation functions will be sparse (i.e. mostly 0-valued), non-negative, and locally smooth[4]. The graph in figure 2 shows a function $f(\mathbf{x})$ with these three properties plotted over a two-dimensional state-space; in relation to the current example, functions of this form were commonly observed in a connectionist learning system that mapped robot-arm joint angles into the visual field activity produced by the arm as it moved in front of a camera (Mel, 1990; see also Kuperstein, 1989). These simple restrictions on functional form will prove to have important consequences for our synaptic learning scheme.

What may the unconditioned activation functions $f(\mathbf{x})$ look like for real neurons in cerebral cortex? The answer to this question is clearly speculative at present, but it is useful to consider two very different kinds of associative learning problems that are likely to be important in cortical processing. We define $\mathcal{X} \in R^d$ to be the underlying $d$-dimensional input space associated with population $X$ (we hereafter denote the dimension by $d_{\mathcal{X}}$). The first class of functions with which we are concerned are smooth, nonlinear functions of relatively low underlying dimension (e.g. for $d_{\mathcal{X}} < 10$). For example, as above $X$ may be devoted to the representation of the joint angles of a forelimb, which, regardless of the number of neurons devoted to its representation, has only a few underlying degrees of freedom. A second example would be the association of a motor representation of the speech articulators to the auditory representation of incoming speech sounds (Rumelhart, 1986).

[4]Sparseness implies that any individual cortical neuron is activated relatively infrequently during learning, e.g. 5-10% of the time; non-negativity derives from the standard frequency coding assumption in which neurons can only have positive activation values; smoothness is a consequence of the assumption that small changes in firing frequencies in $\mathbf{x}$ will generally be associated with small changes in the firing frequency of $y$, due to the smooth coding properties of individual neurons.

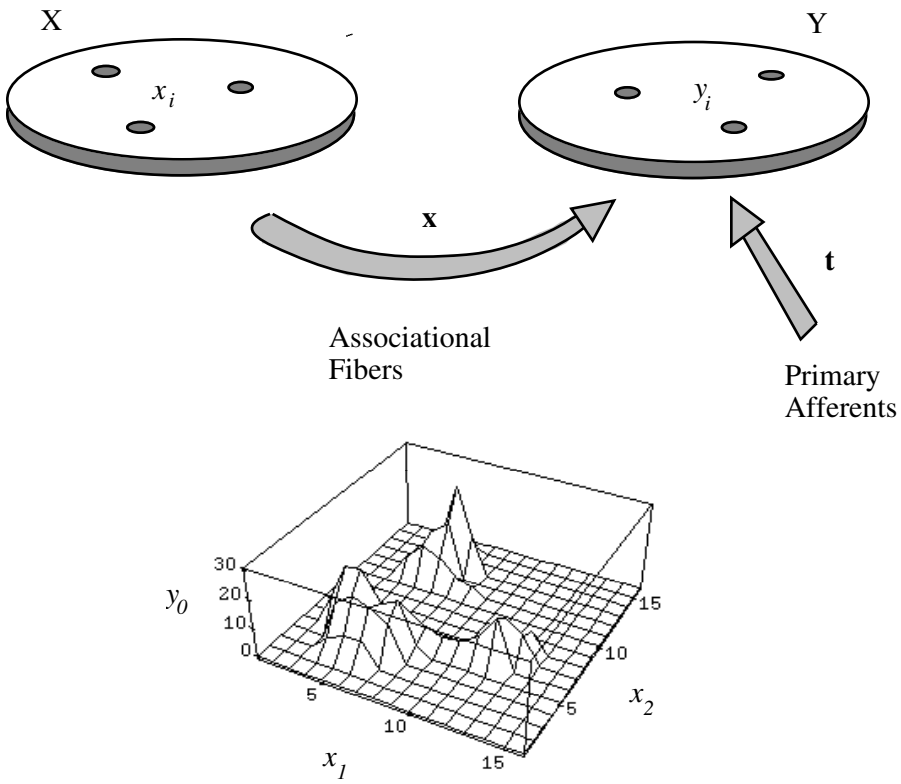


Figure 2: The standard connectionist associative learning problem is mapped into a simplified cortical context. Cortical area(s) $X$ projects to area $Y$ via a plastic associational pathway; during learning population $Y$ is "clamped" by unconditioned teacher input vector $\mathbf{t} = f(\mathbf{x})$ as population $X$ moves through its state space. The task of each individual output unit $y_i$ is to learn it's unconditioned activation function $y_i = f_i(\mathbf{x})$ given the training data consisting of a set of $(\mathbf{x}, y_i)$ pairs. In short, each of the $y_i$ must learn to predict the intensity of its unconditioned activation solely on the basis of the conditioned input vector from $X$. Following the example from the text, $X$ might represent the positions of the joints of a forelimb while $Y$ represents the state of the visual field. The graph below illustrates an unconditioned activation function $t_0 = f_0(\mathbf{x})$ for a single post-synaptic unit $y_0$, which in this case would correspond to the intensity of $y_0$'s unconditioned visual activation as a function of the positions of (two of) the joints. Each unit in $Y$ has a distinct teacher function, though those of neighboring units may in many cases be highly similar. Function learning of this type can be viewed as the process of building up a complicated "receptive field" over an input space.

& Rumelhart, 1989); the potential biological utility of multiplicative dendritic nonlinearities in real neurons has also been pointed out by numerous other authors (e.g. Poggio & Torre, 1978; Koch et al., 1983; Shepherd et al., 1985; Koch & Poggio, 1987; Rall & Segev, 1987, Shepherd & Brayton, 1987). Third, we propose an enhanced single-layer synaptic learning rule called *sigma-pi* learning, in which weights are modified for *groups* of neighboring synapses, rather than for individual synapses, according to a standard Hebb-type learning rule. This general possibility of local interactions in synaptic learning, too, has been discussed elsewhere (e.g. Finkel & Edelman, 1987; Gamble & Koch, 1987; Brown et al., 1988b). Since all three of these assumptions are crucial to the present model, they may serve as predictions that may be tested against the biology.

## 3. Associative Learning in a Cortical Context

In the following, we map the standard connectionist, feedforward supervised associative learning problem (see Rumelhart et al., 1986) onto cerebral cortex, acknowledging that many real complexities are ignored here. A more detailed discussion of the biological basis of this analogy is the subject of section 6.

Consider two cortical populations $X$ and $Y$ (fig. 2), each of which encodes some aspect of an animal's state, e.g. sensory, motor, or internal. The activity states of $X$ and $Y$ are represented by the vectors $\mathbf{x} = \{x_1, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, \ldots, y_m\}$, where the $x_i$ and $y_j$ represent the instaneous activity levels (i.e. average firing rate) of the $i$th and $j$th units in populations $X$ and $Y$ respectively. Population $Y$ receives an "unconditioned" teacher input vector $\mathbf{t}$ that may be thought of as a primary afferent channel[3]. In addition, $X$ projects to $Y$ via a single layer of "conditioned" or "associational" synapses that are modifiable.

During learning, input population $X$ moves through a sequence of states $\mathbf{x}^{(p)}$, $p = 1, \ldots, N$; for each, $Y$ is "clamped" by its teacher input to an "associated" state $\mathbf{y}^{(p)} = \mathbf{t}^{(p)}$. The underlying relation $\mathbf{y} = \mathbf{t} = f(\mathbf{x})$ is assumed to be a function, nonlinear in general, where the form of $f$ is determined by the "semantics" of $X$ and $Y$. For example, $X$ might encode the positions of the joints in a forelimb, while $Y$ encodes the state of the visual field, which depends in part on the configuration of the limb. The learning problem, then, is to build or modify synapses on the association pathway from $X$ to $Y$ such that each trained pattern $\mathbf{x}^{(p)}$ on $X$ gives rise to a good approximation of $\mathbf{y}^{(p)} = f(\mathbf{x}^{(p)})$ on $Y$, typically one that minimizes mean squared error over the training set. The learned map should also generalize such that novel $\mathbf{x}$

[3] A question that has been raised often for supervised connectionist learning systems regards the source of the teacher input that tells the network either i) how specifically to respond to each trained input, or ii) how well it is currently responding to a trained input in comparison to some target. In the cerebral cortex, a wide variety of behavioral state information is simultaneously present in many different areas. Any two (or more) can, interconnections permitting, act as unconditioned "teacher" inputs to the pre- and post-synaptic populations involved in learning.

to the output layer. The rule for synaptic modification no longer depends on information available directly at the synapse, as it does for Hebb-type learning rules, but rather relies on information accumulated from units and synapses across the network as a whole[1]. Extra circuitry must therefore exist for this purpose, or the existing circuitry must act in much more complex ways under global network control. In addition, each model synapse must be able to move freely through both excitatory and inhibitory values; since it is known with relative certainty that biological synapses are not capable of changing sign in this way, it becomes necessary to posit the existence of a network of "sign change" neurons. Unfortunately, this possibility too is unlikely from a biological perspective (Crick & Asanuma, 1986). Yet another possible stumbling block lies in the needed accuracy of the computations that underlie globally-generalizing multilayer learning schemes, in relation to the small number of bits that may be available in single-neuron "arithmetic" hardware.

*Some Updated Assumptions*

Is it, then, a fundamental law that neural associative learning algorithms must be either representationally impoverished or mechanistically overcomplex? We suggest that there is a good resolution to this quandry, achieved by upgrading several basic assumptions of neuronal processing that have run implicitly throughout the above discussion and much of the work discussed therein. First, we place strict representational conditions on the response properties of both pre- and post-synaptic populations participating in the associative learning process, immediately simplifying the task of the synaptic modification rule. Specifically, we require sparsely active populations, and certain styles of "place" coding (also *coarse-coding* or *value-coding*) within these populations (Ballard, 1986) which are known to be ubiquitous in the CNS[2]. Second, we reject the standard summing-unit model neuron (dating to the McCulloch & Pitts neurons (1943), the Adaline (Widrow & Hoff, 1960), and the Perceptron (Rosenblatt, 1962)), in favor of the *sigma-pi* model neuron, which computes its output as a sum of independent multiplicative (or locally-thresholded) clusters of excitatory synapses, distributed throughout its dendritic tree. In so doing we provide a concrete scenario in which *dendritic space*, mostly ignored in current connectionist modeling efforts, is nontrivially utilized as an integral part of the learning computation. For reasons of computational efficacy, model neurons with multiplicative terms (often called "higher-order" units) have been studied by numerous workers (Barron & Barron, 1988; Giles & Maxwell, 1987; Mel, 1988, 1990; Durbin

[1] In the case of back-propagation, this entails that derivatives of the response error with respect to each weight in the network be propagated backwards from outputs toward inputs.

[2] A "coarse-coded" representation consist of units with peaked receptive fields that overlap among neighboring units, and whose peak sensitivities vary systematically from unit to unit across the population.

Hinton & Sejnowski, 1986). The advantage of multiple layers of weights is that the effect of a given input unit on the output can in principle be made an arbitrary nonlinear function (Le Cun, 1985). One example of such a learning scheme is *back-propagation*, a weight-space gradient-descent algorithm designed to minimize response errors over an input-output training set (Rumelhart et al., 1986b; Lecun, 1985; Parker, 1985). For each input-output pair presented to such a network (often called a multi-layered Perceptron or MLP), synaptic weights can in principle be modified throughout the entire multilayer network, such that responses to all subsequent inputs are slightly altered. This property affords the MLP's its power: what is learned from a single isolated input-output example contributes to the development of "hidden" global feature detectors that simplify the learning problem for all inputs (Rumelhart et al., 1986b). The automatic construction of global features that simplify an input-output relation is an extremely exciting possibility, since it provides both a means for reducing the dimensionality of the learning problem (making learning much faster), and good generalizations to novel inputs.

Unfortunately, while the new generation of more powerful neural-network learning schemes have overcome certain limitations of their single-layer predecessors, they have introduced "anatomical" and "physiological" complexities that, coupled with poor scaling behavior and the fundamental problem of local minima, make them highly improbable as biological models (Hinton, 1988). The principal difficulty lies in the nature of the learning algorithm—a gradient-descent search through the parameter space of a high-dimensional nonlinear model. Searches of this kind are well know to be expensive at best, and intractable at worst. Compounding the search problem is the fact that the space of synaptic weights is not fixed, since the number of internal units and layers needed to make learning work well is variable and strongly problem dependent. Moreover, when the relationship between inputs and outputs is complicated (i.e. nonlinear), the synaptic weight space to be searched becomes pocked with local minima that frustrate any "weak method", such as gradient descent. It has been observed often among neural network afficianados that one of the most potent engineering technique in the case of a very difficult learning problem is to build in the best possible features to start, in order that the gradient descent learning process have the best chance of succeeding (Tesauro & Sejnowski, 1989; Denker et al., 1989). Thus careful choice of input and output representations is, in many cases, a far more important ingredient in the success of neural learning than the sheer horsepower of the underlying synaptic modification scheme.

The second drawback of multi-layered synaptic learning schemes when considered as biological models concerns the significant complexities of implementation. In order to build internal feature detectors that help to fit input-output pairs seen during training, it is necessary that each synapse be modified based on its effect on all subsequent synapses and units on the path

abandoned by the theoretical learning community in the early 1970's, due to profound limitations in their ability to learn general-purpose associations among patterns (Minsky & Papert, 1969).

This point is crucial and deserves elaboration. In contrast to the approach of experimental neurobiologists, the thrust of ongoing computational efforts in neural learning has been to explore a range of abstract "neural network" models that relate specific rules for synaptic weight modification to their consequences at the network level, such as their ability to encode patterns and associations of varying complexity and under varying conditions of supervision (for review of concepts see Minsky & Papert, 1969; Rumelhart et al., 1986a; Hinton, 1987). Hebb-type learning rules have been extensively studied in this way (e.g. Palm, 1982; Kohonen, 1984; Sejnowski & Tesauro, 1988), most often for linear or pseudo linear "neuron-like" units, and their scope and limitations are now well understood. They may generally be classified as "single-layer" learning rules, since they only specify changes in a single layer of synaptic weights between inputs and outputs. Single layer learning mechanisms have proven useful for simple supervised pattern association problems (Kohonen, 1985), unsupervised clustering based on statistics of input patterns (Grossberg, 1976; Kohonen, 1985; Rumelhart & Zipser, 1986), and as the basis of certain developmental mechanisms, such as the establishment of connections (Willshaw & von der Marlsburg, 1976) and receptive field properties in visual cortex (Linsker, 1988; Kammen & Yuille, 1988; Miller, et al., 1989; Softky & Kammen, 1990; Moody, 1990). The notorious limitation of single-layer learning rules derives from the fact that they can capture only very simple functional relationships between a given pre-synaptic and post-synaptic unit, i.e. that which may be represented by a simple positive or negative "weight". The exclusive-or (XOR) function is often cited as an example of a very simple function that cannot be learned in this way, but is representative of *all* nonlinear functions in the sense that the direction and degree of influence a given input has on the output is not expressible as a fixed weight, but depends on the values of other inputs (e.g. consider $y = x_1 x_2$; the effect of positive $x_1$ on $y$ can range from excitatory to none to inhibitory depending on the sign of $x_2$, and vice versa). Thus, while single-layer, Hebb-type learning rules are attractively simple from a biological perspective, and are now known to exist in the brain based on solid physiological evidence, they are, under the usual assumptions of the modeling community, provably inadequate for general-purpose associative learning tasks of which the brain is presumed to be capable.

*New Algorithms Are More Powerful, Less Biological*

This limitation, first formalized by Minsky & Papert (1969) for the case of Perceptron learning, has led to the development of more powerful abstract network learning schemes capable of modifying synaptic weights in *multi*-layered networks of pseudo-linear units (e.g. Rumelhart et al., 1986b;

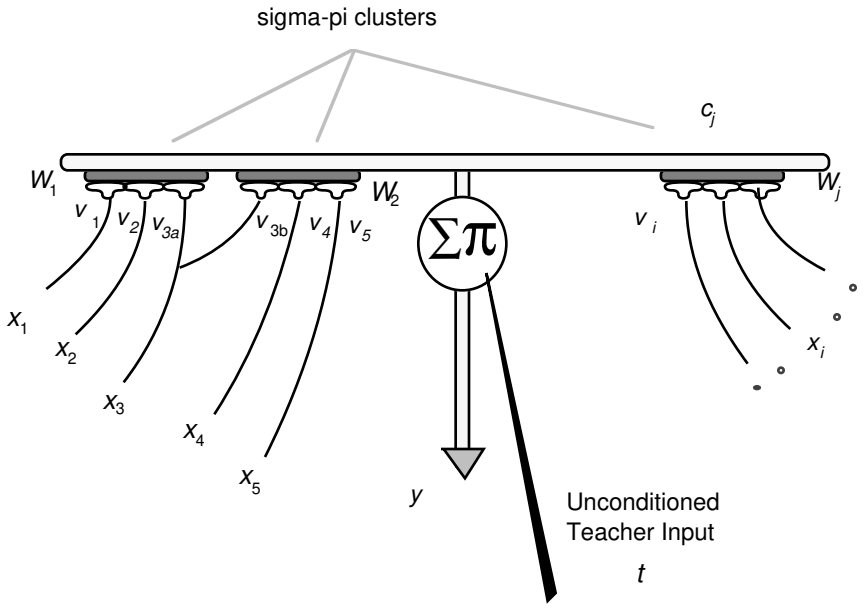


Figure 1: A *sigma-pi* unit computes its level of activation by summing the contributions of a set of independent, multiplicative (or locally thresholded) clusters of input weights. The output $c_j$ of cluster $j$ is computed by multiplying the input signals $x_i$ to that cluster together with their input weights $v_i$, giving $c_j = \prod_i v_i x_i$. The output $y$ of the entire unit is computed as a weighted sum of the outputs of each cluster, giving $y = \sum_j w_j c_j$. Depending on several details of the learning problem, a *sigma-pi* cluster may be viewed as a multi-dimensional receptive field centered on a specific input pattern, or as a multiplicative "higher-order" feature computed over the input space, useful in pattern discrimination problems. See section 5.1 for history of use of this type of model unit.

and are well suited to the broad range of low-dimensional function learning problems likely to crop up in a structure as representationally diverse as cerebral cortex. As abstractions, these learning methods have a long history in statistics and approximation theory (see Atkeson, 1989; Barron & Barron, 1988).

Radial Basis Function methods are essentially similar (see Broomhead & Lowe, 1988) and have recently been discussed by Poggio and Girosi (1989) in relation to regularization theory. Locally-generalizing learning algorithms as neurobiological models date at least to Albus (1971) and Marr (1969); they have also been explored more recently by a number of workers with a more purely computational bent (Broomhead & Lowe, 1988; Lapedes & Farber, 1988; Mel, 1988, 1990; Miller et al., 1988; Moody & Darken, 1989; Poggio & Girosi, 1989). For learning problems of the second class, i.e. involving high-dimensional pseudo-random associations (such as face recognition), *sigma-pi* units can be viewed as standard "higher-order" units that perform Perceptron-style linear discrimination on an input space enriched with higher-order features (Duda & Hart, 1973; Barron & Barron, 1988; Giles & Maxwell, 1987; Durbin & Rumelhart, 1989).

The first part of this paper is devoted to background, motivation, and a description of the *sigma-pi* learning scheme as a computational abstraction. The second part contains a detailed review of the biological evidence for each crucial feature of the model.

## 2. A Historical Quandry

*Hebbian Learning Is Not Enough*

In the past two decades, electrophysiologists have begun to work out the rules that govern synaptic plasticity during learning in the vertebrate central nervous system (Bliss & Lømo, 1973; McNaughton et al., 1978; Baranyi & Feher, 1981; Levy & Desmond, 1985; Kelso et al., 1986; Artola & Singer, 1987; Ito et al., 1989). In confirmation of Donald Hebb's (1949) conjecture, long-term potentiation of synapses that depends of simultaneous pre- and post-synaptic activation has been demonstrated in hippocampus and and other cortical structures (for reviews see Brown et al., 1988a; Kennedy, 1989). Hebb hypothesized that such a rule would lead to the formation of mutually-excitatory "cell assemblies", which he argued could be a means for long-term storage of brain states and/or associations. The term "Hebb-type" will hereafter be used to describe any use-dependent rule for synapse modification that relies only upon locally-available pre- and post-synaptic activity levels (Palm, 1982; Kohonen, 1984; Sejnowski & Tesauro, 1988). While these recent experimental discoveries of Hebb-type learning rules are undisputed milestones in the study of neural learning mechanisms, it is sobering to note that synaptic learning rules of this simple type, as they have been conventionally used in model neural networks (e.g. Rosenblatt, 1962), were all but

# 1. Introduction

Synaptic modification is widely believed to be one of the brain's primary mechanism for long-term information storage. The enormous practical and theoretical importance of synaptic plasticity has stimulated interest among both experimental neuroscientists and neural network modelers, and has provided strong incentive for the development of computational models that can both explain and predict aspects of the neurobiology of learning and memory. With an advancing understanding of structure and function in the central nervous system, the cerebral cortex, consisting of a large number of distinct, interconnected areas that represent many aspects of an animal's behavioral state, has become the generally accepted probable locus for much of the brain's associative learning capacity.

How does this structure, comprising almost 90% of the primate brain, "associate" information from different stimulus modalities? One approach to this problem is to cast cortical associative learning in terms of one or more of the formalisms for neural learning and computation that have been developed in the past 30 years (Anderson & Rosenfeld, 1988). Several questions are thus raised: What are the underlying rules for synaptic modification and synaptogenesis in cerebral cortex? What is the proper computational abstraction for the single cortical output neuron—simple threshold summing unit, such as Rosenblatt's Perceptron (1962), or complex logic processor, as discussed in possibility in (Koch & Poggio, 1987; Shepherd & Brayton, 1987)? With an eye to the specific details or cortical anatomy, we may further ask how the asymmetry in both number and pattern of termination of excitatory vs. inhibitory synapses onto cortical neurons are to be accounted for within the learning scheme? How, if at all, is the modular anatomical structure of cerebral cortex relevant to associative learning?

We present here a model for associative learning in cerebral cortex. Our goal has been to build a coherent framework that assigns computational significance to an array of features of cortical anatomy, physiology, and biophysics that are well documented, but that remain disconnected and/or poorly understood in their relation to learning and to each other. The main hypothesis of this work is that the output neurons of a cortical association area can be modeled as *sigma-pi* units, which compute their activation levels as a sum of contributions from a set of independent multiplicative clusters of synapses scattered throughout their dendritic trees (fig. 1). Their attractiveness from a modeler's perspective stems from the fact that a *sigma-pi* unit can, with only a single layer of excitatory synapses, learn both smooth nonlinear functions of low-dimension, exemplary of an important class of sensory-motor learning problems, as well as the irregular high-dimensional associative maps typical in general mnemonic function (e.g. the association of names with faces). In the first case, *sigma-pi* units learn functions of their inputs as locally-generalizing lookup tables. Table-based learning algorithms both lend themselves to straightforward biological network implementation,

# The Sigma-Pi Column: A Model of Associative Learning in Cerebral Neocortex

**Bartlett W. Mel**
*Computation and Neural Systems Program, 216-76*
*California Institute of Technology*
*Pasadena, California 91125*
*(818)356-3643, mel@aurel.caltech.edu*

April 30, 1990

CNS Memo 6

# Abstract

In this paper we present a model of associative learning in cerebral neocortex. The extrinsically-projecting pyramidal cells of layers 2, 3, and 5 of association cortex are modeled as *sigma-pi* units, where a *sigma-pi* unit computes its activation level as a sum of contributions from a set of multiplicative (or locally-thresholded) clusters of synapses distributed throughout its dendritic tree. The model demonstrates how a broad class of biologically-relevant nonlinear associative learning problems can be solved in this system by modifying only a single layer of excitatory synapses under the control of a Hebb-type learning rule. The model also accounts for a variety of features of cortical anatomy, physiology, and biophysics whose relations to learning have remained poorly understood. These include, (1) three learning-related roles for the NMDA channel, including one new, (2) the gross asymmetry in number and patterns of termination of excitatory vs. inhibitory synapses onto cortical pyramidal cells, as well as the apparent lack of plasticity at inhibitory synapses, (3) the replication of like-activated neurons beneath a single point in cerebral cortex, and in particular the clumping of apical dendrites of pyramidal cells on their rise to the cortical surface, (4) the complex 3-dimensional arborizations of axons and dendrites in layer 1, which give rise to a rich "combinatorial" association interface crucial to the current model, and (5) putative rules for activity-dependent axon growth and synaptogenesis during associative learning.