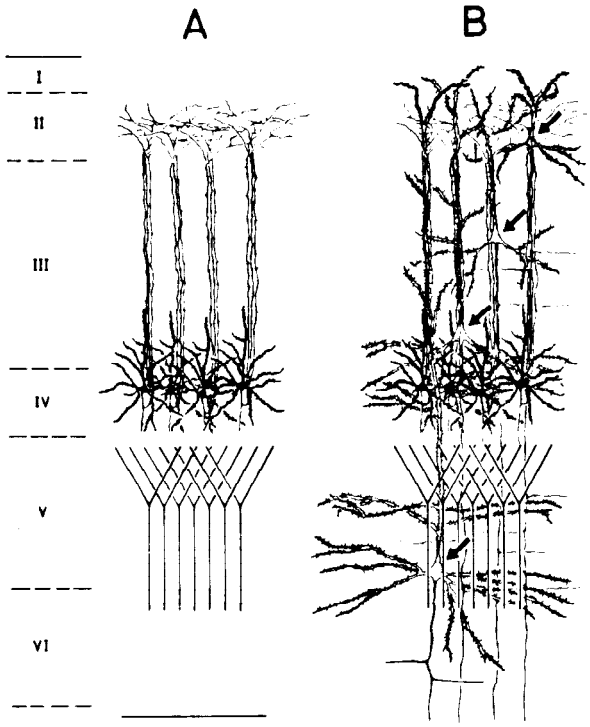


Figure 8: Evidence for an unconditioned "teacher" input pathway with priveledged access to modules of cortical pyramidal cells. One class of small spiny cells which receive thalamocortical input send their axons vertically to enclose, as if in a sleeve, the clumped apical dendrites of a group of pyramidal cells. From (Jones, 1981).

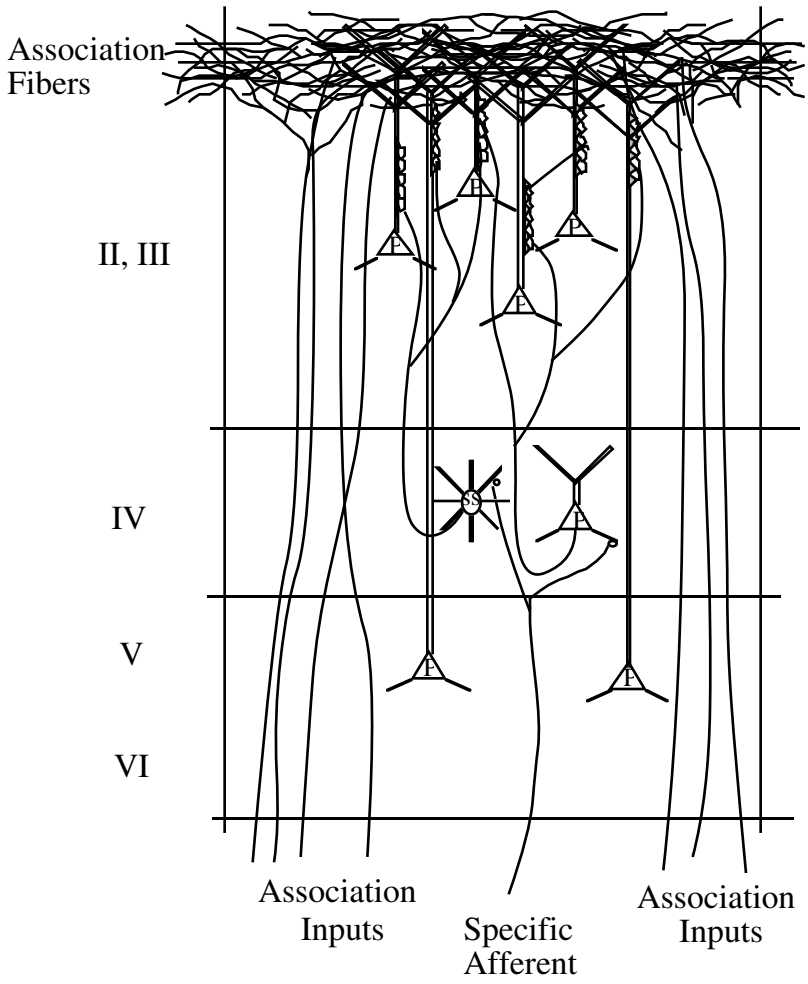


Figure 7: Schematized view of the cellular elements in a generic association cortex. Each "module" of layer 2, 3, and 5 pyramidal cells whose apical dendrites are spatialy clumped on their rise toward the cortical surface (Peters, 1985) are proposed to form a "*sigma-pi* column" (only one module/column is shown). The module is driven by two distinct types of input: Inputs terminating onto granule cells in the middle layer(s) of cortex are presumed to contribute relatively powerful, relatively focal excitation to the module. The long-range excitatory association fibers that climb to layer 1 (though they may for synapses in other layers as well) and arborize amongst the apical dendrites of pyramidal cells are assumed to contribute the conditioned, or associational input to the module.

and known or suspected intracortical circuits are ignored here, such as all types of inhibitory interneurous and their connections, and all locally directed excitatory feedback pathways. Since it is currently impossible to cope at once with the full complexity of cortical cytoarchitecture, we proceed instead with a directed search for each of the six essential elements of the *sigma-pi* learning abstraction. These are, i) an appropriate population of output neurons, ii) an unconditioned teacher input, iii), a diffuse association input, iv) Hebb-type synaptic plasticity, v) local dendritic multiplication (or thresholding), and vi) a cluster reservoir. We conclude, surprisingly, that only a very simple subset of known cortical circuitry can in principle underlie a very general associative learning capacity.

6.1 Gross Anatomy

We concern ourselves here with the cytoarchitecture of "generic" association cortex, and propose the cortical circuit of fig. 7 to contain all of the basic elements necessary for associative learning (closely paralleling the accounts of Marr (1970) and Eccles (1981) at this level of description). We limit our focus to the "output" pyramids of layers II, III, and V, which are posited to be *sigma-pi* units. These cells are a likely locus of associative learning as they are well situated to receive both teacher and associational input pathways.

6.1.1 The Teacher Input

A prominent feature of the micro-anatomy of cerebral cortex is the tight columnar aggregation of the shafts of apical dendrites of layer 2, 3, and 5 pyramidal cells as they rise to the cortical surface (Jones, 1981; Peters, 1987, 1989), where they ultimately form a "dense dendritic mat" in layer 1 (Feldman, 1984). These clumps of apical dendrites occur on 50-60 μm centers in the rat (30 μm in the monkey, Alan Peters, personal communication), and derive from approximately 50 layer 5 and 90 layer 2/3 pyramidal cells (Peters, 1989). These bundles of apical dendrites seem to be specifically designed to receive a powerful common input, and in fact are enclosed, as if in a "sleeve" (Jones, 1981), by the axons of the excitatory granule cells of layer 4, which, roughly speaking, consist in primary sensory areas of spiny stellate cells and elsewhere in morphologically similar small pyramidal cells with similar vertically-oriented axons (Lund, 1984). It thus appears plausible that afferent terminations onto layer 4 granule cells, which may be of both thalamo-cortical (Herkenham, 1986) and cortico-cortical origin (Van Essen, 1985), contribute to a powerful, relatively focal unconditioned "teacher" input pathway acting on distinct modules of cortical output neurons (fig. 8). The common physiological responses of neurons in a narrow vertical column (Mountcastle, 1978) supports the idea that a tight, excitatory columnar integration process, beginning with the afferent terminations in layer 4, exists as a general feature of cerebral cortex (Jones, 1975, 1981).

6. Toward A Neurobiological Foundation

Given its sheer size and importance to higher vertebrates, cerebral cortex is in all likelihood the locus of much of the brain's associative learning capacity. It is therefore of great theoretical and practical importance to develop models for the synaptic basis of cortical associative learning. Such models should minimally explain how a nonlinear map can be learned in cerebral cortex, in a way that is consistent with known constraints of cortical anatomy and physiology, and neuronal biophysics.

The *sigma-pi* learning scheme has *a priori* biological appeal from several standpoints. It is fast, requiring on the order of a single pass through a training set for learning to occur (though training sets can grow to be large—see footnote 5). Second, it can account for the learning of nonlinear functions with only a single layer of plastic feedforward synapses between input and output units—the algorithmic and structural complexities associated with multi-layered learning schemes are thus avoided. Third, *sigma-pi* units are modular, such that their usefulness can be directly multiplied by simply using more of them. This feature would seem to be extremely advantageous for a biological learning system, for which simple replication of structure during development is likely to be a relatively simple genetic adjustment.

In addition to these *a priori* advantages, the algorithm has several additional features that are highly suggestive from a neurobiological perspective. For example, the *sigma-pi* learning scheme deals strictly with excitatory synapses, which, while leaving the possible functions of intracortical inhibition entirely unspecified within the model, is at least consistent with anatomical and physiological data in two important respects: (1) Excitatory synapses outnumber inhibitory by a factor of between 5 and 8 in cerebral cortex (Beaulieu & Colonnier, 1985). There is likewise a pronounced asymmetry in the physical distribution of excitatory and inhibitory synapses on cortical neurons, strongly suggesting that inhibition and excitation do not play mirror image roles in synaptic learning. (2) In keeping with this possibility, while there is a growing body of direct evidence of Hebb-type plasticity in excitatory synaptic pathways (see Kennedy, 1989), there is simultaneously a growing *lack* of evidence for plasticity at inhibitory synapses (Singer, 1977; Bear et al., 1985; Mower et al., 1987). The *sigma-pi* model is thus consistent with the observed asymmetries in number, disposition, and plasticity of excitatory vs. inhibitory synapses in cerebral cortex (a similar point is made by Bear & Cooper, 1990). A further immediate suggestion of biological relevance lies in the known existence of active, i.e. voltage-sensitive, membrane channels in cortical dendrites, such as that mediated by the NMDA-type glutamate receptor (see Cotman et al, 1988), potentially capable of carrying out the local multiplicative interaction upon which the *sigma-pi* learning scheme depends (e.g. Koch, 1987).

In the remainder of this paper we examine the hypothesis that *sigma-pi* units underlie associative learning in cerebral cortex. Numerous cell types

training set, while minimizing total synaptic weight for all clusters weighted by their average levels of activation[8]. Note that this learning rule calls for a *weakening* of synapses, corresponding roughly to the condition of "pre-without-postsynaptic" activation (governed by the second term of equation 5.3). The theoretical motivation for this weakening condition is most easily seen in the lookup-table analogy for function approximation. The weight given to each cluster receptive field should be proportional to the expected intensity of the teacher signal *when in that cluster's receptive field.* To this end, the decay term in equation 5.3 is proportional to the cluster's activity c_j, insuring that those clusters are relatively "punished" that fire too often without reinforcement from the post-synaptic teacher signal (i.e. for whom the expected intensity of teacher reinforcement is low). This can occur if the cluster receptive field is too large and non-selective, such as a cluster that *always* fires, independently of the teacher signal, or for clusters that are poorly correlated with the teacher signal. Note that the *other* phase of Hebbian anti-correlation, i.e. post-without-presynaptic activation, does *not* result in synaptic weakening according to 5.3. This is because the weight on a receptive field is only "relevant", and should only be modified, when the system is "in" that receptive field, i.e. when $c_j > 0$. We touch upon this Hebb-rule asymmetry again in section 6.2.2. We also briefly note that because this Hebb-type learning rule is truly local, i.e. depends only upon activity levels available directly at a synapse to be modified, it may be applied transparently to a group of neurons driven by the same global teacher input (see above discussion of *sigma-pi* modularity). Error-correcting rules that modify synapses based on a difference between desired vs. actual neural output do not share this property, since they require that the effects of all contributing inputs, putatively scattered across a group of neurons, be integrated in order to generate an error signal for learning.

[8] More formally, the learning rule can be seen to minimize the cost functional,

$$E = -\sum_p (\alpha\, t^{(p)} y^{(p)} - \frac{\beta}{2} \sum_j c_j^{(p)} w_j^{(p)\,2}),$$

where the first term grows as the product of the desired and actual output of the unit, and the second, punitive, term tends to limit the total weight allocated to all clusters, where a larger cost is accrued for clusters that are more often active.

that each cluster input x_i represents a gaussian receptive field in the underlying input space of dimension $\delta_i, 1 < \delta_i < \delta_{max}$, and, for reasons discussed in section 6.2.3, we assume a hardware bound on cluster size $k = k_{max}$. A *sigma-pi* cluster can therefore represent a "receptive field" of dimension at most

$$d_\Pi = \sum_{i=1}^{k} \delta_i,$$

such that the maximum dimensionality attainable for any single cluster receptive field is given by $d_{\Pi max} = k_{max} \cdot \delta_{max}$. In this way, receptive fields of between 6 and 9 dimensions are easily attainable with small values of δ_{max} and k_{max}. When $d_{\Pi max} \geq d_{\mathcal{X}}$, a *sigma-pi* unit can directly implement a RBF decomposition of the function to be learned (see section 4.1, fig. 4), since full-rank receptive fields of dimension $d_{\mathcal{X}}$ can be constructed within the constraints of maximum cluster size and maximum input-line receptive field dimension. When $d_{\Pi max} \ll d_{\mathcal{X}}$, then a *sigma-pi* cluster "degenerates" to a single conjunctive higher-order features in support of a high-dimensional pattern discrimination problem in $\mathcal{X}$. This is to say that as a single cluster can no longer by itself recognize a specific input vector (as is possible in the low-dimensional case), it acts only as a source of evidence for post-synaptic activity, in concert with a large number of other clusters (see use of "codons", Marr, 1970).

5.2 The Learning Rule

Initially, we assume all possible clusters of size $1 < k < k_{max}$ pre-exist with random (or 0) weight. The cluster weights w_j are modified during training according to the following self-normalizing Hebb rule:

$$\dot{w}_j = \alpha c_j t_j - \beta c_j w_j, \tag{5.3}$$

where α and β are small positive constants, c_j is the cluster response, and t_j is the teacher input to cluster j. The inputs weights v_i are unmodified during learning, representing the degree of cluster membership for each input line (this simplification is reconsidered in section 6.2.3). The steady state of this learning rule occurs when

$$w_j^{ss} = \frac{\alpha}{\beta} \frac{< c_j t >}{< c_j >}, \tag{5.4}$$

i.e. when w_j converges, very roughly speaking, to the expected value of the teacher signal within the cluster's receptive field, normalized by the size of the receptive field. Equation 5.3 includes both a standard Hebb-type reward term as well as a decay term, and tries to maximize the correlation[7] of pre-synaptic cluster activation and post-synaptic teacher signal over the

[7] Strictly speaking, the average product.

5. Sigma-Pi Learning

5.1 The Sigma-Pi Unit

Unlike the classic thresholded linear unit that is the mainstay of many current connectionist models, the output of a *sigma-pi* unit is computed as a sum of contributions from a set of independent multiplicative clusters of input weights:

$$y = \sum_j w_j c_j, \tag{5.1}$$

where

$$c_j = \prod_{i=1}^{k} v_i x_i \tag{5.2}$$

is the product of k inputs x_i with their weights v_i within cluster j, and w_j is the weight on cluster j as a whole. During learning, the output may be clamped by an unconditioned teacher input, $y = t$ (fig. 1). It is important to note that equation 5.1 is inherently *modular*, in the sense that a group of *sigma-pi* units can, by simply adding their outputs together, act as a single much larger virtual *sigma-pi* unit with correspondingly increased cluster storage capacity. A neural architecture that allows system storage capacity to be multiplied by a factor of k by growing k neurons in the place of one, is one that should be strongly preferred by biological evolution.

Units of this general type were first proposed by Feldman & Ballard (1982), and have been used occasionally by other connectionist modelers, most commonly to allow certain inputs to gate others or to allow the activation of one unit to control the strength of interconnection between two other units (Rumelhart & McClelland, 1986). The use of *sigma-pi* units as function lookup tables was suggested by Feldman & Ballard (1982), who cited a possible relevance to local dendritic interactions among synaptic inputs. Volper & Hampson have also discussed the use of related OR-of-AND neurons (1987) as well as neurons with polynomial inputs (1990) for the learning of Boolean functions from examples, and have made certain ties to properties of biological learning. *Sigma-pi* units are also often described as "higher order" units, since the multiplicative operation can be used to implement global second and higher-order polynomial relations among a set of inputs (Giles, 1987). Durbin & Rumelhart (1989) have recently used related "product" units in an extension of the back-propagation algorithm to learn arbitrary polynomial functions over the input space; they too suggest possible relevance of their multiplicative nonlinearity to local processing operations in dendritic trees.

In the present work, the specific nonlinear interaction among inputs to a *sigma-pi* cluster is not of primary theoretical importance. The crucial property of a cluster is that its output should be AND-like, i.e. selective for the simultaneous activity of *all* of its k input lines. Following our previous discussion of low-dimensional value-coding in cerebral cortical neurons, we assume

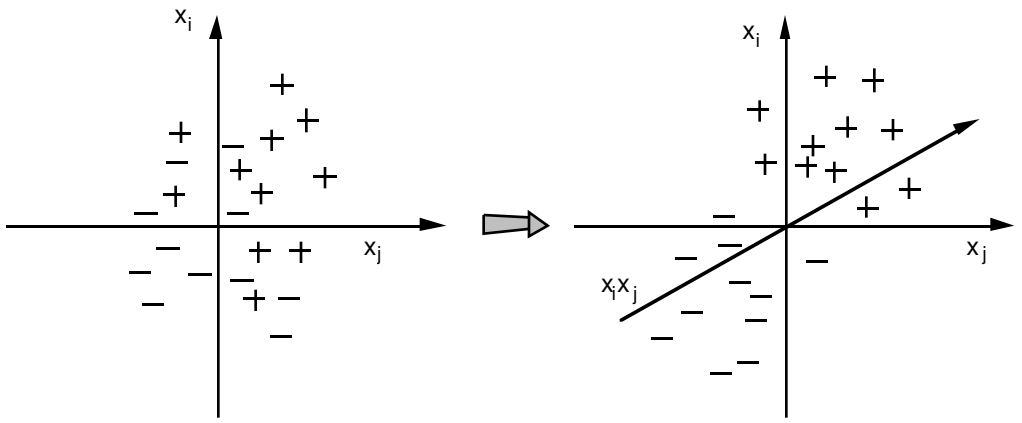
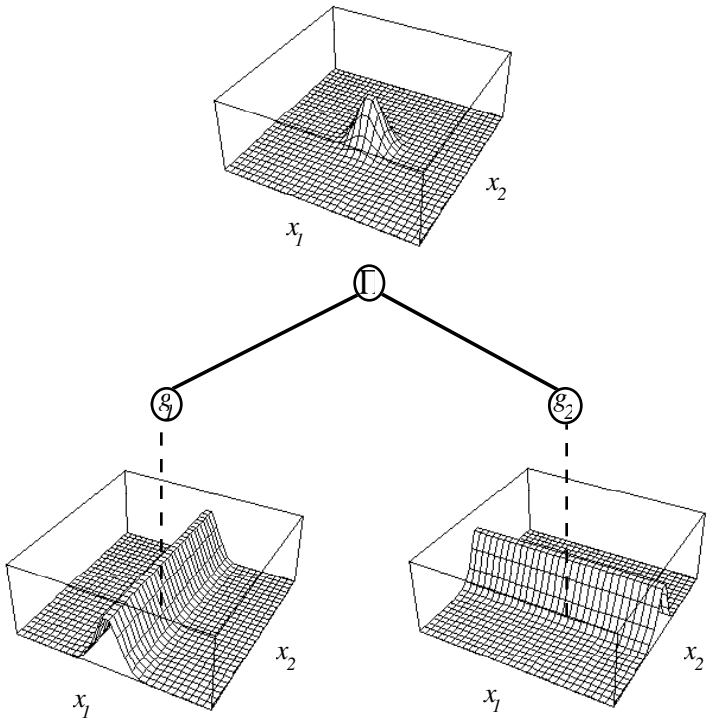


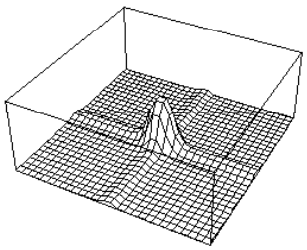
Figure 6: An input space in which pattern classes are easily confusable (left) can be "orthogonalized" by mapping the input patterns into a space of higher dimension in which the patterns are more easily dichotomized. One method of increasing the dimensionality of the input space is with polynomial (i.e. multiplicative) combinations of the original input space parameters.

rithm discussed above, this would entail that each output unit activated in the presence of a particular input image construct a multiplicative "receptive field" selective simultaneously for all of the thousands of features that define that particular image. This approach to learning would be highly inefficient and unbiological. In cases such as these, a standard computational technique involves mapping the original input patterns into a space of even higher dimension in which they become more easily discriminable (Duda & Hart, 1973). One way to construct a new high-dimensional feature space is to add to the original set of elemental features a set of small multiplicative conjunctions (i.e. polynomial terms) of the elemental features (Marr, 1970; Ballard, 1986, 1987; Giles, 1987; Barron & Barron, 1988; Durbin & Rumelhart, 1989; Volper & Hampson, 1990). The addition of such higher-order features, which are more "selective" than the elemental set (i.e. they "fire" much less often and to a much smaller set of input patterns), is a powerful technique for "orthogonalizing" the set of input patterns (Kohonen, 1985), i.e. making them look less similar to each other than before (fig. 6).

In summary, the act of "decoding" an input space, by creating multiplicative conjuncts of input features, facilitates learning for two classes of learning problems at opposite ends of the spectrum of continuity and dimensionality. In the first case, multiplication is useful for the establishment of localized receptive fields in low-dimensional learning problems where smooth interpolation is crucial. In the second case, multiplication is useful for the creation of higher-order features that increase the discriminability of input patterns. This utility of multiplication as an elemental "dendritic" computation has motivated the *sigma-pi* model neuron discussed below.



(a)



(b)

Figure 5: (a) A 2-dimensional Gaussian receptive field may be constructed by multiplying together the outputs of two units with 1-dimensional Gaussian receptive fields. Multidimensional neural response selectivities can be synthesized in this way. For example, a unit simultaneously selective for the positions of two joints can be gotten by multiplying together the outputs of two units, each of which is selective for the position of only one of the joints. (b) The multiplication operation from (a) can be approximated by a sum followed by a soft threshold.

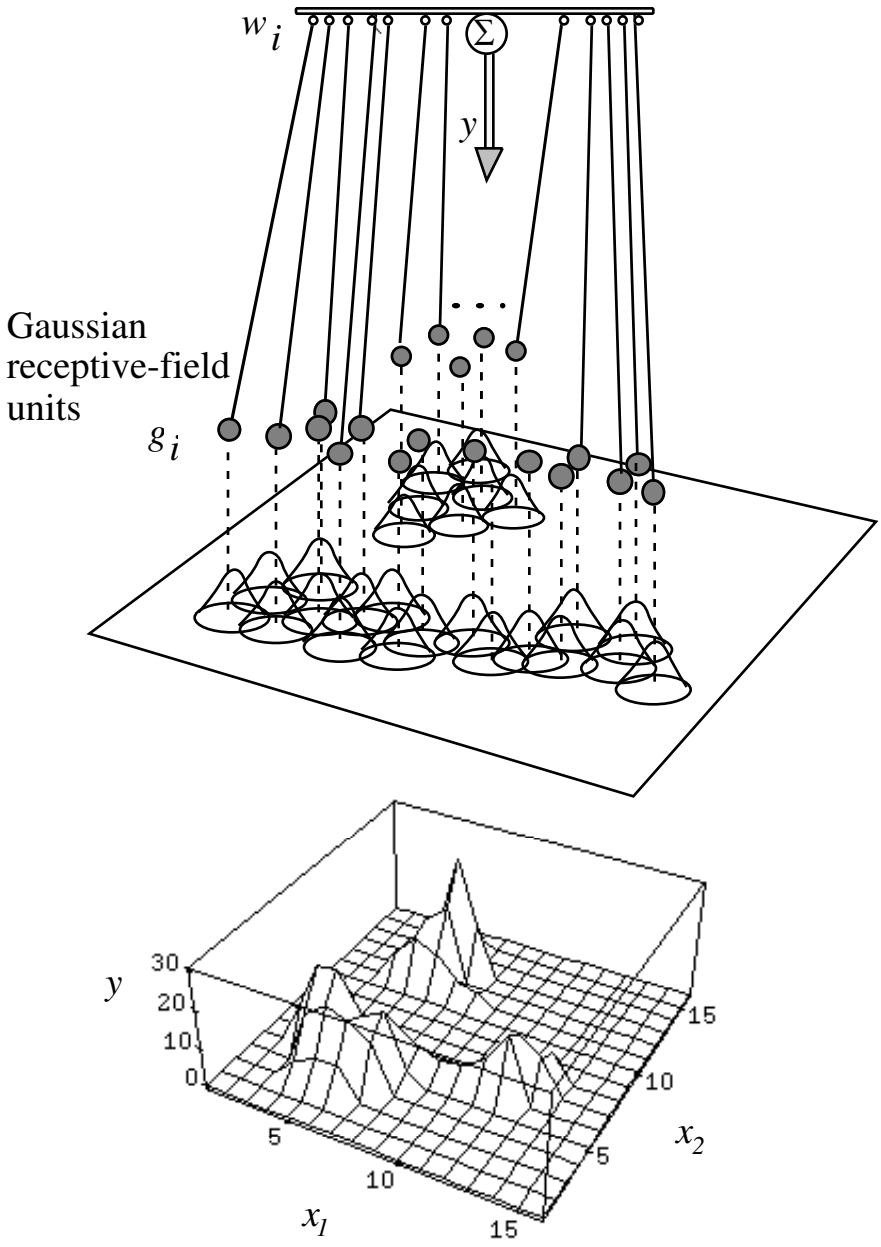


Figure 4: A neural lookup table. A linear output neuron (at top) approximates the graph of a nonlinear function by summing the weighted outputs of a set of localized receptive field units distributed across the 2-dimensional input space. Where the function is 0-valued, no receptive field units are needed. Also, fewer (broader) receptive field units are needed, in general, when the function to be learned is relatively smooth.