Network 1 (1990) 407–421. Printed in the UK

# The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain

Edmund T Rolls and Alessandro Treves

University of Oxford, Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, UK

Received 19 July 1990

**Abstract.** In some neuronal networks in the brain which are thought to operate as associative memories, a sparse coding of information can enhance the storage capacity. The extent to which this statement is valid in general is discussed here, by considering some simple formal models of associative memory which include different neurobiological constraints. In nets of linear neurons, trained with either a Hebbian (purely incremental) or a Stanton and Sejnowski learning rule, sparse coding increases the number of independent associations that can be stored. When neurons are nonlinear, for a diversity of learning rules, sparse coding may result in an increase in the number of patterns that can be discriminated. The analysis is then used to help interpret recent evidence on the encoding of information in the taste and visual systems, as obtained from recordings in primates. Following the taste pathway, it is found that the breadth of tuning of individual neurons becomes progressively finer, consistent with the idea that sparser representations become advantageous as the taste information is eventually associated with that coming from other sensory modalities. In the visual system, considering a population of neurons in the temporal cortex that respond preferentially to faces, it is argued that their breadth of tuning represents a compromise between fully distributed encoding, and a grandmother cell type of encoding, which would result in a given neuron responding only to an individual face.

## 1. Introduction

It has been suggested that the encoding of information towards the end of sensory processing systems in the cerebral cortex is a delicate compromise between very fine tuning, which has the advantage of low interference in associative neuronal network operations but the disadvantage of losing the emergent properties of storage in such neuronal networks, and broad tuning, which has the advantage of allowing the emergent properties of neuronal networks to be realized but the disadvantage of leading to interference between the different memories stored in an associative network (Rolls 1987, 1989a). There is reasonable support that the response of neurons in the taste and visual systems does become tuned in this way before it is interfaced to associative memories in structures such as the amygdala and orbitofrontal cortex, and autoassociative memories in structures such as the hippocampus (Rolls 1987, 1989a,b,c, 1990b,d). The purpose of this paper is to consider in more detail the exact conditions under which sparse representations do have advantages for information storage in neuronal networks of types which might be implemented in the brain, and why these advantages occur.

A remark should be made immediately concerning the above notions. A sparse representation, or sparse coding, of a signal is one in which only a small fraction of the neurons in a network is activated by a given stimulus. In the context of single-cell

0954-898X/90/040407+15$03.50  

recording, one often uses the concept of fine tuning, which refers to a given neuron being activated by only a small proportion of the stimuli belonging to a certain set. If the set of stimuli considered, on the whole, activates neurons distributed evenly over the network, the two concepts can be taken as equivalent, and they will be used as such in this paper.

The arguments presented here are intentionally kept general so that they apply to many types of (feedforward) associative network, and do not involve fully specifying the details of particular network models. While particular models describing some similar systems have been analysed extensively in the literature (e.g. by Nadal and Toulouse 1990), relatively little attention has been paid to the implications of a few simple constraints on the operation of associative neuronal networks in the brain.

## 2. Associative neuronal nets with linear neurons

The statement that relatively finely tuned neurons, which imply a sparse representation, might be advantageous for inputs to neuronal networks in the brain came from a consideration of storage in linear associative networks using a Hebb rule. Such networks have been studied, for example, by Kohonen and his colleagues (Kohonen *et al* 1981). A set of axons from a population of pyramidal cells makes modifiable synaptic contacts onto a set of dendrites belonging to cells of a later processing stage. Each dendritic tree, receiving inputs from $N$ axons, is taken to compute a linear summation of post-synaptic potentials (PSPs):

$$h_i = \sum_k J_{ik} V_k \tag{1}$$

where $h_i$ represents a term in the membrane potential of cell $i$, and each PSP is written as the product of a synaptic efficacy $J_{ik}$ times the firing rate $V_k$ of axon $k$ ($k = 1, \ldots, N$). The full expression for the membrane potential might include other terms, such as a threshold $h_T$, and PSPs mediated by non-modifiable synapses (or non-modifiable components of synapses), collectively denoted as $h_i^0$. The single cell input–output function is assumed linear for this class of model, at least in some normal 'operating' range for $h_i$, i.e. the firing rate of cell $i$ is simply proportional, in this range, to $(h_i + h_i^0 - h_T)$. Note that this implies that the cell is normally operating above threshold.

This linear behaviour makes the response of the output cells to an arbitrary signal predictable, once one knows their response to $N$ linearly independent input patterns. For example, the response ($R$) to the presentation in the input of the pattern $V = c_1 V^1 + c_2 V^2$ will be just $R = c_1 R^1 + c_2 R^2$ (linear generalization). While the input patterns are embedded in a space of dimension $N$, the effective dimensionality they really span, as evidenced for example by factor analysis, might be smaller, say $p$. A conceivable requirement for a network that learns input–output associations, then, is that for each effective dimension there is a prototype input pattern, whose association to a particular output pattern is learned independently of other associations. One may now consider the constraints that the mechanisms subserving learning pose on the capacity $p$ thus defined.

### 2.1. *Hebb (incrementing) learning rule*

The modifiable synaptic efficacies are taken to encode $p$ learned associations of axonal firing patterns with dendritic activation patterns, according to a Hebb rule:

$$J_{ik} = \text{constant} \times \sum_m (h_i^m + h_i^0) V_k^m \tag{2}$$

where $m = 1, \ldots, p$. It should be noted that each learnt association results in an increment in $J_{ik}$, which in this particular case is simply linear both in the presynaptic firing rate $V_k^m$ and in the post-synaptic integrated potential $h_i^m + h_i^0$.

In order for the network to retrieve a particular learnt pattern (say, $m = 1$) upon presentation of the corresponding axonal pattern, to avoid any crosstalk from other stored associations one should have

$$\sum_k V_k^m V_k^1 = 0 \text{ for each } m > 1 \tag{3}$$

i.e. the axonal firing patterns should be orthogonal to each other. If the $V$ were to assume arbitrary real values (i.e. negative as well as positive), one could find up to $N$ mutually orthogonal vectors of length $N$. As each component $V_k$, however, in the brain represents a firing rate (typically in the range between 0 and 100 spikes $s^{-1}$), it is constrained to be zero or positive, and this fact sets a limit (which is much below $N$) on the maximum number $p$ of independent associations stored in this type of network. In order to keep this number high it becomes necessary to have rather sparse encoding, to ensure that the different pattern vectors are relatively orthogonal to each other. Thus, if only $Na$ axons are activated (i.e. fire with non-zero rate) in any given pattern, where $a$ parametrizes the sparseness of the coding scheme, the maximum $p$ to strictly avoid crosstalk between memories become $1/a$, corresponding to the simplest situation in which different patterns activate non-overlapping sets of input fibres. The suggestion was thus that to exploit the capacity of the matrix of connections, with minimal interference between the stored patterns, the coding would be very sparse, $a \ll 1$. The advantage for having some distribution in the encoding, that is not with just one 'grandmother' neuron representing the input to the associative net ($a = 1/N$), was accounted for (Rolls 1987, 1989a,c) on the grounds that this enables the 'emergent' properties of distributed systems to be realized (Kohonen *et al* 1981, Rolls 1987, 1989a,c). These properties include robustness of performance with respect to the presence of noise in the input (which can be expressed as an ability to generalize and to complete partial information) and with respect to moderate lesions in the network (graceful degradation).

### *2.2. Learning rules modified in the presynaptic factor*

The above argument does not apply if the form of the synaptic learning rule is modified, so that the summed postsynaptic potential $h_i$ does not depend only on the scalar products expressing the correlation between the current axonal firing pattern and the stored firing patterns as set out in expressions (2) and (3) (see e.g. Brown *et al* 1990). A possible modification would be to substitute expression (2) with one like

$$J_{ik} = \text{constant} \times \sum_m (h_i^m + h_i^0)(V_k^m - v) \tag{4}$$

where the efficacy $J_{ik}$ increases if there is postsynaptic activity and the presynaptic activity is above a value $v$, and decreases if there is postsynaptic activity and the presynaptic activity is below the value $v$. At this stage $v$ is an arbitrary constant, possibly different for each presynaptic input. The net result is that when pattern $m$ is learned and depending on the degree of postsynaptic activation, $J_{ik}$ increases if $V_k^m > v$, and decreases otherwise. (A rule of this type is sometimes known as a Singer–Stent rule (Singer 1987, Stent 1973), and accounts for many findings on the plasticity of the visual system during development (Singer 1987) and on long-term potentiation in the hippocampus (Levy

1985, Levy and Desmond 1985).) In this case to avoid crosstalk upon presentation of pattern 1 one would require

$$\sum_k V_k^m V_k^1 = v \sum_k V_k^1 \tag{5}$$

and it becomes again possible to find $N$ vectors that satisfy these conditions, provided now that the average activity of these vectors is of the order of $v$ (see also Willshaw and Dayan, 1990). In other words, the network can make use of its potential capacity for independently storing associations if some mechanism ensures that $v$ is held close to the average firing activity of the presynaptic axon. In this particular case there appears to be no special advantage to sparse representations.

Expression (4) is closer to the 'covariance' learning rule postulated in many theoretical models (Sejnowski 1977). A covariance rule would involve subtracting from both the pre- and the postsynaptic factor in $J_{ik}$ their respective average values over a prescribed set of pairings. As concerns the capacity of feedforward networks of linear neurons, however, it is not crucial whether there is or there is not a subtractive term included in the postsynaptic factor (for the reasons described above). It should be noted that such a postsynaptic subtraction would imply a decrease in synaptic efficacy whenever strong presynaptic firing is paired with weak postsynaptic activation. Evidence for a mechanism of long-term depression which could be expressed by a learning rule of this type has been claimed recently by Stanton and Sejnowski (1989).

An alternative modification would be to leave expression (2) as it stands, and write expression (1) as

$$h_i = \sum_k J_{ik}(V_k - v) \tag{6}$$

where the subtracted term $\Sigma_k J_{ik} v$ could represent, for example, some form of inhibition which depends on the modified synaptic strengths and therefore could be produced by feedback inhibition, or else a modulation of the threshold of cell $i$ tuned to the total modification over all the $N$ synapses on each output neuron. Although specifying the way in which this inhibitory effect or this threshold modulation could be implemented in practice in the brain requires additional assumptions, whose validity is yet to be tested, mechanisms of this type have been hypothesized by modellers since the time of Marr (1970, 1971). It should be noted in particular that for a large net the sum $\Sigma_k J_{ik} v$ may be substituted with an average over the patterns, leading to a subtraction mechanism which need not reflect synaptic modification and therefore might be much more easily realized in the brain.

### 2.3. *Learning rules of a type which could be implemented by NMDA receptors*

A major model of neural mechanisms involved in learning in the brain is provided by long-term potentiation. In some brain systems studied intensively, such as parts of the hippocampus, long-term potentiation occurs when the postsynaptic membrane becomes so strongly activated and thus depolarized that the NMDA receptors, which are voltage sensitive, are activated. Activation of the NMDA receptors then allows $Ca^{++}$ to enter the cell, which is necessary for the long-term change in synaptic efficacy (Collingridge and Singer 1990). There is thus a strong nonlinearity in synaptic modification, in that synaptic modification only occurs for synapses onto neurons that are strongly activated. This learning rule may be expressed by

$$J_{ik} = \text{constant} \times \sum_m F(h_i^m) V_k^m \tag{7}$$

where $F$ is a nonlinear function of the postsynaptic activation which mimics the operation of the NMDA receptors in learning (Collingridge and Singer 1990). The above form may again include a decrementing term as in expression (4). The main effect of a rule of this type is that the postsynaptic activation elicited by the presentation of one of the learned axonal firing patterns would not be the same as the activation occurring during the learning of that pattern. In fact, in the absence of interference effects, the former would be

$$h_i^m{}_{\text{LEARNED}} = \sum_k J_{ik} V_k^m = F(h_i^m{}_{\text{LEARNING}}) \tag{8}$$

i.e. the activation after learning would be the same nonlinear function $F$ applied to the activation during learning. As far as the effective usage of the matrix memory goes, however, the arguments mentioned in the first two cases (subsections 2.1 and 2.2) still hold, as they depend only on the linearity of the synaptic modifications with respect to the presynaptic firing rates. Thus, there is a potential advantage in terms of storage capacity to sparse representations in the simpler case described by expression (2), which disappears in the presence of a presynaptic decrementing term in $J_{ik}$, expression (4).

## 3. Associative nets with nonlinear neurons

If the relation between the postsynaptic activation due to modifiable synapses and the firing rate of each output neuron is not linear, it is necessary to find a more appropriate criterion, to evaluate the capacity of the associative network, than the maximum number of independent associations which can be stored. In fact, the mapping from the axonal firing pattern in the input to the one in the output becomes nonlinear, which makes it less meaningful to regard the patterns themselves as embedded in a vector space, and less meaningful to look at the effective dimensionality they span. The choice of an alternative criterion depends on the functional role the network is supposed to play, and on the form of the nonlinear transfer function at the single-neuron level. In other words, the notion of the capacity of the network acquires a different meaning, and the new meaning has to be determined by understanding what the network does, and how single cell characteristics enable it to do that.

A possible starting point is to assume (a) that the main feature of the above transfer function is a threshold effect, whereby any subthreshold activation results in a firing rate of the output neurons equal to zero, and (b) that one of the roles of the network as a whole is to map the input pattern distribution into an output distribution clustered around a discrete set of prototype output patterns, which are themselves encoded in the synaptic efficacies. If $V^1$ and $V^2$ are two of these prototype patterns the network response to $V = c_1 V^1 + c_2 V^2$ need no longer be $R = c_1 R^1 + c_2 R^2$, as with the linear system. In fact requiring a degree of generalization within clusters would now mean that if $V$ is close to $V^1$ (i.e. it belongs to its 'basin of attraction') then $R$ should be even closer to $R^1$. Or $V$ could belong to a third cluster with an independently associated response. Within this framework one might measure the capacity in terms of the number of prototype patterns that can be discriminated by the network. Then, considering these patterns as drawn at random from a given distribution, one can perform a signal-to-noise analysis to estimate the capacity. It should be noted that this is a different measure from the total amount of stored and retrievable information, which, in some cases, might turn out to have a very different dependence on the sparseness of coding (see e.g. Treves 1990).

In order to work out a generic expression for the signal-to-noise ratio, which can then be analysed in specific cases, it is useful to assume the following. The contribution to the membrane potential from modifiable synapses is expressed again as in (1), and each synaptic efficacy is written:

$$J_{ik} = \frac{1}{N} \sum_m F(h_i^m) G(V_k^m) \tag{9}$$

where $F(h)$ and $G(V)$ are generic functions of their arguments (Brown *et al* 1990). The input patterns $V_k^m$ and output activations (during learning) $h_i^m$ are taken as drawn at random, and independently for each $i$ and $m$, from their respective distributions. It is convenient to introduce a notation for the averages and variances of the relevant quantities over these distributions:

$$\begin{aligned} \langle V \rangle &= a_V \qquad \langle V^2 \rangle - \langle V \rangle^2 = c_V \\ \langle G \rangle &= a_G \qquad \langle G^2 \rangle - \langle G \rangle^2 = c_G \\ \langle F \rangle &= a_F \qquad \langle F^2 \rangle - \langle F \rangle^2 = c_F. \end{aligned} \tag{10}$$

To obtain a measure of the sensitivity of the network to effects of interference between $p + 1$ prototype memories, one may apply one of the learned patterns in the input, and split the resulting output activation into the contribution due to the storage of that pattern, and that due to all the ($p$) others. The latter acts as noise, and its variance can be evaluated as

$$c_N = p c_F a_G^2 a_V^2 + (p^2/N) a_F^2 a_G^2 c_v + (p/N)[(a_F^2 + c_F) c_G (a_V^2 + c_V) + c_F a_G^2 c_V]. \tag{11}$$

Neglecting all other sources of noise that may affect the network, the signal from the pattern being recalled has to be at least of the same order of magnitude as this crosstalk noise to be effectively detected by the nonlinear output cells.

### 3.1. Hebb rule

One may consider again the simplest case, in which $G(V) = V$. Then $a_G = a_V$ differs from zero, and all terms survive in the above expression for the variance of the noise. In a parameter region in which the coding is not too sparse, the first term is the most important one. The square of the amplitude of the signal, $S^2$, on the other hand, can be evaluated as

$$S^2 = (F(h_{\max}) - F(h_{\min}))^2 (a_V^2 + c_V)^2. \tag{12}$$

This leads to the conclusion, neglecting the next two terms in $c_N$, that the maximum number of prototype patterns is of the order of

$$p < (F(h_{\max}) - F(h_{\min}))^2 [1 + (c_V/a_V^2)]^2 / c_F. \tag{13}$$

As $V$ is a positive number, the ratio $c_V/a_V^2$ (and therefore $p$) becomes large only if the coding is very sparse. For example, if the distribution for $V$ is binary, and $V = 0$ with probability $1 - a$, and $V = V_{\max}$ with probability $a$, then $p$ can grow as $1/a^2$ and become large in the sparse coding regime $a \ll 1$. This argument showing the advantage of sparse coding can be made more precise by further defining a particular model (see e.g. Nadal and Toulouse 1990).

### *3.2. Presynaptic decrementing term*

As with linear neurons, it is interesting to consider the effects of a presynaptic decrementing term in the learning rule analogous to that in the Singer–Stent rule specified in (4). A marked suppression of crosstalk results from this decrementing term being tuned to the average presynaptic firing rate. This case may be modelled by writing, as in (4),

$$G(V) = V - a_V \tag{14}$$

in which case $a_G = 0$ and $c_G = c_V$. Then the covariance of the noise reduces to

$$c_N = (p/N)(a_F^2 + c_F)(a_V^2 + c_V)c_V \tag{15}$$

and the signal is

$$S^2 = (F(h_{max}) - F(h_{min}))^2 c_V^2 \tag{16}$$

which leads to the estimate

$$p < \frac{N(F(h_{max}) - F(h_{min}))^2}{(a_F^2 + c_F)[1 + (a_V^2/c_V)]}. \tag{17}$$

The above expression indicates that now the number of prototype patterns can grow with $N$, the number of inputs to a cell. Moreover, the proportionality factor can be very large, i.e. $p$ can be much larger than $N$, if the following two criteria are met:

(i) the representation in the input is sparsely coded, so that $a_V^2 \ll c_V$,

(ii) the representation in the output is also sparsely coded, so that only few cells are subject to the activation $h_{max}$, and thus contribute little to the averages $(a_F^2 + c_F)$, which can be further reduced by setting the threshold $h_T$ appropriately. We note that the nonlinearity inherent in the NMDA learning rule would tend to produce a more sparse representation in the output pattern than occurred during learning, and that this would increase the capacity of the memory.

Sparse coding then enhances the storage capacity of a network with this type of learning rule.

### *3.3. Non-additive learning rules*

Among the learning rules which cannot be written in the form considered above, but are still of interest for a biological pattern associator, a particularly simple and appealing one has been studied by Willshaw and others (Willshaw *et al* 1969). Both synaptic efficacies and axonal firing rates are taken to assume only binary values, and in particular synaptic efficacies acquire the higher value if there is a conjunction of pre- and postsynaptic activity in at least one of a set of $p$ prototype patterns, and remain with the lower value otherwise. This type of network can be shown again to have an enhanced storage capacity if the coding is very sparse, as essentially the maximum number of prototypes grows as

$$p < 1/a^2 \tag{17}$$

where $a$ is the probability of a neuron (both in the input or in the output) being activated.

## 4. The representation and processing of information in some real neuronal networks in the primate taste and visual systems

The analyses given above show that in the associative nets with nonlinear neurons trained with all the learning rules considered (Hebb, Singer–Stent, Stanton and Sejnowski, and Willshaw), there is an advantage in sparse encoding (as compared with fully distributed encoding) in that this increases the number of learned patterns that can be discriminated, by decreasing the noise effects due to the storage of a large number of patterns. Further, in associative nets with linear neurons with positive continuous rates trained with the Hebb learning rule or the Stanton and Sejnowski rule, is there an advantage in sparse encoding in that this increases the number of independent associations that can be stored, by not allowing interference during recall. Only in associative nets with linear neurons trained with the Singer–Stent learning rule is there no advantage in sparse encoding with respect to increasing the number of stored associations. Nevertheless, encoding which is partly distributed and uses at least an

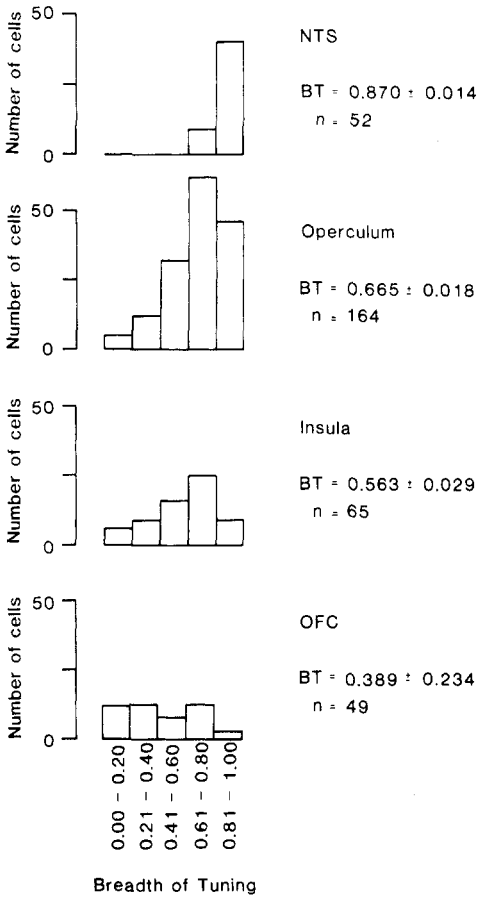


**Figure 1.** The breadths of tuning of neurons in different stages of the taste system. A value of 1 represents equal responses to all stimuli (i.e. very broad tuning), and a value of 0 represents a response to only 1 of the stimuli (see text). The stimuli used were 1 M glucose, 1 M NaCl, 0.01 M HCL, 0.001 M quinine HCl, water and 20% blackcurrant juice.

ensemble of active neurons to represent a pattern has the advantages of graceful degradation and generalization to near patterns, where nearness is measured by the correlation between the test input and an input learned previously. It also has the advantage that with nonlinear neurons more patterns can be stored than when only one neuron is firing (because the combination of inputs is stored). We thus see that in general, as noted previously (Rolls 1987, 1989a,c), the representation of information in an associative net will in many cases be a compromise between rather sparse representations with relatively few neurons active in each pattern to increase the number of patterns stored, and rather more distributed encoding to allow generalization and graceful degradation. It should be noted that the whole of this discussion applies to the representation of information in the axonal input which conveys the learned or conditioned input to the modifiable synapses, for it is in this access to the memory that the number of different keys must be maximized and interference minimized.

With this background, we now consider some evidence on encoding in the primate taste (see Rolls 1989b) and visual (Rolls 1990b,d) systems, and some additional ideas raised by the actual neuronal responses found. One reason for considering the taste system is that the range of stimuli to be encoded may be relatively limited and the processing may not involve complicated transforms, so that in this system it may be relatively straightforward to describe the encoding of information being performed by neurons.

### *4.1. The taste system*

The first central synapse of the gustatory system is in the rostral part of the nucleus of the solitary tract (NTS) (Beckstead and Norgren 1979, Beckstead *et al* 1980). In order to investigate the tuning of neurons in the nucleus of the solitary tract, the response of single NTS neurons to the prototypical stimuli NaCl, glucose, HCl, and quinine, and to water and a complex stimulus, blackcurrant juice, were measured in the macaque monkey. It was found that NTS neurons are relatively broadly tuned to the prototypical taste stimuli (Scott *et al* 1986a; see figure 1)†. The NTS projects via the thalamic taste area to the frontal opercular taste cortex and to the insula (Beckstead *et al* 1980). In these regions, gustatory areas were found, and it was discovered that the breadth of tuning of the neurons in these areas was finer than in the NTS (Scott *et al* 1986b, Yaxley *et al* 1990; see figure 1). The frontal opercular taste cortex projects into a fourth order gustatory area in the caudolateral orbitofrontal cortex (see Rolls 1989b), and here it was found that on average the tuning of the gustatory neurons was even finer (Rolls *et al* 1990; see figure 1). This analysis shows that one change which takes place in the representation of information in the gustatory system is that the breadth of tuning becomes finer. It is suggested that the reasons for this change in the breadth of tuning are as follows.

First, tuning may become fine by the secondary taste cortex because it is here, but not at earlier processing stages in primates, that the taste system is interfaced by association memories to other modalities. Consistent with this, in the rostral NTS, the frontal opercular taste cortex, and the insular taste cortex, the neurons found are mainly

†As noted above, what is observed in single-cell recordings is the degree to which different stimuli in a certain set activate a given cell. This is quantified by the breadth of tuning index introduced by Smith and Travers (1979). It is a measure of entropy derived from information theory, and calculated as $H = -k \Sigma_i p_i \log p_i$ where $p_i$ is the response to stimulus $i$ expressed as a fraction of the total response to all the stimuli in the set, and $k$ is a scaling constant, set so that $H = 1$ when the neuron responds equally well to all stimuli in the set. $H$ thus ranges from 0 to 1, and $H = 0$ when there is total specificity to one of the stimuli.

gustatory, and other modalities are not strongly represented. On the other hand, after these stages, i.e. when tuning has become fine, taste processing is interfaced to other modalities. For example, in and near to the orbitofrontal taste area, neurons with olfactory, with visual, and with somatosensory responses are found (see Rolls 1989b). Indeed, in a part of the orbitofrontal cortex, neurons with bimodal responses, for example to olfactory and taste stimuli, or to visual and taste stimuli, are found (see Rolls 1989b, Thorpe *et al* 1983). Also, in the amygdala, which receives inputs from the insular taste cortex (Mufson and Mesulam 1982), different modalities are brought together (Rolls 1981). Indeed, the visual projections to the amygdala follow the same rule, in that visual projections are not found to the amygdala from early stages of sensory analysis, but only from temporal lobe visual areas, that is, after much earlier processing (Van Hoesen 1981, Turner *et al* 1980). The reason for this is again probably that it is only possible to allow modalities to interact, in order to form associations for example, after much processing in each modality, so that interference due to lack of sparseness in the representation of the stimuli can be minimized.

A second important principle in leading to fine tuning may be related to the categorization of stimuli, and the need to prevent the categories from interacting or interfering with each other too strongly. For example, in the primate taste system, it is only after several stages of sensory information processing (which produce efficient categorization of the stimulus) that there is an interface to motivational systems. Thus in the primate, neuronal responses to gustatory stimuli in the NTS, the opercular taste cortex, and in the insular taste cortex, are not affected by hunger (see Rolls 1989b). It is only in the orbitofrontal taste area that neuronal responses are modulated by hunger, ceasing to occur, for example, to glucose if glucose has just been eaten to satiety (Rolls *et al* 1989). The explanation for this is probably as follows. If satiety were to operate at an early level of sensory analysis, then because of the broadness of tuning of neurons, responses to non-foods would become attenuated as well as responses to foods (and this could well be dangerous if poisonous non-foods became undetectable). This argument becomes even more compelling when it is realized that satiety typically shows some specificity for the particular food eaten, with other foods not eaten in the meal remaining relatively pleasant (Rolls 1984a, 1989b). Unless tuning was relatively fine, this mechanism could not operate, for reduction in neuronal firing after one food had been eaten would inevitably reduce behavioral responsiveness to other foods. Indeed, it is of interest to note that such a sensory-specific satiety mechanism can be built by arranging for tuning to particular foods to become relatively specific at one level of the nervous system (as a result of categorization processing in earlier stages), and then at this stage (but not at prior stages) to allow habituation to be a property of the synapses. This would result in a decreased response to one taste, but not to another taste unless they were very similar. It appears that precisely this mechanism is found in the primate secondary taste cortex in the caudolateral orbitofrontal region (Rolls *et al* 1989). In effect, given that many tastants can be represented in a low-dimensional space (bounded by sweet, salt, bitter and sour, and possibly one or two other primary tastes), in early stages of neural transmission in the taste nerve, rather broad tuning will maximize information transfer in a relatively limited number of fibres, will allow fine discriminations to be made by small alterations in the rate of firing of many nerve fibres, and allow useful redundancy in case of fibre loss. Such processing may often be approximately linear. Later on, during cortical processing, it may become important to set up further categories formed by nonlinear combinations of the prototypical tastants. For example, sweet and sour could be combined nonlinearly to make a new taste which might be processed separately from both sweet taste and sour

taste (so that sensory-specific satiety might occur for the combination of sweet and sour, but not for sweet alone or sour alone). Such combinations to form new categories are likely to become even more evident when the modalities are combined, for example, with the flavour of a food being determined by taste, smell and touch inputs.

### *4.2. The encoding of visual stimuli in the temporal lobe cortical areas*

Another way in which the representation of information across ensembles of neurons in the brain is being investigated is by analysing the responses of single neurons in the temporal lobe visual cortex which respond preferentially to faces. The question considered is whether information which could specify the face of one individual is represented by the firing of one neuron, or whether the pattern of firing of an ensemble is needed to enable identification of the individual being seen.

Neurons which respond preferentially or selectively to faces are found in certain areas of the temporal lobe visual cortex, which receive their inputs via a number of cortico–cortical stages from the primary visual cortex, the striate cortex, through prestriate visual areas (Cowey 1979, Desimone and Gross 1979, Seltzer and Pandya 1978). The responses of these neurons to faces are selective in that they are 2–10 times as large to faces as to gratings, simple geometrical stimuli, or complex 3D objects (Perrett *et al* 1982, Baylis *et al* 1985, 1987). They are probably a specialized population for processing information from faces, in that they are found primarily in architectonic areas TPO, TEa and TEm, and are not just the neurons with the most complex types of response found throughout the temporal lobe visual areas (Baylis *et al* 1987). The advantage of such a specialized system in the primate may lie in the importance of rapid and reliable recognition of other individuals using face recognition so that appropriate social and emotional responses can be made (Rolls 1984b, 1990b–d).

In experiments to determine how information which could be used to specify an individual is represented by the firing of these neurons, it has been shown that in many cases (77% of one sample), these neurons are sensitive to differences between faces (Baylis *et al* 1985), but that each neuron does not respond only to one face. Instead, each neuron has a different pattern of responses to a set of faces, as illustrated in figure 2. Such evidence shows that the responses of each of these neurons in the cortex in the superior temporal sulcus does not code uniquely for the face of a particular individual. Instead, across a population of such cells information is conveyed which would be useful in making different behavioral responses to different faces. Thus information which specifies an individual face is present across an ensemble of such cells. In that each neuron does not respond to only one face, and in that a particular face can activate many neurons, these are not 'grandmother' cells (Barlow 1972). However, in that their responses are relatively specialized both for the class 'faces' and within this class, they could contribute to relatively sparse coding of information over relatively few cells (Barlow 1972). It may be noted that even if individual neurons in this population are not tuned to respond completely specifically to only face stimuli, it is nevertheless the case that the output of such an ensemble of neurons would be useful for distinguishing between different faces. The appropriateness of these neurons for such a function is enhanced by their relative constancy of response over some physical transforms, such as size, contrast, and colour (Perrett *et al* 1982, Rolls and Baylis 1986). These findings lead to the hypothesis that these neurons are filters, the output of which could be used for recognition of different individuals, and in emotional responses made to different individuals.

It is unlikely that there are further processing areas beyond those described where

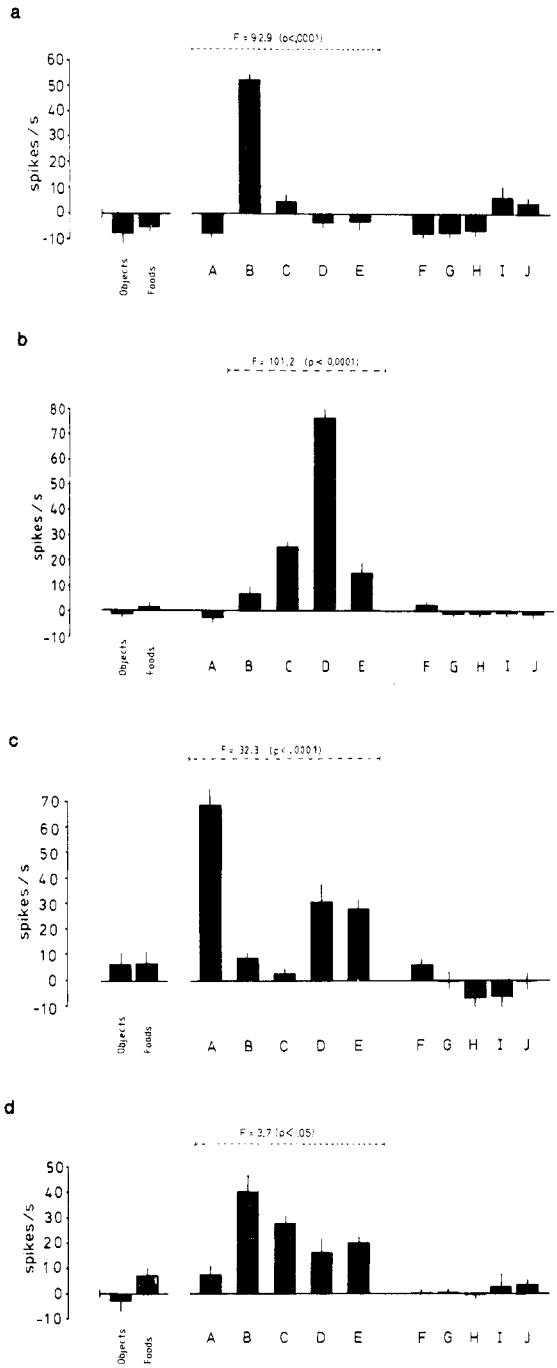


**Figure 2.** The responses of four cells (*a*)–(*d*) in the cortex in the superior temporal sulcus to a variety of face (A–E) and non-face (F–J) stimuli. The bar represents the mean firing rate response above the spontaneous baseline firing rate with the standard error calculated over 4–10 presentations. The *F* ratio for the analysis of variance calculated over the face sets indicates that the units shown range from very selective (top) to relatively non-selective (bottom). (From Baylis *et al* 1985.)

ensemble coding changes into grandmother cell encoding. Part of the evidence for this is that anatomically there does not appear to be a whole further set of visual processing areas present in the brain. Indeed, from the temporal lobe visual areas such as those described, outputs are taken to limbic and related regions such as the amygdala where it is believed that cross-modal associations to reinforcing stimuli (e.g. vision to taste) are made (see Rolls 1990c); and via the parahippocampal gyrus and entorhinal cortex to the hippocampus where it is suggested that autoassociation networks underly the formation of episodic memories (Rolls 1989a, 1990a, Treves and Rolls 1990). Indeed, tracing this pathway onwards, Leonard *et al* (1985) have found a population of neurons with face-selective responses in the amygdala, and in the majority of these neurons different responses occur to different faces, with ensemble encoding rather than grandmother encoding still being present. Thus, in at least this part of the visual system, it appears that neurons become relatively finely tuned, so that across a sparsely encoded ensemble they convey information which could specify an individual, and would be useful for interfacing to association memories, found not in the visual cortex itself, but instead in limbic structures such as the amygdala and hippocampus.

## 5. Conclusions

It is conceivable that the coding of information that has to be processed by associative memory networks in the brain is designed to optimize a number of different and sometimes conflicting measures of performance. What has been shown here is that under rather general conditions sparse codings enhance the storage capacity of these systems, as defined in terms of a discrete number of memorized associations. The biological relevance of the performance measure adopted, and of the way it might depend on the coding of information, has been substantiated by considering specific sensory systems and the functions they may implement in primates. A verification of the ideas presented here entails a more advanced knowledge of the operation of the relevant neuronal networks, both at the systems level and at the level of the basic circuitry, and offers a challenge for research in several branches of neuroscience. It will be of interest, for example, to analyse the circuitry, the mechanisms of synaptic plasticity, and the representations of information found in the amygdala, which is implicated in associations of the type discussed in this paper (Rolls 1990c). A particular empirical investigation, that theoretical considerations suggest as important, is to assess how finely neurons are tuned by using large sets of input stimuli, and measuring the degree to which each neuron found in a given brain region responds to all the members of the set of stimuli.

## References

Barlow H B 1972 Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1** 371–94

Baylis G C, Rolls E T and Leonard C M 1985 Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey *Brain Res.* **342** 91–102

—— 1987 Functional subdivisions of temporal lobe neocortex *J. Neurosci.* **7** 330–42

Beckstead R M and Norgren R 1979 An autoradiographic examination of the central distribution of the trigeminal, facial, glossopharyngeal, and vagal nerves in the monkey *J. Comp. Neurol* **184** 455–72

Beckstead R M, Morse J R and Norgren R 1980 The nucleus of the solitary tract in the monkey: Projections to the thalamus and brainstem nuclei *J. Comp. Neurol.* **190** 259–82

Brown T H, Kairiss, E W and Keenan C L 1990 Hebbian synapses: biophysical mechanisms and algorithms *Ann. Rev. Neurosci.* **13** 475–511

Collingridge G L and Singer W 1990 Excitatory amino acid receptors and synaptic plasticity *Trends Pharm. Sci.* **11** 290–96

Cowey A 1979 Cortical maps and visual perception *Quart. J. Exp. Psychol.* **31** 1–17

Desimone R and Gross C G 1979 Visual areas in the temporal lobe of the macaque *Brain Res.* **178** 363–80

Kohonen T, Oja E and Lehtio P 1981 Storage and processing of information in distributed associative memory systems *Parallel Models of Associative Memory*, ed G E Hinton and J A Anderson (Hillsdale, NJ: Erlbaum) ch 4, pp 105–43

Leonard C M, Rolls E T, Wilson F A W and Baylis G C (1985) Neurons in the amygdala of the monkey with responses selective for faces *Behav. Brain Res.* **15** 159–76

Levy W B 1985 Associative changes in the synapse: LTP in the hippocampus *Synaptic Modification, Neuron Selectivity, and Nervous System Organization* ed W B Levy, J A Anderson and S Lehmkuhle (Hillsdale, NJ: Erlbaum) ch 1, pp 5–33

Levy W B and Desmond N L 1985 The rules of elemental synaptic plasticity *Synaptic Modification, Neuron Selectivity, and Nervous System Organization* W B Levy, J A Anderson and S Lehmkuhle (Hillsdale, NJ: Erlbaum) ch 6, pp 105–21

Marr D 1970 A theory for cerebral neocortex *Proc. R. Soc.* B **176** 161–234

—— 1971 Simple memory: a theory for archicortex. *Phil. Trans. R. Soc.* B **262** 24–81

Mufson E J and Mesulam M-M 1982 Insula of the old world monkey. II: Afferent cortical input and comments on the claustrum *J. Comp. Neurol.* **212** 23–37.

Nadal J and Toulouse G 1990 Information storage in sparsely coded memory nets *Network* **1** 61–74

Perrett D I, Rolls E T and Caan, W 1982 Visual neurons responsive to faces in the monkey temporal cortex *Exp. Brain Res.* **47** 329–42

Rolls E T 1981 Responses of amygdaloid neurons in the primate *The Amygdaloid Complex* ed Y Ben-Ari (Amsterdam: Elsevier) pp 383–93

—— 1984a The neurophysiology of feeding *Int. J. Obesity* **8** suppl 1 139–50

—— 1984b Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces *Human Neurobiol.* **3** 209–22

—— 1987 Information representation, processing and storage in the brain: analysis at the single neuron level *The Neural and Molecular Bases of Learning* ed J-P Changeux and M Konishi (Chichester: Wiley) pp 503–40

—— 1989a The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus *The Computing Neuron* ed R Durbin, C Miall and G Mitchison (Wokingham: Addison-Wesley) ch 8, pp 125–59

—— 1989b Information processing in the taste system of primates *J. Exp. Biol.* **146** 141–64

—— 1989c Functions of neuronal networks in the hippocampus and neocortex in memory *Neural Models of Plasticity: Experimental and Theoretical Approaches* ed J H Byrne and W O Berry (San Diego: Academic) ch 13, pp 240–65

—— 1990a Functions of the primate hippocampus in spatial processing and memory *Neurobiology of Comparative Cognition* ed D S Olton and R P Kesner (Hillsdale, NJ: Erlbaum) ch 12, pp 339–62

—— 1990b The processing of face information in the primate temporal lobe *Processing Images of Faces* ed V Bruce and M Burton (Norwood, NJ: Ablex) in press

—— 1990c A theory of emotion, and its application to understanding the neural basis of emotion *Cognition and Emotion* **4** 161–90

—— 1990d The representation of information in the temporal lobe visual cortical areas of macaques *Advanced Neural Computers* ed R Eckmiller (Amsterdam: North-Holland) pp 69–78

Rolls E T and Baylis G C 1986 Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey *Exp. Brain Res.* **65** 38–48

Rolls E T, Sienkiewicz Z J and Yaxley S 1989 Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey *Eur. J. Neurosci.* **1** 53–60

Rolls E T, Yaxley S and Sienkiewicz Z J 1990 Gustatory responses of single neurons in the orbitofrontal cortex of the macaque monkey *J. Neurophysiol.* **64** 1055–66

Scott T R, Yaxley S, Sienkiewicz Z J and Rolls E T 1986a Taste responses in the nucleus tractus solitarius of the behaving monkey *J. Neurophysiol.* **55** 182–200

—— 1986b Gustatory responses in the frontal opercular cortex of the alert cynomolgus monkey *J. Neurophysiol.* **56** 876–90

Sejnowski, T J 1977 Storing covariance with nonlinearly interacting neurons *J. Math. Biol.* **4** 303–21

Seltzer B and Pandya D N 1978 Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey *Brain Res.* **149** 1–24

Singer W 1987 Activity-dependent self-organization of synaptic connections as a substrate for learning *The Neural and Molecular Bases of Learning* ed J-P Changeux and M Konishi (Chichester: Wiley) pp 301–35

Smith D V and Travers J B 1979 A metric for the breadth of tuning of gustatory neurons *Chem. Senses Flavour* **4** 215–29

Stanton P K and Sejnowski T J 1989 Associative long-term depression in the hippocampus indiced by Hebbian covariance *Nature* **339** 215–8

Stent G S 1973 A physiological mechanism for Hebb's postulate of learning *Proc. Natl Acad. Sci.* USA **70** 997–1001

Thorpe S J, Rolls E T and Maddison S 1983 Neuronal activity in the orbitofrontal cortex of the behaving monkey *Exp. Brain Res.* **49** 93–115

Treves A 1990 Graded-response neurons and information encodings in autoassociative memories *Phys. Rev.* A **42** 2418–30

Treves A and Rolls E T 1990 Neuronal networks in the hippocampus involved in memory *Proc. 11th Sitges Conf. on Neural Networks* ed L Garrido (Berlin: Springer) in press

Turner B H, Mishkin M and Knapp M 1980 Organization of the amygdalopetal modality-specific cortical association areas in the monkey *J. Comp. Neurol* **191** 515–43

Van Hoesen G W 1981 The differential distribution, diversity and sprouting of cortical projections to the amygdala in the rhesus monkey *The Amygdaloid Complex* ed Y Ben-Ari (Amsterdam: Elsevier) pp 79–90

Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 Non-holographic associative memory *Nature* **222** 960–1

Willshaw D and Dayan P 1990 Optimal plasticity from matrix memories: what goes up must come down *Neural Comput.* **2** 85–93

Yaxley S, Rolls E T and Sienkiewicz Z J 1990 Gustatory responses of single neurons in the insula of the macaque monkey *J. Neurophysiol.* **63** 689–700