Neural Networks and Brain Function

Edmund T. Rolls

University of Oxford
Department of Experimental Psychology
Oxford
England

Alessandro Treves

International School of Advanced Studies
Programme in Neuroscience
34013 Trieste
Italy

OXFORD UNIVERSITY PRESS • OXFORD 1998

Preface

This document provides Appendix A3 of Rolls and Treves (1998) *Neural Networks and Brain Function* published by Oxford University Press.

This appendix is being made easily available, for it contains material on the quantitative analysis of the capacity of pattern association networks not published elsewhere, and so that this material is easily available as this Appendix is referred to by Rolls (2016) *Cerebral Cortex: Principles of Operation* published by Oxford University Press.

In producing this version, we have corrected minor typographical errors. To facilitate comparison with Rolls (2016) (and earlier books Rolls and Deco (2002), Rolls (2008), Rolls and Deco (2010), Rolls (2012), and Rolls (2014)), we have also used, in the first part of this text, an alternative notation for the presynaptic and postsynaptic firing rates, as follows:

The presynaptic firing rate r' in Rolls and Treves (1998) is also denoted as x in this document.

The postsynaptic firing rate r in Rolls and Treves (1998) is also denoted as y in this document.

In addition, this Appendix in Rolls and Treves (1998) was Appendix A3 with sections A3.1, equations A3.1 etc, and becomes Appendix C with sections C.1, equations C.1 etc in this document.

Appendix 3 Pattern associators

Chapters 2 and 3 provide introductory accounts of the operation of pattern associators and of autoassociative memories; networks comprising a few units, often just binary ones, are used to demonstrate the basic notions with minimal complication. In later chapters, instead, these notions are argued to be relevant to understanding how networks of actual neurons in parts of the brain operate. It is obvious that real networks are much more complex objects than are captured by the models used as illustratory examples. Here we address some of the issues arising when considering the operation of large networks implemented in the brain.

Beyond providing an intuitive introduction to the real systems, a second important reason for striving to simplify the complexities of neurobiology is that formal, mathematical models of sufficient simplicity are amenable to mathematical analysis, which is much more powerful, particularly in extracting quantitative relationships, than either intuition or computer simulation (which also requires a degree of simplification, anyway). Providing all the necessary tools for formal analyses of neural networks is largely outside the scope of this book, and many of these tools can be found, for example, in the excellent books by Amit (1989), and by Hertz, Krogh and Palmer (1991), and in the original and review literature. Nevertheless, we shall sketch some of the lines of the mathematical approaches and refer back to them in discussing issues of realism and simplification, because some appreciation of the analytical methods is important for an understanding of the domain of validity of the results we quote.

C.1 General issues in the modelling of real neuronal networks

C.1.1 Small nets and large nets

In Chapter 2 we used as an example of an associative memory a network with 6 input axons and 4 output cells. It is easy to simulate larger nets on a computer, and to check that the same mechanisms, associative storage and retrieval, with generalization, fault tolerance, etc., can operate just as successfully. In the brain, we may be considering local populations comprising tens of thousands of neurons, each receiving thousands of synaptic inputs just from the sensory stream carrying the conditioned stimulus. Formal mathematical analyses can be worked out in principle for nets of any size, but they are always much simpler and more straightforward (and this is a point which may not be known to experimentalists) in the limit when the size approaches infinity. In practice, infinite means very large, so that finite size effects can be neglected and the central limit theorem applied to probabilities, whenever needed. Usually, what makes life simpler by being very large is not just the number of cells in the population, but also the number of inputs per cell. This is because the precise identity and behaviour of the individual cells feeding into any given neuron become unimportant, and what remain important are only the distributions characterizing, for example, input firing rates and synaptic efficacies; often, not even the full distributions but only their means and maybe a few extra moments. Therefore, while for those conducting very detailed conductance-based simulations, or even more for those developing artificial networks in hardware, the problem

tends to be that of reaching sizes comparable to those of brain circuits, and the number of units or of inputs per unit simulated is a score of success for the enterprise, for the theoretician the problem with size, if there is any, is to check that actual networks in the brain are large enough to operate in the same way as the infinitely large formal models – which is usually concluded to be the case.

C.1.2 What is the output of a neuron?

Neurons may communicate, or more in general interact with each other in a variety of ways. The only form of communication considered in this book is the emission of action potentials carried by the axon and resulting in release of neurotransmitter at synaptic terminals. This is clearly the most important way in which central neurons affect one another, but one should note that there are alternatives that appear to be important in certain systems (see for example Shepherd (1988), such as dendro-dendritic synapses (described in the olfactory bulb by Rall and Shepherd (1968)), gap junctions in various parts of the developing nervous system, or ephaptic interactions, that is the (minor) electrical couplings produced by sheer proximity, for example among the tightly packed cell bodies of pyramidal neurons in the hippocampus. The role that these alternative forms of interaction (which, to be noted, are still local) may play in operations implemented in the corresponding networks remains to a great extent to be elucidated.

One should also note that an experimenter can access a larger set of parameters describing the 'state' of the neuron than is accessible to the postsynaptic neurons that receive inputs from it. Thus, one can measure membrane potential with intracellular electrodes in the soma, calcium concentration, various indicators of metabolic activity, etc., all variables that may be correlated to some degree with the rate of emission of action potentials. Action potentials are what triggers the release of neurotransmitter that is felt, via conductance changes, by receiving neurons, that is they are the true output, and since they are to a good approximation self-similar, only their times of occurrence are relevant. Therefore, the fuller description of the output of cell i that we consider here is the list of times $\{t^k\}_i$ for the emission of each of the action potentials, or spikes, which are indexed by k.

Reduced descriptions of the output of a cell are often sufficient, or thought to be sufficient. For example, if one looks at a given cell for 20 ms, records its action potentials with 1 ms resolution, and the cell never fires at more than 500 Hz, the full description of the output would be a vector of 20 binary elements (each signifying whether there is a spike or not in the corresponding 1 ms bin), which could in principle have as many as $2^{20} \approx 1{,}000{,}000$ different configurations, or values (most of which will never be accessed in practice, particularly if the cell never fires more than 10 spikes in the prescribed 20 ms window). If the precise emission time of each spike is unimportant to postsynaptic cells, a reduced description of the cell's output is given by just the number of spikes emitted in the time window, that is by specifying one of the 11 values from 0 to 10, with a nearly 10^5 reduction in the number of possible outputs.

Whether a reduction in the output space to be considered is justified is a question that can be addressed experimentally case by case, by using the information theoretic measures introduced in Appendix A2. One may ask what proportion of the information conveyed by the full output is still conveyed by a particular reduced description (Optican and Richmond 1987), and whether any extra information that is lost would in any case be usable, or decodable, by receiving neurons (Rolls, Treves and Tovee 1997). In higher sensory cortices, at least in parts of the visual, olfactory, and taste systems, it appears (Rolls, Critchley and Treves 1996, Tovee and Rolls 1995, Tovee, Rolls, Treves and Bellis 1993) that the simple firing rate, or equivalently the number of spikes recorded in a time window of a few tens of ms, is indeed a reduced

description of neuronal output that preserves nearly all the information present in the spike emission times (at the level of single cells). This finding does not of course preclude the possibility that more complex descriptions, involving the detailed time course of neuronal firing, may be necessary to describe the output of single cells in other, for example more peripheral, systems; nor does it preclude the possibility that populations of cells convey some information in ways dependent on the precise relative timing of their firing, and that could not be revealed by reporting only the firing rate of individual cells, as suggested by findings by Abeles and collaborators (Abeles, Bergman, Margalit and Vaadia 1993) in frontal cortex and by Singer and collaborators (Gray, Konig, Engel and Singer 1989) in primary visual cortex.

Turning to the question of what is decodable by receiving neurons, that is of course dependent on the type of operation performed by the receiving neurons. If this operation is simple pattern association, then the firing rates of individual input cells measured over a few tens of ms is all the output cells are sensitive to, because the fundamental operation in a pattern associator is just the dot product between the incoming axonal pattern of activity and the synaptic weight vector of each receiving cell. This implies that each input cell contributes a term to a weighted sum (and not a factor that interacts in more complex ways with other inputs) and precise emission times are unimportant, as thousands of inputs are effectively integrated over, in space along the dendritic tree, and in time (allowing for leakage) between one action potential of the output cell and the next.

The firing rate of each cell i, denoted y_i (or x_j for an input), is therefore what will be considered as a suitable description of its output, for pattern associators. Since these networks, moreover, are feedforward, and there are no loops whereby the effect of the firing of a given cell is felt back by the cell itself, that is there is no recurrent *dynamics* associated with the operation of the network, it does not really matter whether the inputs are considered to persist only briefly or for a prolonged period. Neglecting time-dependent effects such as adaptation in the firing, the outputs are simply a function of the inputs, which does not involve the time dimension. Pattern associators, like other types of feedforward nets, can thus be analysed, to a first approximation, without having to describe the detailed time course of a cell's response to inputs that vary in time: the inputs can be considered to be carried by steady firing rates, and likewise the outputs.

C.1.3 Input summation and the transfer function

Specifying the way inputs are summed and transduced into outputs means in this case assigning functions that transform a set of input rates $\{x_j\}$ into one rate for each output cell i, that is functions $y_i(\{x_j\})$. This is to be a very compressed description of what in real neurons is a complex cascade of events, typically comprising the opening of synaptic conductances by neurotransmitter released by presynaptic spikes, the flow of synaptic currents into the cell, their transduction through the dendrites into the soma possibly in conjunction with active processes or non-linear interactions in the dendrites themselves, and the initiation of sodium spikes at the axon hillock. A simple summary of the above which is useful in basic formal models is as follows. Input rates from C presynaptic cells are summed with coefficients w representing the efficacy of the corresponding synapses into an activation variable h for each of N output cells:

$$h_i = \sum_{j=1,C} w_{ij} x_j = w_{i1} x_1 + ... + w_{ij} x_j + ... + w_{iC} x_C \tag{C.1}$$

One should note that more complex forms of input integration have been suggested as alternative useful models in the connectionist literature (cf. the Sigma-Pi units in Rumelhart

and McClelland (1986)), and that for some of them it might even be argued that they capture what is observed in certain special neuronal systems.

The activation is then converted into the firing rate of the output cell via a transfer function that includes at least a non-linearity corresponding to the threshold for firing, and an approximately linear range above threshold. A simple one is

$$y_i = g_i(h_i - \theta_i)\,\Theta(h_i - \theta_i) \tag{C.2}$$

where $\Theta(x)$ is the step function which equals 1 for $x > 0$ and 0 for $x < 0$, and g is the gain of the transfer function. Since both in the input and output one deals with steady rates, the activation is generally thought to represent the current flowing into the soma (Treves 1990), and hence θ is a threshold for this current to elicit a non-zero steady response. If more transient responses were considered, it would be appropriate to consider, instead, the membrane potential and its corresponding threshold (Koch, Bernander and Douglas 1995). This is because the triggering of a single action potential is due to the establishment, at the local membrane level, of a positive feedback process involving the opening of voltage-dependent sodium channels together with the depolarization of the membrane. Hence, what is critical, to enter the positive feedback regime, is reaching a critical level of depolarization, that is reaching a membrane potential threshold. What we call (in a somewhat idealized fashion) *steady* state, instead entails a constant current flowing into the soma and a constant rate of emission of action potentials, while the membrane potential particularly at the soma is *oscillating* with each emission. Hence, the threshold for firing is at steady state a threshold for the current. This point is actually quite important in view of the fact that, as explained in Appendix A5, currents are attenuated much less, in reaching the soma, than voltages. Therefore, the effectiveness of inputs far out on the apical dendrites and close to the soma of a pyramidal cell may be much more comparable, at steady state, than might be concluded from experiments in which afferent inputs are transiently activated in order to measure their strengths. In addition to being on average more similar relative to each other, both far and near inputs on a dendrite are stronger in absolute terms, when they are measured in terms of steady-state currents. This should, however, not be taken to imply that very few typical synaptic inputs are sufficient to fire a pyramidal cell; the number of required inputs may still be quite large, in the order of several tens. Precise estimates have been produced with studies *in vitro*, but in order to understand what the conditions are *in vivo*, one must take into account typical sizes of Excitatory Post-Synaptic Currents (EPSCs) elicited in the soma by synaptic inputs to different parts of a dendrite, and the effects of inhibitory inputs to the cell (see for example Tsodyks and Sejnowski (1995).

For an isolated cell, the gain g_i (in Hz/mA) and the threshold θ_i (in mA) could be made to reproduce the corresponding parameters measured from the response to current injection, for example in a slice. Here, however, gain and threshold are supposed to account not only for the properties of the cell itself, but also for the effects of non-specific inputs not explicitly included in the sum of Eq. C.1, which only extends over principal cells. That is, the networks we consider are usually networks of principal cells (both in the input and in the output), and local interneurons (for example those mediating feedforward and feedback inhibition) are not individually represented in the model, but only in terms of a gross description of their effect on the transduction performed by principal cells. Divisive inhibition may be represented as a modulation of the gain of pyramidal cells, and subtractive inhibition as a term adding to their threshold. Both may be made a function of the mean activity on the input fibres, to represent feedforward control of overall level of activation; and of the mean activity of output cells, to represent feedback effects. Likewise, the general, non-specific effects of neuromodulatory afferents may be represented as modifications in the parameters of the transfer function.

The threshold linear transfer function of Eq. C.2 is the simplest one that reproduces the two basic features of a threshold and a graded range above threshold. Many other choices are of course possible, although they may be less easy to treat analytically in formal models. The two most widely used forms of transfer function are in fact even simpler, but at the price of renouncing representing one of those two features: the binary transfer function captures only the threshold but not the graded response above it, and the linear transfer function only the graded response and not the threshold. The so-called logistic (or sigmoid) function (Fig. 1.3) is derived as a 'smoothing' of the sharp threshold of a binary unit, and is also widely used especially in connectionist models. It can have the undesirable feature that the resulting firing rates tend to cluster around both zero and the maximal rate, producing an almost binary distribution of values that is most unlike the typical firing statistics found in real neurons.

C.1.4 Synaptic efficacies and their modification with learning

The effect of the firing of one spike on the activation of a postsynaptic cell i is weighted by a single coefficient w_{ij}, parametrizing the efficacy of the synapse between the firing cell j and the receiving one. Even in more detailed models in which synaptic inputs are represented in terms of conductances and their respective equilibrium potentials, the size of the open conductance can still be parametrized by a weight coefficient. Modifiable synapses, in the models, are those in which these coefficients are taken to change in the course of time, depending for example on the pre- and post-synaptic cell activity and their relation in time. A mathematical representation of such modifiability is usually called a learning rule.

The modifiability expressed in formal models by learning rules is, at a general level, implemented in the real nervous system by various forms of synaptic plasticity. Among them, the one with the most interesting properties is the phenomenon, or group of phenomena, called long-term potentiation, or LTP (and its counterpart long-term depression, LTD), first discovered in the hippocampus by Bliss and Lømo (1973). LTP has long been a focus of attention because it appears to fulfil the desiderata for a learning mechanism formulated on a conceptual basis by the psychologist Donald Hebb (1949): essentially, its sustained nature and the fact that its induction depends in a conjunctive way on both activity in the presynaptic fibre and activation (depolarization) of the postsynaptic membrane. One crucial component that senses this conjunction, or AND function, is the so-called NMDA receptor. This receptor opens when glutamate released from an activated presynaptic terminal binds at the appropriate site, but to let ions pass through it also requires that magnesium ions be expelled by sufficient depolarization of the postsynaptic membrane. (At normal potentials magnesium ions block the channel by entering it from the outside of the cell and acting effectively as corks.) The entry of calcium ions through the unblocked channel appears to release a complex cascade of events that result in the LTP of the synapse, possibly expressed as a combination of increased average release of neurotransmitter by each incoming spike, and an increased postsynaptic effect of each quantum of neurotransmitter binding at the ordinary (AMPA) glutamate receptors (co-expressed on the postsynaptic membrane with the NMDA ones). The precise nature and relevance of the steps in the cascade, and in general the mechanisms underlying different subforms of LTP, the mechanisms involved in LTD, the differences between the plasticity evident during development and that underlying learning in the adult, are all very much topics of current attention, still at one of the frontiers of neurobiological research. New important phenomena may still be discovered if the appropriate experimental paradigms are employed, and overall, it is fair to say that present knowledge about synaptic plasticity does not tightly constrain theoretical ideas about its occurrence and its roles. Although the discoveries that have been made have provided powerful inspiration for the further development and refinement of

theoretical notions, the latter may potentially exert an even more profound stimulus for the design and execution of the crucial experiments.

The main features of any learning rule are (a) which factors, local or not, the modification depends on, (b) what precise form it takes and (c) its time course of induction and expression. On point (a), experimental evidence does provide some indications, which are however much more vague on points (b) and (c).

A learning rule is called local if the factors determining the modification refer only to the pre- and postsynaptic cells; most commonly the firing rates of the two cells are taken to be these factors. It should be noted that whereas the firing rate of the presynaptic cell may be directly related to the size of the modification, in that each spike may have the potential for contributing to synaptic change, it is unclear whether the postsynaptic site may be sensitive to individual spikes emitted by the postsynaptic cell (and retrogradely transmitted up to the synapse as sharp variations in membrane potential), or rather more to a time averaged and possibly partly local value of the membrane potential, such as that which, in the operation of NMDA receptors, is thought to be responsible for relieving the magnesium block. It is noted in Appendix A5 that voltage transmission from the cell body or proximal part of the dendrite towards the apical part of the dendrite is not severely attenuated (Carnevale and Johnston 1982). An implication of this is that much of the postsynaptic term in a learning rule can be felt throughout all the dendrites, so that depolarization of the cell soma associated with fast firing *in vivo* would lead to sufficient depolarization of the dendrite to relieve the block of NMDA receptors. In any case, the term local implies that the relevant factors are available at the synapse, but it does not rule out the necessity, under a strict analysis of the molecular mechanisms of the change, for short-distance messengers. Thus, to the extent that the change is expressed on the presynaptic site, there is a need for a retrograde messenger that conveys presynaptically the signal that the proper factors for induction are present postsynaptically. Nevertheless, local rules do not need long-distance messengers that bring in signals from third parties, such as the error signals backpropagating across cell populations required in the non-local and rather biologically implausible learning rule employed in backpropagation networks.

In the standard learning rules we consider, modifications are expressed as sums of individual synaptic changes Δw, each supposed to occur at a different moment in time. Δw is a product of a function of the presynaptic firing rate, and of a function of the postsynaptic rate at that moment in time. The two functions may be different from each other, but in one common form of learning rule they are in fact equal: they are just the rate itself minus its average

$$\Delta w_{ij} = \gamma(y_i - < y_i >)(x_j - < x_j >) \tag{C.3}$$

so that, in this simple model, any fluctuations from the average values, when occurring both pre- and postsynaptically, are sufficient to elicit synaptic changes. The plasticity γ is a parameter quantifying the average amount of change. The change Δw is then independent of the value of the synaptic efficacy at the time it occurs and of previous changes (the overall modification is just a sum of independent changes). The modification is to be conceived as being from a baseline value large enough to keep the synaptic efficacy, at any moment in time, a positive quantity, since it ultimately represents the size of a conductance (in fact, of an excitatory one). The important aspects of the precise form of Eq. C.3 are the linear superposition of successive changes, and the fact that these are both positive and negative, and on average cancel out, thus keeping the synaptic value fluctuating around its baseline. Minor modifications that are broadly equivalent to Eq. C.3 are expressions that maintain these two aspects. For example, the postsynaptic factor may be a different function of the postsynaptic rate, even a positive definite function (the presynaptic factor is sufficient to ensure the average cancellation of different contributions), and it may also be a function not of the

postsynaptic rate but of a variable correlated with it, such as a time-averaged value of the postsynaptic membrane potential. This time average may even be slightly shifted with respect to the presynaptic factor, to represent what happens to the postsynaptic cell over the few tens of ms that follow the presynaptic spike. Learning rules, instead, that violate either the linear superposition or the average cancellation aspects represent major departures from Eq. C.3, and may lead to different functional performance at the network level. This is evident already in the later part of this appendix, where a signal-to-noise analysis reveals the importance of the average cancellation in the presynaptic factor, leading to Eq. C.15. Thus some form of synaptic plasticity analogous to what is referred to as heterosynaptic LTD (see Chapter 2) is found to be computationally crucial in providing a balance for increases in synaptic efficacies, modelled after LTP. Learning rules that do not fall into this general class are discussed in this book under separate headings.

The third feature of the learning rule, its time course, is expressed in the standard rule we consider by taking each change to be induced over a short time (of less than 1 s, during which the synaptic efficacy is still effectively the one before the change), and then persisting until forgetting, if a forgetting mechanism is included, occurs. The simplest forgetting mechanism is just an exponential decay of the synaptic value back to its baseline, which may be exponential in time or in the number of changes incurred (Nadal, Toulouse, Changeux and Dehaene 1986). This form of forgetting does not require keeping track of each individual change and preserves linear superposition. In calculating the storage capacity of pattern associators and of autoassociators, the inclusion or exclusion of simple exponential decay does not change significantly the calculation, and only results, as can be easily shown, in a different prefactor (one 2.7 times the other) for the maximum number of associations that can be stored. Therefore a forgetting mechanism as simple as exponential decay is normally omitted, and one has just to remember that its inclusion would reduce the critical capacity obtained by roughly 0.37. Another form of forgetting, which is potentially interesting in terms of biological plausibility, is implemented by setting limits to the range allowed for each synaptic efficacy or weight (Parisi 1986). As a particular synapse hits the upper or lower limit on its range, it is taken to be unable to further modify in the direction that would take it beyond the limit. Only after modifications in the opposite direction have taken it away from the limit does the synapse regain its full plasticity. A forgetting rule of this sort requires a slightly more complicated formal analysis (since it violates linear superposition of different memories), but it effectively results in a progressive, exponential degradation of older memories similar to that produced by straight exponential decay of synaptic weights. A combined forgetting rule that may be particularly attractive in the context of modelling synapses between pyramidal cells (the ones we usually consider throughout this book) is implemented by setting a lower limit, that is, zero, on the excitatory weight (ultimately requiring that the associated conductance be a non-negative quantity!), and allowing exponential decay of the value of the weight with time (all the way down to zero, not just to the baseline as above). Again, this type of combined forgetting rule places demands on the analytical techniques that have to be used, but leads to functionally similar effects.

C.1.5 The statistics of memories

As synaptic modification reflects, with the simplest learning rule, the firing rates of different cells at the same moment in time, the information stored in associative memories is in the form of distributions, or patterns, of firing rates. Such rates are of course the result of processing in the networks upstream of the one considered, and in that sense they do not comprise completely arbitrary sets of numbers. In fact, we have argued that interesting constraints arise, in autoassociators, from the way patterns of firing rates emerge in the learning of new

memories (Treves and Rolls 1992). Specifically, the autoassociator posited to be implemented in the CA3 region of the hippocampus may require a special preprocessing device, the dentate gyrus, that ensures the decorrelation of patterns of firing that must be stored in CA3 (see Chapter 6). Nevertheless, when considering solely the *retrieval* of information from associative memories, and not its *encoding*, one usually avoids the need to analyse how patterns of firing arise, and takes them as given, with certain statistics. In pattern associators, each firing pattern is actually a pair, with one pattern for the afferent axons and another for the output cells. We shall consider that p patterns have been stored at a given time, and label them $\mu = 1, ..., p$. Each pattern is assumed to have been generated independently of others. This is an approximation, whose validity can be checked in specific cases. Further, the firing of each cell (we use the index i for output cells and j for input ones) in each pattern is also assumed to be independent of that of other cells. This is also an approximation, which is both very handy for the analysis of mathematical models and appears to be an excellent one for some firing rates recorded in higher sensory and memory areas. Probably the agreement is due to the high number of (weak) inputs each real neuron receives, which effectively removes substantial correlations among different cells in those cases. In any case, the probability of a given firing pattern is expressed mathematically, given our independence assumption, as

$$\mathrm{P}(\{x_j^\mu\}, \{y_i^\mu\}) = \Pi_\mu \Pi_j \mathrm{P}(x_j^\mu) \Pi_i \mathrm{P}(y_i^\mu). \tag{C.4}$$

For simplicity, the probability distribution for the rates r of each cell in each pattern are usually taken to be the same, hence denoted simply as $\mathrm{P}(r)$ in the following.

By considering different forms for $\mathrm{P}(r)$ one can to some extent explore the effect of different firing statistics on the efficient coding of memories in pattern associators (the full range of possibilities being restricted by the assumptions of independence and homogeneity among cells, as explained above). As any probability distribution, $\mathrm{P}(r)$ is constrained to integrate to 1, and since it is a distribution of firing rates, it takes values above zero only for $r \geq 0$, therefore $\int_0^\infty \mathrm{d}r\, \mathrm{P}(r) = 1$. Different values for the first moment of the distribution, that is for the mean firing rate, are equivalent to simple rescalings, or, when accompanied by similar rescalings for thresholds and synaptic efficacies, are equivalent to using different units than s^{-1}. They are not therefore expected to affect storage capacity or accuracy, and in fact, when threshold-linear model units are considered, these limits on performance (which presuppose optimal values for thresholds and gain) are independent of the mean rate in $\mathrm{P}(r)$. Other aspects of the probability distribution are instead important, and it turns out that the most important one is related to the second moment of the distribution. It is convenient to define the *sparseness* (see Chapter 1) as

$$a = (\int_0^\infty \mathrm{d}r\, r\mathrm{P}(r))^2 / \int_0^\infty \mathrm{d}r\, r^2 \mathrm{P}(r) = < r >^2 \, / \, < r^2 > . \tag{C.5}$$

The sparseness ranges from 0 to 1 and it parametrizes the extent to which the distribution is biased towards zero: $a \approx 0$ characterizes distributions with a large probability concentrated close to zero and a thin tail extending to high rates, and $a \approx 1$ distributions with most of the weight away from zero.

Defined in this way, in terms of a theoretical probability distribution of firing rates assumed to be common to all cells in a given population, the sparseness can be interpreted in two complementary ways: either as measuring roughly the proportion of very active cells at any particular moment in time, or the proportion of times that a particular cell is very active (in both cases, by saying *very active* we wish to remind the reader that a actually measures the ratio of the average rate squared to the square rate averaged, and not simply the fraction of active cells or active times, which is just what the definition reduces to for binary units).

Usually, when talking about sparse coding, or, conversely, about distributed coding, one refers to an activity pattern occurring at a moment in time, and in which few or many cells fire, or fire strongly; and when instead referring to the distribution of rates of one cell at different times or, for example, in response to different stimuli, one sometimes uses the terms fine tuning, or broad tuning.

Measuring the sparseness of real neuronal activity

In analysing formal network models, especially when something like Eq. C.4 is taken to hold, the two interpretations are essentially equivalent. In the observation of real firing statistics, on the other hand, the interpretation that considers at a moment in time the firing across cells has been applied less, partly because the simultaneous recording of many cells is a relatively recent technique, partly because it is non-trivial to identify and 'count' cells that fire rarely or never, and last because the firing statistics of even similar-looking cells is expected a priori to be somewhat different due to differences in their exact physiological characteristics. What one does is thus normally to record the distribution of firing rates, say collected as the number of spikes in a fixed time window, of a particular cell across different time windows. This procedure can itself be carried out in at least two main ways. One is to record from the cell as the animal interacts with a large variety of external correlates (for example, in the case of visual stimulation, sees a movie), and to collect a large number K of time windows, for example indexed by k. The sparseness of the distribution is then

$$a = \left[\sum_k r_k/K\right]^2 / \left[\sum_k r_k^2/K\right]. \tag{C.6}$$

This particular way of measuring sparseness is affected by the length of the time window considered, in the sense that if the window is much shorter than the typical interspike interval, most of the times it will contain no spikes, and as a result the distribution will appear artificially sparse; progressively increasing the window the measured sparseness parameter, a, increases, typically saturating when the window is long enough that the above artefact does not occur. The value obtained is also obviously affected by the nature and variety of the external correlates used.

A second way to measure sparseness is to use a fixed set of S well-defined correlates, each one occurring for a number of repetitions (for example, in the case of visual stimulation, these could be a number of static visual stimuli presented each for a number of trials, with the number of spikes emitted in a given peristimulus interval quantifying the response). In this case the sparseness can be measured from the mean response rate r_s to each stimulus s, (averaged across trials with the same stimulus), as

$$a = \left[\sum_s r_s/S\right]^2 / \left[\sum_s r_s^2/S\right]. \tag{C.7}$$

This measure of sparseness is less affected by using a short time window, because the average across trials implies that non-zero mean responses can be produced even to stimuli for which most trials carry no spikes in the window; on the other hand, this measure is affected in the opposite direction by the number of correlates in the set, in the sense that the minimal value of the sparseness parameter measured in this way is just $a = 1/S$, and therefore one will find an artificially high value if using a small set.

Both artefacts can be easily controlled by measuring the effect of the size of the time window and of the size of the set of correlates. In both cases, the sparseness value obtained will nevertheless remain, obviously, a measure *relative* to those particular correlates.

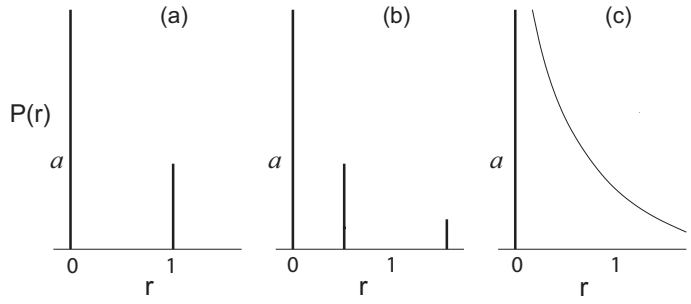

Fig. C.1 Examples of firing rate probability distributions used in the text: (a) binary; (b) ternary; (c) exponential. The ordinate shows probability, and the abscissa the firing rate. The 'a' on the ordinate referring to the sparseness is at the height of the bars that represent the fraction of active cells in (a), and of cells with low nonzero activity in (b). (From Treves and Rolls, 1991, Fig. 3.)

Specific models of firing probability distributions

One particular type of probability distribution, which arises inevitably when considering binary model units, is a binary one, with the whole weight of the distribution concentrated on just two values, one of which is usually set to zero (a non-firing unit) and the other to one (the maximal rate in arbitrary units). The sparseness a is then just the fraction of the weight at $r = 1$; in the usual mathematical notation

$$\mathrm{P}(r) = (1 - a)\delta(r) + a\delta(r - 1), \tag{C.8}$$

where the δ-functions indicate concentrated unitary weight on the value that makes their argument zero. Binary distributions (with one of the output values set at zero firing) maximize the information that can be extracted from each spike (see Appendix A2 and Panzeri, Biella, Rolls, Skaggs and Treves (1996)), among distributions with a fixed sparseness. Since the information per spike is closely related to the breadth of tuning, it is easy to see that binary distributions also minimize the breadth of tuning (again, among distributions with a fixed sparseness). In this quantitative sense, binary distributions provide what amounts to a rather precise code; on the other hand, they do not allow exploiting the graded or nearly continuous range of firing rate that a real neuron has at its disposal in coding its output message.

Other types of probability distributions that we have considered, still parametrized by their sparseness, include *ternary* ones such as

$$\mathrm{P}(r) = (1 - 4a/3)\delta(r) + a\delta(r - 1/2) + a/3\delta(r - 3/2), \tag{C.9}$$

and continuous ones such as the exponential distribution

$$\mathrm{P}(r) = (1 - 2a)\delta(r) + 4a\exp(-2r). \tag{C.10}$$

In both these distributions the first moment is conventionally set so that, as for the binary example, $< r >= a$. The ternary distribution is a conveniently simple form that allows analysis of the effects of departing from the more usual binary assumption, while the exponential form is one which is not far from what is observed in at least some parts of the brain. These three types of distribution are represented in Fig. C.1. Other types of distribution can be considered, made to model in detail experimentally observed statistics. The distributions observed for example in the primate temporal visual cortex are always unimodal, with the mode at or close to the spontaneous firing rate and sometimes at zero, and a tail extending to higher rates, which is often close to exponential (Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman

and Rolls 1997). It may be possible to understand this typical form of the firing distribution in terms of a random, normally distributed input activation, with additional normally distributed fast noise, resulting in an asymmetric rate distribution after passing through a threshold-linear input-output transform (Treves, Panzeri, Rolls, Booth and Wakeman 1999).

C.2 Quantitative analyses of performance

Having made a series of simplifying assumptions, and postponing judgement on whether each such assumption is appropriate or not when the results are applied to specific networks of real neurons, one may now evaluate the performance of pattern associators, instantiated as precisely defined mathematical models. Two aspects of their performance are to be examined: how many associations between input and output firing patterns they can store (so that each input pattern can retrieve its output one), and how accurate the retrieval of the output pattern is (which of course depends on the accuracy of the input). A third aspect, that of the time over which retrieval occurs, is not really relevant for pattern associators, because the operation of the feedforward network adds no dynamics of its own to that characterizing its constituent units.

Storage capacity and retrieval accuracy (the asymmetry in the traditional terminology hiding the fact that both aspects imply both the storage and retrieval of associations) are two faces of the same coin. This is because if more memories are stored one has poorer retrieval, that is an output with a weaker resemblance to the stored output. With nets whose dynamics leads to a steady state, such as autoassociators, the progressive deterioration is abruptly interrupted when, above a critical storage load, the network fails to evolve to a retrieval state. This is because the iterative dynamics of autoassociators converges towards one of its attractors, which can be pictured as the states lying at the bottom of valleys – their basins of attraction – with the network rolling down the slopes as if pulled by gravity. Not all of the attractors correspond to retrieval of one firing pattern, and increasing the storage load leads indeed to more retrieval attractors, but with progressively shallower valleys, until at a critical load they cease to be attractors at all; the network then ends up in states with no strong correlation to any of the stored patterns. With pattern associators that whole picture is meaningless, if anything because the starting state, that is the input firing pattern, belongs to a different panorama, as it were, from that of the output state, and there is no rolling down because the whole dynamical trajectory reduces to a single step. The absence of a critical load can still be understood intuitively, however, by realizing that by making one step from a state with some correlation to an input pattern, the network will still be within some correlation, stronger or weaker, from the corresponding output pattern. No matter how many associations are stored, which result in interfering influences on the direction of that one step, a single step will always maintain some initial correlation, if altered in magnitude. Therefore the storage capacity of a pattern associator can only be defined as that storage load which preserves, *up to a given degree*, the correlations present in the input; and since the degree of preservation (or enhancement) depends on the input correlations themselves, one usually defines capacity in terms of the constraint that the network should at least preserve *full* input correlations. In other words, this amounts to using a complete and noiseless input pattern and checking what is the maximal load such that the network still produces a noiseless output. (When using continuous output units, a strictly noiseless output is non-physical, and obviously one should set a finite tolerance level.)

C.2.1 Signal-to-noise ratios

A signal-to-noise analysis is a simple approach to analyse how retrieval accuracy depends on storage load, and on all the other characteristic parameters of the network. A pattern associator is considered to have stored $p + 1$ associations with the learning rule

$$w_{ij} = w^0 + \gamma \sum_{\mu} F(r_i^{\mu}) G(r_j^{\mu}) \tag{C.11}$$

which is a somewhat more general expression than Eq. C.3 above, with F and G arbitrary pre- and postsynaptic factors. The averages and variances of these factors over the distribution of firing rates on the input (for G) and output (for F) are denoted as

$$\begin{aligned} &< F >= m_F && < F^2 > - < F >^2 = \sigma_F^2 \\ &< G >= m_G && < G^2 > - < G >^2 = \sigma_G^2. \end{aligned} \tag{C.12}$$

The network is taken to receive in the input one of the stored input patterns, say $\mu = 1$. What each output cell receives from the C input cells that feed into it is, due to the linear superposition in the learning rule, Eq. C.11, a signal S coming from the $\mu = 1$ term in the sum, plus noise N coming from the other p stored associations. The mean value of these noise terms, plus the w^0 baseline terms, can be subtracted out by appropriate thresholds. Their variance, instead, can be evaluated by using the assumption of lack of correlations among cells in the same pattern and among patterns, yielding

$$\sigma_N^2 = p\gamma^2 C < r >^2 \{C\sigma_F^2 m_G^2 + p(1/a - 1) m_F^2 m_G^2 + (m_F^2 + \sigma_F^2)\sigma_G^2/a + (1/a - 1)\sigma_F^2 m_G^2\}. \tag{C.13}$$

The variance of the noise has to be compared with the mean square amplitude of the signal, which can be taken to be the square difference of the signal received by a cell that fires at the conventional unitary rate in the pattern being retrieved, and that of a cell that does not fire. This is

$$S^2 = \gamma^2 C^2 [F(1) - F(0)]^2 < G(r) r >^2 . \tag{C.14}$$

The argument now is that, for any given level of required retrieval accuracy, the signal must be at least of the same order of magnitude as the noise. If p is a large number, this can only happen if the first two terms in the mean square noise vanish: in that case the noise is of order $(pC)^{1/2}$ and the signal of order C, and they can be comparable up to p of order C. Clearly this requires that the mean of the presynaptic G factor be negligible: $m_G \approx 0$. If this mean is really close to zero, only the third term in the noise is substantial, and the signal-to-noise ratio is of order $(C/p)^{1/2}$. To be more precise, one may specify for the presynaptic factor the form $G(r) = r - < r >$, to find

$$S/N \approx [C(1 - a)/p]^{1/2} [F(1) - F(0)] / [< F(r)^2 >]^{1/2} \tag{C.15}$$

which indicates the potentially beneficial effect of sparseness in the output pattern: if it is sparse, and the postsynaptic factor is such that

$$[< F(r)^2 >]^{1/2} \ll [F(1) - F(0)], \tag{C.16}$$

then the number of stored associations p can be much larger than the number of inputs per unit C while preserving the signal-to-noise ratio of order one. While Eq. C.15 is valid for any form of the postsynaptic factor F and of the distribution of output firing rates, it is helpful to assume specific forms for these in order to gain an intuitive understanding of that equation.

Taking output units to be binary, with a distribution of output rates of sparseness a, as in Eq. C.8, and the postsynaptic factor to be simply proportional to the output rate, $F(r) = r$, yields

$$S/N \approx [C(1-a)/pa]^{1/2}, \tag{C.17}$$

which is of order unity (implying that the signal is strong enough to effect retrieval despite the noise) for

$$p \approx C(1-a)/a. \tag{C.18}$$

The above analysis illustrates the origin of the enhancement in storage capacity (as measured by the number of associations stored) produced by sparseness in the output distribution. The exact same effect holds for an autoassociator, in which, however, capacity is measured by the number of memory patterns stored, and the sparseness that matters is just the sparseness of the patterns, since there is no distinction between an input pattern and an output pattern being associated with it. More careful analyses have been worked out for binary units, especially in the limiting case of very sparse coding, both for autoassociators with a covariance learning rule (Tsodyks and Feigel'man 1988, Buhmann, Divko and Schulten 1989, Evans 1989) and for pattern associators or autoassociators with an optimal synaptic matrix, following the approach mentioned below (Gardner 1988). These more precise calculations enable one to extract in more detail the exact dependence of the maximum value allowed for p, on the firing sparseness, in the form

$$p \approx C/[a \ln(1/a)], \tag{C.19}$$

which is valid in the limit $a \to 0$. Eq. C.19 has been found to be valid also for autoassociators with graded response units (Treves 1990, Treves and Rolls 1991), as described in Appendix A4. For pattern associators, the lack of a well-defined notion of storage capacity, independent of retrieval quality and of cue quality, makes the approaches based on binary units not applicable to more realistic systems, as discussed earlier in this section, but the semiquantitative signal-to-noise analysis shows that the important determinants of the capacity are the same.

C.2.2 Information-theoretic measures of performance

The basic signal-to-noise analysis sketched above may be elaborated in specific cases and made more accurate, even without having to resort to binary units. To obtain more precise quantitative results, one needs to look not at the *input* to the output cells, but at the *output* of the output cells, and consider a definite measure of the performance of the system, in particular a measure of the correlation between the output firing at retrieval and the firing in the associated output pattern. The correct measures to use are information measures, as discussed in Appendix A2, and in particular for this correlation the right measure is the mutual information between the two patterns, or, expressed at the single cell level, the average information any firing rate r_R^μ which the output cell sustains during retrieval of association μ conveys on the firing rate r^μ during the storage of that association

$$I = \int \mathrm{d}r \int \mathrm{d}r_R \mathrm{P}(r, r_R) \ln_2[\mathrm{P}(r, r_R)/\mathrm{P}(r)\mathrm{P}(r_R)]. \tag{C.20}$$

In addition, one should also consider the information provided by the cue, which is to be subtracted from the above to quantify the actual 'yield' (per cell) of the memory system. Moreover, some information is also associated with the selection of the pattern being retrieved. This last quantity is at most (when the *correct* pattern is always selected) of order $\log(p)$ and therefore usually much smaller than that reflecting the full correlation in the retrieved pattern, or that in the cue, which are both quantities expected to be proportional to the number of

cells involved. However, there are cases (Frolov and Murav'ev 1993) when binary units are used with very sparse coding, in which all such quantities could be of similar magnitude; and, more importantly, it is also true that what can be measured in neurophysiological experiments is, in practice, the information associated with the selection of the output pattern, not the one expressed by Eq. C.20 (since the firing rates recorded are typically only those, as it were, at retrieval, r_R).

In general, one is interested in the value of I, in the equation above, averaged over a large number of variables whose specific values are irrelevant: for example, the firing rates of both output and input cells in all the other stored associations. One is therefore faced with the computational problem of calculating the average of a logarithm, which is not trivial. This is where mathematical techniques imported from theoretical physics become particularly useful, as a body of methods (including the so-called replica method) have been developed in statistical physics specifically to calculate averages of logarithms. These methods will be briefly sketched in Appendix A4, on autoassociators, but the motivation for resorting to them in the quantitative analysis of formal models is essentially identical in the case of pattern associators. An example of the application of the replica method to the calculation of the information retrieved by a pattern associator is given by Treves (1995). In that case, the pattern associator is in fact a slightly more complex system, designed as a model of the CA3 to CA1 Schaffer collateral connections, and cells are modelled as threshold-linear units (this is one of the few calculations carried out with formal models whose units are neither binary nor purely linear). The results, which in broad terms are of course consistent with a simple signal-to-noise analysis, detail the dependence of the information gain on the various parameters describing the system, in particular its degree of plasticity, as discussed in Chapter 6.

C.2.3 Special analyses applicable to networks of binary units

Networks of binary units, which of necessity support the storage and retrieval of binary firing patterns, have been most intensely studied both with simulations and analytically, generating an abundance of results, some of which have proved hard to generalize to more realistic nets with continuously graded units. One of the peculiarities of memories of binary units operating with binary patterns is that one can define what one means by precise, exact retrieval of a memorized pattern: when all units are in the correct 0 or 1 state, bit by bit. This greatly helps intuition, and we have used this fact in the illustratory examples of Chapter 2, but it may also generate misconceptions, such as thinking that 'correct retrieval', a notion which makes sense in the technological context of digital transmission of information, is a meaningful characterization of the operation of realistic, brain-like networks.

The prototypical pattern associator, or Associative Net, originally introduced by Willshaw, Buneman and Longuet-Higgins (1969), and often referred to as the Willshaw net; and the slightly different version used at about the same time in the extensive analyses by the late David Marr (1969, 1970, 1971), are both rather extreme in their reliance on binary variables, in that not only the processing units, but also the synaptic weights are taken to be binary. This is in contrast with the models considered by Little (1974) and Hopfield (1982), and analysed mathematically by Amit, Gutfreund and Sompolinsky (1985, 1987), in which the units are taken to be binary, but the synaptic weights, which are determined by a Hebbian covariance rule, are in principle real-valued. The learning rule employed in the Willshaw net may be considered a clipped version of a Hebbian, but non-LTD-balanced, learning rule. A synapse in the Willshaw associator can be in one of two states only, off or on, and it is on if there is at least one input-output association in which both pre- and postsynaptic units are on, and it is off otherwise. If a is, again, the average proportion of units on in both the input and output

patterns, the number of associations that can be stored and retrieved correctly is

$$p \approx 1/a^2 \tag{C.21}$$

which can be understood intuitively by noting that the probability that a particular synapse be on is, from simple combinatorics,

$$\mathrm{P}(w_{ij} = 0) \quad = (1 - a^2)^p \quad \approx \exp -(pa^2) \tag{C.22}$$

and that this probability must be of order unity for the network to be in a regime far from saturation of its potential for synaptic modification. A detailed analysis of the Willshaw net (interpreted as an autoassociator) has been carried out by Golomb, Rubin and Sompolinsky (1990) using the full formal apparatus of statistical physics. A particularly insightful analysis is that of Nadal and Toulouse (1990), where they compare a binary pattern associator operating with a purely incremental, Hebbian rule and one operating with a clipped version of this rule, that is a Willshaw net. The interesting message to take home from the analysis is that, in the regime of very sparse patterns in which both nets operate best, in terms of being able to retrieve many associations, the clipping has a very minor effect, resulting in a slightly inferior capacity. In terms of information, if the pattern associator with graded synapses is shown to utilize up to $1/(2 \ln 2) \simeq 0.721$ bits per synapse, the Willshaw net is shown to utilize $\ln 2 \simeq 0.693$ bits per synapse, with a reduction of only 4% (Nadal 1991, Nadal and Toulouse 1990). This can be converted into a statement about the precision with which a biological mechanism such as LTP would have to set synaptic weights in associative learning, that is, it would not have to be very precise. In general, when a learning rule that includes heterosynaptic LTD is used, it remains true that not much resolution is required in the synaptic weights to achieve near optimal storage capacity (cf. Sompolinsky (1987). Having very few bits available produces only a moderate decrease in the number of associations that can be stored or in the amount of information stored on each synapse. In the particular case of the Willshaw type of learning rule, however, it should be noted that the near equivalence between binary-valued and continuously-valued synapses only holds in the very extreme sparse coding limit. Moreover, in the Willshaw type of network a price for being allowed to use weights with only 1 bit resolution is that the thresholds have to be set very precisely for the network to operate successfully.

The way in which the binary nature of the units explicitly enters the calculation of how many associations can be stored in a binary pattern associator is typically in finding the best value of the threshold that can separate the input activations that should correspond to a quiescent output from those that should correspond to a firing output. (Differences in specifications arise in whether this threshold is allowed to vary between different units, and in the course of learning, or rather it is held fixed, and/or equal across output units.) In a Willshaw net if the input is exactly one of the stored patterns, the activation of the output units that should be 'on' is exactly the same (a distribution of width zero), because all the synapses from active input units are themselves in the 'on' state, and the number of active inputs is, at least in the original Willshaw net, fixed exactly. This allows setting the threshold immediately below this constant activation value, and the only potential for interference is the extreme tail of the distribution of activation values among output units that should be 'off' in the pattern. Only if at least one of these output units happens to have all the synaptic weights from all active inputs on, will an error be generated. Thus one factor behind the surprisingly good capacity achieved by the Willshaw net in the very sparse coding regime is its exploiting the small amount of probability left at the very tail of a binomial distribution (the one giving the probability of being 'on' across all synapses of the same output unit). This condition allowing relatively large capacity would seem not to carry over very naturally to

more realistic situations in which either synaptic values or firing rates are taken to have not fully binary distributions, but rather continuous distributions with large widths around each peak. (In fact, realistic distributions of values for both synaptic efficacies and firing rates are more likely to be unimodal, and not to show any more than one peak.) A quantitative analysis of the Willshaw net, adapted to include finite resolution widths around two dominant modes of synaptic values, has never been carried out, but in any case, binary or nearly binary-valued synaptic weights would be expected to function reasonably only in the limit of very sparse coding. This can be understood intuitively by noting again that with more distributed coding, any particular synaptic value would have a fair chance of being modified with the learning of each new association, and it would quickly become saturated. Even if synaptic values are allowed to modify downwards, the availability of only two synaptic levels implies that few successive associations are sufficient to 'wash out' any information previously stored on a particular synapse, unless the coding is very sparse.

The Gardner approach

The approach developed by the late Elizabeth Gardner (1987, 1988) represented an entirely novel way to analyse the storage capacity of pattern associators. Instead of considering a model network defined in terms of a given learning rule, she suggested considering the 'space' of all possible networks, that is of all possible synaptic matrices, independently of their having been produced, or not, by a learning mechanism, and selecting out among all those the ones that satisfied certain constraints. For a pattern associator with binary units, these constraints can be specified by requiring that for each association the net is taken to have stored, the activation resulting from the presentation of the correct input pattern be above threshold for those output units that should be 'on', and below threshold for those output units that should be 'off'. The average number of synaptic matrices that satisfy these requirements, for each of a number p of stored associations, can be literally counted using statistical physics techniques, and the maximum p corresponds to when this average number of 'viable' nets reduces to one. Such an approach, which is too technical to be described more fully here (but an excellent account is given in the book by Hertz, Krogh and Palmer (1991)) has been generalized in an endless variety of ways, generating a large body of literature. Ultimately, however, the approach can only be used with binary units, and this limits its usefulness in understanding real networks in the brain.

Nevertheless, the Gardner calculations, although strictly valid only for binary units, are often indicative of results that apply to more general models as well. An example is the increase in storage capacity found with sparse coding, Eq. C.19 above, which is a result first obtained by Gardner (1987, 1988) without reference to any specific learning rule, later found to hold for fully connected autoassociators with Hebbian covariance learning (Tsodyks and Feigel'man 1988, Buhmann et al. 1989), then to apply also to autoassociators with sparse connectivity (Evans 1989), and subsequently generalized (modulo the exact proportionality factor) to non-binary units and non-binary patterns of activity (e.g. Treves and Rolls (1991). A point to be noted in relation to the specific dependence of p on a, expressed by Eq. C.19, is that the original Gardner (1988) calculation was based on steady-state equations that would apply equally to a pattern associator and to an autoassociator, and hence did not distinguish explicitly between the sparseness of the input pattern and that of the output pattern. In the Hertz, Krogh and Palmer (1991) account of the same calculation, a different notation is used for input and output patterns, but unfortunately, when sparseness is introduced it mistakenly appears that it refers to the input, whereas by repeating the calculation one easily finds that it is the sparseness of the output pattern that matters, and that enters Eq. C.19. One can also be convinced that this should be the case by considering a single binary output unit (the exact calculation is in fact carried out independently for each output unit) and what the learning of

different associations entails for the connection weights to this unit. If the output statistics is very sparse, the unit has to learn to respond (that is to be in the 'on' state) to very few input patterns, which can be effected simply by setting a large threshold and the weights appropriate for a superposition of a few AND-like functions among the input lines (the functions would be AND-like and not AND proper in the sense that they would yield positive output for a specific conjunction of 'on' and 'off', not just 'on', inputs).

The crucial point is that just superimposing the weights appropriate to each AND-like operation only works if the functions to be superimposed are few, and they are few if the output statistics is sparse. The full calculation is necessary to express this precisely in terms of the $[a \ln(1/a)]^{-1}$ factor.

The Gardner approach has been fruitful in leading to many results on networks with binary output units (see Hertz, Krogh and Palmer (1991)).

References

Abeles, M., Bergman, H., Margalit, E. and Vaadia, E. (1993). Spatiotemporal firing patterns in the frontal cortex of behaving monkeys, *Journal of Neurophysiology* **70**: 1629–1638.

Amit, D. J. (1989). *Modeling Brain Function. The World of Attractor Neural Networks.*, Cambridge University Press, Cambridge.

Amit, D. J., Gutfreund, H. and Sompolinsky, H. (1985). Spin-glass models of neural networks, *Physical Review, A* **32**: 1007–1018.

Amit, D. J., Gutfreund, H. and Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation, *Annals of Physics (New York)* **173**: 30–67.

Baddeley, R. J., Abbott, L. F., Booth, M. J. A., Sengpiel, F., Freeman, T., Wakeman, E. A. and Rolls, E. T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes, *Proceedings of the Royal Society B* **264**: 1775–1783.

Bliss, T. V. P. and Lømo, T. (1973). Long lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimulation of the perforant path, *Journal of Physiology* **232**: 331–356.

Buhmann, J., Divko, R. and Schulten, K. (1989). Associative memory with high information content, *Physical Review A* **39**: 2689–2692.

Carnevale, N. T. and Johnston, D. (1982). Electrophysiological characterization of remote chemical synapses, *Journal of Neurophysiology* **47**: 606–621.

Evans, M. R. (1989). Random dilution in a neural network for biased patterns, *Journal of Physics A* **22**: 2103–2118.

Frolov, A. A. and Murav'ev, I. P. (1993). Informational characteristics of neural networks capable of associative learning based on Hebbian plasticity, *Network* **4**: 495–536.

Gardner, E. (1987). Maximum storage capacity in neural networks, *Europhysics Letters* **4**: 481–485.

Gardner, E. (1988). The space of interactions in neural network models, *Journal of Physics A* **21**: 257–270.

Golomb, D., Rubin, N. and Sompolinsky, H. (1990). Willshaw model: associative memory with sparse coding and low firing rates, *Physical Review A* **41**: 1843–1854.

Gray, C. M., Konig, P., Engel, A. K. and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties, *Nature* **338**: 334–337.

Hebb, D. O. (1949). *The Organization of Behavior: a Neuropsychological Theory*, Wiley, New York.

Hertz, J., Krogh, A. and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*, Addison Wesley, Wokingham, U.K.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci. USA* **79**: 2554–2558.

Koch, C., Bernander, O. and Douglas, R. J. (1995). Do neurons have a voltage or a current threshold for action potential initiation?, *Journal of Computational Neuroscience* **2**: 63–82.

Little, W. A. (1974). The existence of persistent states in the brain, *Mathematical Bioscience* **19**: 101–120.

Marr, D. (1969). A theory of cerebellar cortex, *Journal of Physiology* **202**: 437–470.

Marr, D. (1970). A theory for cerebral cortex, *Proceedings of The Royal Society of London, Series B* **176**: 161–234.

Marr, D. (1971). Simple memory: a theory for archicortex, *Philosophical Transactions of the Royal Society, London [B]* **262**: 23–81.

Nadal, J. P. (1991). Associative memory: on the (puzzling) sparse coding limit, *Journal of Physics A* **24**: 1093–1102.

Nadal, J. P. and Toulouse, G. (1990). Information storage in sparsely coded memory nets, *Network* **1**: 61–74.

Nadal, J. P., Toulouse, G., Changeux, J. P. and Dehaene, S. (1986). Networks of formal neurons and memory palimpsests, *Europhysics Letters* **1**: 535–542.

Optican, L. M. and Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: III. Information theoretic analysis, *Journal of Neurophysiology* **57**: 162–178.

Panzeri, S., Biella, G., Rolls, E. T., Skaggs, W. E. and Treves, A. (1996). Speed, noise, information and the graded nature of neuronal responses, *Network* **7**: 365–370.

Parisi, G. (1986). A memory which forgets, *Journal of Physics A* **19**: L617–L619.

Rall, W. and Shepherd, G. M. (1968). Theoretical reconstruction of field potentials and dendrodendritic synaptic interactions in olfactory bulb, *Journal of Neurophysiology* **31**: 884–915.

Rolls, E. T. (2008). *Memory, Attention, and Decision-Making. A Unifying Computational Neuroscience Approach*, Oxford University Press, Oxford.

Rolls, E. T. (2012). *Neuroculture: On the Implications of Brain Science*, Oxford University Press, Oxford.

Rolls, E. T. (2014). *Emotion and Decision-Making Explained*, Oxford University Press, Oxford.

Rolls, E. T. (2016). *Cerebral Cortex: Principles of Operation*, Oxford University Press, Oxford.

Rolls, E. T. and Deco, G. (2002). *Computational Neuroscience of Vision*, Oxford University Press, Oxford.

Rolls, E. T. and Deco, G. (2010). *The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function*, Oxford University Press, Oxford.

Rolls, E. T. and Treves, A. (1998). *Neural Networks and Brain Function*, Oxford University Press, Oxford.

Rolls, E. T., Critchley, H. D. and Treves, A. (1996). The representation of olfactory information in the primate orbitofrontal cortex, *Journal of Neurophysiology* **75**: 1982–1996.

Rolls, E. T., Treves, A. and Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex, *Experimental Brain Research* **114**: 149–162.

Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing*, Vol. 1: Foundations, MIT Press, Cambridge, MA.

Shepherd, G. M. (1988). *Neurobiology*, New York, Oxford University Press.

Sompolinsky, H. (1987). The theory of neural networks: The Hebb rule and beyond, *Heidelberg colloquium on glassy dynamics*, Springer, pp. 485–527.

Tovee, M. J. and Rolls, E. T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex, *Visual Cognition* **2**: 35–58.

Tovee, M. J., Rolls, E. T., Treves, A. and Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex, *Journal of Neurophysiology* **70**: 640–654.

Treves, A. (1990). Graded-response neurons and information encodings in autoassociative memories, *Physical Review A* **42**: 2418–2430.

Treves, A. (1995). Quantitative estimate of the information relayed by the Schaffer collaterals, *Journal of Computational Neuroscience* **2**: 259–272.

Treves, A. and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain?, *Network* **2**: 371–397.

Treves, A. and Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network, *Hippocampus* **2**: 189–199.

Treves, A., Panzeri, S., Rolls, E. T., Booth, M. and Wakeman, E. A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli, *Neural Computation* **11**: 601–631.

Tsodyks, M. V. and Feigel'man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level, *Europhysics Letters* **6**: 101–105.

Tsodyks, M. V. and Sejnowski, T. (1995). Rapid state switching in balanced cortical network models, *Network* **6**: 111–124.

Willshaw, D. J., Buneman, O. P. and Longuet-Higgins, H. C. (1969). Non-holographic associative memory, *Nature* **222**: 960–962.