Communicated by Mikhail Tsodyks

Stable and Rapid Recurrent Processing in Realistic Autoassociative Memories

Francesco P. Battaglia
Alessandro Treves
SISSA - Cognitive Neuroscience, Trieste, Italy

It is shown that in those autoassociative memories that learn by storing multiple patterns of activity on their recurrent collateral connections, there is a fundamental conflict between dynamical stability and storage capacity. It is then found that the network can nevertheless retrieve many different memory patterns, as predicted by nondynamical analyses, if its firing is regulated by inhibition that is sufficiently multiplicative in nature. Simulations of a model network with integrate-and-fire units confirm that this is a realistic solution to the conflict. The simulations also confirm the earlier analytical result that cued-elicited memory retrieval, which follows an exponential time course, occurs in a time linearly related to the time constant for synaptic conductance inactivation and relatively independent of neuronal time constants and firing levels.

1 Introduction

Autoassociative networks, or free simple memories in David Marr's terms (Marr, 1971), have been considered one of the fundamental building blocks of brain function (Little, 1974; Kohonen, 1977; Hopfield, 1982). In an autoassociative memory, all the components of a distributed representation of a memory item are associated together by Hebbian (Hebb, 1949) synaptic plasticity, enabling the modified synaptic matrix to retrieve the whole representation when some of the components are presented later as a cue. The difference with heteroassociative memories (Willshaw, Buneman, & Longuet-Higgins, 1969) is that in the latter, a representation of item X, or part of it, is used as a cue to retrieve the representation of a different item, Y. The representation of X is distributed over the input lines to a group of units, the output lines of which encode Y. In autoassociative memories, the item is the same, but the degree to which there is a differentiation between input and output may vary across possible autoassociative architectures. Two limiting cases are useful to illustrate the spectrum of possibilities (Treves & Rolls, 1991).

In a purely recurrent network, the output lines from a group of units have recurrent collateral branches that provide inputs to the same units. One can think of each memory item as having a unique representation, defined

as the pattern of activity distributed across the units. A representation is retrieved by repeatedly circulating activity through the recurrent loops until a steady state is reached. The existence, stability, and characteristics of steady states can be analyzed mathematically with self-consistent statistics (Amit, Gutfreund, & Sompolinsky, 1987).

A different architecture consists of several (L) groups of units in cascade, with purely feedforward connections from one group to the next one. The retrieval operation is, as it were, unwrapped along the L stages, with the ongoing pattern of activity at each stage increasingly approaching the target representation. Independent representations of the same memory item exist at each stage. Since the system is feedforward, simpler statistics are sufficient to analyze its operation (Domany, Kinzel, & Meir, 1989).

Many intermediate possibilities exist, of course, but already to a very abstract level, one can point to three advantages that favor architectures closer to the recurrent than to the feedforward limit: (1) the number of units and connections required is reduced by a factor L; (2) if feedback reverberation is sufficient, it can sustain the activation of a representation over time, endowing the network with an additional capability for short-term memory (Amit, 1995); and (3) it is easier to store each item in the memory by forming the one required representation rather than L representations over L groups of units. A disadvantage of architectures dominated by feedback is that they suffer more from crosstalk, or interference, when many items are stored on the same connections; this disadvantage disappears if the coding becomes sparse, as revealed by analyses of the storage capacity (Tsodyks & Feigelman, 1988; Treves & Rolls, 1991).

These characterizations of the advantages of different architectures, which have been derived using simple formal models, are important in understanding the autoassociative function that may be served by real networks in the brain. When considering autoassociative memory as implemented in the brain, two additional aspects need to be studied that cannot be fully understood using models that are too simple.

The first aspect is the time required for the retrieval of a representation. In the simpler models, time is typically discretized into time steps. If one then contrasts a recurrent network, in which a representation is retrieved (to a required degree of accuracy) after L sweeps through the single group of units, with a multilayer feedforward network consisting of L stages activated in succession, the time required for retrieval is apparently the same. Obviously such a description has little to do with the dynamics of real neurons, and trying to construct a biophysical equivalent of the time step (e.g., the membrane time constant, or the typical interspike interval) does not lead to any real understanding. What is needed at the least is a study of formal models based on a description of real neurons as integrate-and-fire units (Lapique, 1907; Eccles, 1957) and of real synaptic transmission as conductance changes (Eccles, 1964). An analysis of the dynamics of an autoassociative recurrent network model built with such components has yielded part of the answer

as to the time scales for retrieval (Treves, 1993). The partial answer is an analytical formula for the time constants of the exponentially decaying transient modes, through which firing activity in the network approaches the firing at steady state. There are many different modes, each of which has a time constant with a real part, describing the rate of decay of the mode, and an imaginary part, specifying the frequency of the oscillations accompanying the decay. An important family of transients has the real part of the time constant determined by the rate of inactivation of the synaptic conductances opened by activity on the recurrent collaterals. Since such a rate of inactivation in the brain is typically short (10–20 msec, even when taking into account the dendritic spread not included explicitly in the integrate-and-fire description (Hestrin, Nicoll, Perkel, & Sah, 1990; Colquhoun, Jonas, & Sakmann, 1992; McBain & Dingledine, 1992), a prediction arising from the analysis is that the contribution of recurrent collaterals to the retrieval of a memory representation may take place in a relatively short time, over a few tens of milleseconds, independent of the prevailing firing rates and the membrane time constants, however defined, of the neurons in the population (Treves, Rolls, & Tovee, 1996). The analysis, however, has remained incomplete, because it describes only the modes close to steady state and not the full dynamics from an arbitrary initial state and because it is unable to tell to what extent each individual mode will be activated when the activity evolves from any initial state. These limitations can be overcome by computer simulations of the same network model considered by the analytical treatment.

A second aspect that has to be addressed by models that aim to be applicable to the real brain is that of the stability of the steady states that are taken to correspond to memory retrieval. As with any other steady state in the dynamics of a system of many units, there are many possible sources of instability. One example is the instability of the steady states in which the firing of different units is asynchronous, to synchronization among groups of units (Tsodyks, Mitcov, & Sompolinsky, 1993; Deppisch et al., 1993; van Vreeswijk, Abbott, & Ermentrout, 1994; Hansel, Mato, & Meunier, 1995). A more basic potential instability arises out of the fact that the Hebbian modifiable connections that are thought to mediate associative memory in the brain are those between pyramidal excitatory cells. Therefore, a recurrent autoassociative memory is in itself a positive feedback circuit, and unless its activity can be tightly controlled by appropriate inhibition, it will tend to explode. Although the stability of realistic networks of excitatory and inhibitory units has been studied (Abbott, 1991), it was not in the context of autoassociative memories. In this article, we show that in such networks there is a fundamental conflict between stability to excitatory explosion and storage capacity. In the next section, we show that the conflict can be avoided by inhibition that is predominantly multiplicative in nature. Then we return to the issue of the time scales for retrieval, with simulations that support and qualify the analytical predictions. The last section discusses

the implications of these results for the operations of associative memories in the brain. A brief report of this work appeared in Battaglia and Treves (1996).

2 The Stability-Capacity Conflict

A full analysis of the stability of asynchronous steady firing states must be carried out using appropriately detailed models, but the requirements for stability against excitatory runaway reverberations can be discussed using a simple two-variable model. In such a model, two variables, ν_E and ν_I, describe, respectively, the average firing rates of excitatory and inhibitory units, which approach their steady-state values with time constants τ_E and τ_I. The steady-state values are determined by these average firing rates and the level of afferent inputs. If we assume that, above threshold, the dependence is approximately linear, the dynamical system can be written (Wilson & Cowan, 1972):

$$\tau_E \dot{\nu_E} = -\nu_E + J_E^E \nu_E - J_E^I \nu_I + \nu_E^{aff} \tag{2.1}$$

$$\tau_I \dot{\nu_I} = -\nu_I + J_I^E \nu_E - J_I^I \nu_I + \nu_I^{aff}, \tag{2.2}$$

where the J's are the adimensional effective couplings (signs are chosen so that they are all positive in value) between the dynamical variables, as they emerge, essentially, from averaging synaptic strengths across pairs of active units, and $\nu^{aff}\nu_{E,I}^{aff}$ are constant terms, which depend on the afferent input activity and are proportional to fixed-point rates. They ensure that equilibrium rates are not zero, even if the network does not receive any input, reflecting the capability of the network to self-sustain its activity. If this system of equations has a fixed point, its stability requires that

$$\mathrm{Tr} = (J_E^E - 1)/\tau_E - (J_I^I + 1)/\tau_I < 0 \tag{2.3}$$

$$\mathrm{Det} = -(J_E^E - 1)(J_I^I + 1) + J_I^E J_E^I > 0. \tag{2.4}$$

Both inequalities can be satisfied for arbitrary values of the mean excitatory-excitatory coupling among active units, J_E^E, provided inhibitory couplings are strong enough to control excitation. If, on the basis of this simple two-variable model, we want to ensure the stability of a real autoassociator, both inequalities must be satisfied with ample margins. The reason is that exactly which units are active will be highly variable, and therefore the effective value of J_E^E at any moment in time will fluctuate substantially. It is easy to realize, then, that for values of the four mean couplings much larger than 1, the determinant appearing in the second condition will be of the order of such large value, squared. Now, the fixed-point firing rates are

$$\nu_E^{fp} = \frac{(J_I^I + 1)\nu_E^{aff} - J_E^I \nu_I^{aff}}{\mathrm{Det}} \tag{2.5}$$

$$\nu_I^{fp} = \frac{J_I^E \nu_E^{aff} - (J_E^E - 1)\nu_I^{aff}}{\mathrm{Det}}, \tag{2.6}$$

which means that if the couplings are large, under conditions of robust stability the mean excitatory firing rate at the fixed point will be much lower than the one determined by afferent inputs alone, $\nu_E^{fp} \ll \nu_E^{aff}$. However, this is incompatible with the effective operation of the network as content-addressable memory, since it makes recurrent processing minor with respect to the feedforward relay of the cue. In fact, when we tried to simulate memory retrieval with large couplings and at the same time insisted on the condition that local intrinsic inputs dominate over external afferent inputs (a condition intended to mimick the observed cortical anatomy; Abeles, 1991), we always ran into large oscillations (Simmen, Treves, & Rolls, 1996), due to even transient imbalances between local excitation and inhibition, which resulted in large fluctuations in the effective couplings, and prevented the network from reaching a steady retrieval state. Only by using as a cue the nearly complete memory pattern could we effect proper retrieval, but then recurrent connections played only a minor role. Therefore, to obtain robust, stable, fixed points, we had to restrict ourselves to smaller effective couplings, in particular to values of J_E^E not much above 1. In that case, since the excitatory self-coupling always appears in the combination $(J_E^E - 1)$, its potentially devastating influence on the stability of the fixed point will be reduced, and at the same time conditions will exist under which even small cues will be sufficient to initiate retrieval. Keeping the excitatory self-coupling low, however, conflicts with ensuring a large storage capacity.

Consider a simple autoassociator in which the weights of the connections among the units are determined by a linear sum of Hebbian-modification terms, as, for example, in the Hopfield model (Hopfield, 1982). If the units represent excitatory cells and the weights ultimately correspond to conductances, one may assume that such a memory structure is superimposed on a baseline connection weight that is large enough to keep positive even the individual weights that happen to undergo more negative modifications.[1] Therefore, one may write for the weight between units i and j,

$$w_{ij} = w^0 + \frac{1}{C}\sum_{\mu=1}^{p}\left(\frac{\eta_i^\mu}{\langle\eta\rangle} - 1\right)\left(\frac{\eta_j^\mu}{\langle\eta\rangle} - 1\right), \tag{2.7}$$

where η_i^μ is the firing rate of unit i in the μth memory pattern, $\langle\eta\rangle$ is the average firing rate, the network stores p patterns with equal strength, and C is the number of inputs per unit. The specific (covariance) form of the Hebbian term and the normalization factor are inessential to the argument

[1] This assumption is made for the sake of clarity. In the simulations that follow, we use an equivalent formulation, although it is less transparent to the analysis.

that follows and were chosen for consistency with previous analyses (Treves & Rolls, 1991). The minimum connection weights will be those between pairs in which the pre- and postsynaptic unit happen to be anticorrelated across all patterns; that is, whenever one of the two is firing, for example, at a typical elevated rate η^*, the other is quiescent. Then the condition that ensures that the underlying conductance remains positive, even in such cases, reads

$$w^0 \geq \frac{p}{C}\frac{\eta^*}{\langle\eta\rangle}. \tag{2.8}$$

On the other hand, the effective excitatory self-coupling—that is, the effect that the average excitatory firing rate exerts on each excitatory unit—is given by summing conductances across input lines and multiplying by the gain γ characterizing the unit's input-output transform in a linear range above firing threshold,

$$J_E^E = \gamma C w^0. \tag{2.9}$$

Note that the Hebbian terms average to zero when summing across the C inputs. Previous analyses (Treves, 1990; Treves & Rolls, 1991) have shown that for the network to be able to retrieve memory patterns, the gain has to be sufficiently strong, as expressed by the condition

$$\gamma \geq \frac{a}{(1-a)}, \tag{2.10}$$

where $0 < a < 1$ is the sparseness of the firing patterns, defines as $a = \langle\eta\rangle^2/\langle\eta^2\rangle$ (Treves, 1990). Putting together now the condition that the effective excitatory self-coupling be at most of order 1 with the last three equations, one realizes why stability conflicts with storage capacity:

$$O(1) \approx J_E^E = \gamma C w^0 \geq p\frac{\eta^*}{\langle\eta\rangle}\frac{a}{(1-a)}; \tag{2.11}$$

that is, in this case, to be stable at retrieval, the network must not store more than a number of memory patterns,

$$p_{max} \simeq \frac{\langle\eta\rangle}{\eta^*}\frac{(1-a)}{a} = O(1)!, \tag{2.12}$$

that is, more than a handful of patterns. In simulations that followed these very specifications, we found it difficult to obtain retrieval in nets storing more than two or three patterns, whatever their size. The conflict arises out of requiring simultaneously dynamical stability and effective retrieval

ability and biological plausibility (in that the memory is stored on the connections between excitatory units and in that each conductance must be a positive quantity). It does not arise in storage capacity analyses based on simplified formal models (Amit et al., 1987; Treves & Rolls, 1991) if one treats connection weights as real variables that can have either sign and can change in sign as more memories are stored.

Recurrent autoassociative memory models based on an alternative simple "learning rule"—the so-called Willshaw models (Willshaw et al, 1969)—although assuming only positive (or zero) weights among excitatory units, still suffer from similar limitations. That class of models, however, is more difficult to treat analytically (Golomb, Rubin, & Sompolinsky, 1990) and does not lend itself to such a simple discussion of the conflict; moreover, what is limited is not simply p, the number of memories that can be stored (which can be well above two or three; Amit & Brunel, 1997), but the total amount of information that can be stored and retrieved, which is proportional to p but also decreases the sparser are memory patterns (and the more information need be provided with the cue).

3 Realistic Inhibition May Avoid the Conflict

A seemingly innocuous assumption that was made in writing equations 2.1 and 2.2 is that excitatory firing rates depend linearly not just on themselves but also, through a separate linear term, on inhibitory rates. This is equivalent to considering what is sometimes called subtractive inhibition. Purely subtractive inhibition is a convenient model for $GABA_B$ inhibition, which acts through K^+ channels of limited total conductance, primarily by hyperpolarizing the receiving cell (Connors, Malenka, & Silva, 1988). If colocated on dendrites along with excitatory inputs, $GABA_B$ can be thought of as providing an additional term that is negative in sign and hence subtractive, and occurs on a slower time scale (Hablitz & Thalmann, 1987).

$GABA_A$ inhibition, which is responsible for fast inhibitory control of the activity level of recurrent networks (Miles & Wong, 1987), is sometimes referred to as multiplicative (or, rather, divisive) in nature. This is because it acts via Cl^- channels of relatively large total conductance (Connors et al., 1988) and inversion potential not far below the resting potential; hence, its effect is more shunting than hyperpolarizing. If located on proximal dendritic branches or on the soma (Andersen, Eccles, & Loyning, 1964), it can be modeled to a first approximation as producing a division of the current resulting from more distal inputs (Abbott, 1991).

Purely multiplicative inhibition acting on excitatory cells would lead to substitute equation 2.1 with

$$\tau_E \dot{v_E} = -v_E + J_E^E(v_I)v_E + v_E^{aff}, \tag{3.1}$$

that is, the excitatory self-coupling is now a function of the average firing

rate of inhibitory units (the second part of the equation can be modified as well, but this is irrelevant for the present discussion). To the extent that afferent inputs are absent or negligible, at the fixed point the self-coupling takes the value 1, thereby automatically ensuring stability, at least in the sense of equations 2.3 and 2.4 (since the terms in $J_E^E - 1$ disappear from the inequalities). Real inhibition, of course, is not purely multiplicative; however, the situation holding in this limit clarifies that under appropriate conditions (if inhibition is multiplicative to a sufficient degree), the stability of recurrent networks against runaway excitation is automatically guaranteed.

As for the upper limit on storage capacity, we have checked, by repeating previous analyses (Treves & Rolls, 1991) of recurrent associative memories of threshold-linear units with a gain γ now dependent on the average inhibitory rate, that the same, exact equations determine the storage capacity. Such a result stems from the fact that by acting on the gain, inhibition now keeps the effective J_E^E entering the stability analysis close to 1, but it leaves identical the capacity equations, as the analytical treatment shows. This confirms that the form of inhibition used has no effect on such absolute limit (a limit that with subtractive inhibition was far beyond what could be achieved in practice). We have also carried out simulations of a simple network model with 3000 to 5000 threshold-linear units as used in the analytical calculation, at several sparseness values. We estimated storage capacity from the simulations by progressively increasing memory load and determining the critical level at which no retrieval of any stored pattern was possible. Results are shown in Figure 1, and confirm the analytical prediction, which is the exact reproduction of previous analyses with subtractive inhibition (Treves & Rolls, 1991). Note that a value of the storage parameter $\alpha = 0.3$, for example, corresponds to 900 stored patterns.

We have carried out simulations of a more detailed network model with spiking units and conductance-based synaptic action, to understand whether realistic inhibition still allows retrieval of more than two or three patterns (the limit we had on similar simulations with purely subtractive inhibition) and, once disposed of this limitation, to address anew, in a realistic context, the issue of the time scales for recurrent memory retrieval.

4 Simulations Show Stability and Fast Retrieval

The simulated network consisted of $N_{ex} = 800$ excitatory units and $N_{in} = 200$ inhibitory ones. Each integrate-and-fire unit represents a neuron as a single-branch, compartmented dendrite through which the cell receives all its input, and a pointlike soma, where spikes are generated. Though very simple, the compartmental model is still computationally demanding and severely limits the size of the network that we could implement on a Linux workstation. The current flowing from each compartment to the external

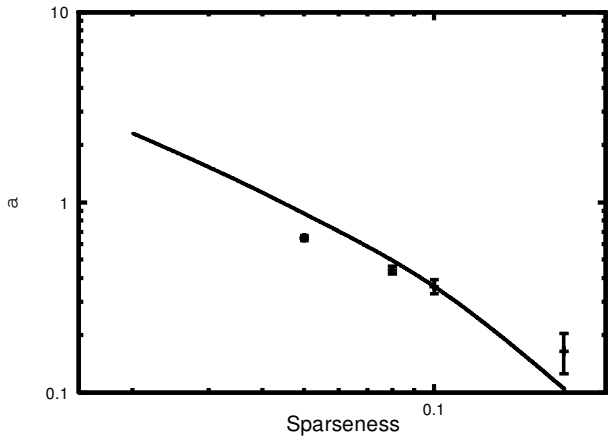

Figure 1: Simulation results for the capacity of a network of 3000 threshold-linear neurons (5000 for $a = 0.05$) are compared with the theoretical prediction (solid line) at different values of the sparseness a. The prediction arises from equations identical to those found by Treves (1990).

medium is written as

$$I(t) = g_{leak}(V(t) - V^0) + \sum_j g_j(t)(V(t) - V_j), \tag{4.1}$$

where g_{leak} is a constant, passive leakage conductance, V^0 the membrane resting potential, $g_j(t)$ the value of the jth synapse conductance at time t, and V_j the reversal potential of the jth synapse. $V(t)$ is the potential in the compartment at time t. Synaptic conductances have an exponential decay time behavior, obeying the equation

$$\frac{dg_j}{dt} = -\frac{g_j}{\tau_j} + \Delta g_j \sum_k \delta(t - t_k^j), \tag{4.2}$$

where τ_j is the synaptic decay time constant and Δg_j is the amount the conductance is increased when the presynaptic unit fires a spike. Δg_j thus represents the (unidirectional) coupling strength between the presynaptic and the postsynaptic cell. t_k^j is the time at which the presynaptic unit fires its kth spike.

For each time step of 1 ms, the cable equation for the dendrite is integrated (MacGregor, 1987) with a finer time resolution of 0.1 ms and the somatic potential is compared with the spiking threshold V^{thr}. When this is exceeded, postsynaptic conductances are updated, and the somatic potential is reset to the after-hyperpolarization value V^{ahp} throughout the neuron.

Connections from excitatory to inhibitory, from inhibitory to excitatory, and between inhibitory units are taken to be homogeneous, that is, all of the same strength. Synaptic parameters depend only on the type of presynaptic and postsynaptic unit. The connectivity level is 0.25 between populations and 0.5 within the inhibitory population; that is, each unit synapses onto a fraction of the units of the receiving population, chosen at random. The excitatory units, in contrast, are all connected to each other. This very high connectivity, out of the actual anatomical range, is necessary, because of the small size of the simulated network, to produce sufficient statistical averaging in the synaptic input to each unit.

Excitatory-to-excitatory connections encode in their strength p memorized patterns of activity η_i^μ, consisting of binary words with sparseness (in this simple binary case, the fraction of 1s, or active cells in the pattern) $a = 0.1$. Encoding is implemented through a modified Hebb rule. In contrast with equation 2.7, which includes a baseline weight, all conductances are initially set to zero, and then, for each pattern, the synapse from the ith to the jth unit is modified by a covariance term,

$$\Delta g = \frac{g_{EE}}{C_{EE}} \left(\frac{\eta_i^\mu}{a} - 1 \right) \left(\frac{\eta_j^\mu}{a} - 1 \right). \tag{4.3}$$

If the conductance becomes negative, it is reset to zero. Memories are therefore stored through a "random walk with one reflecting barrier" procedure. The barrier acts as a "forgetting" mechanism (Parisi, 1986); whenever the conductance value bumps into the barrier, it loses memory about the previously presented patterns. Because there is no upper boundary, the average value of excitatory connection strengths grows with the number of memory items learned. The network is tested at low memory loading ($p = 10$). A systematic study of the storage capacity of the net would not be very meaningful because of the small size of the network.

The excitatory synapses impinge on the distal end compartment of the postsynaptic dendrite, and they have a positive reversal potential (referred to as resting membrane potential). Inhibitory synapses are distributed uniformly along the dendritic body, and they have a reversal potential equal to the resting membrane potential (except for the simulations in Figure 2). Inhibition is therefore predominantly shunting, with a geometry very similar to the one considered in Abbott (1991), leading to a mainly multiplicative effect on the postsynaptic firing rate. Table 1 summarizes the parameters used for the simulations.

Once the connection matrix is constructed, a test of the retrieval dynamics was performed according to the following protocol. The network is activated by injecting a current in a random fraction $a = 0.1$ of the units (see Figure 2A). The excitatory and the inhibitory population become diffusely active. Notice that units active in the memory pattern being tested are on average slightly more active than the other units. This is explained by the fact that they have

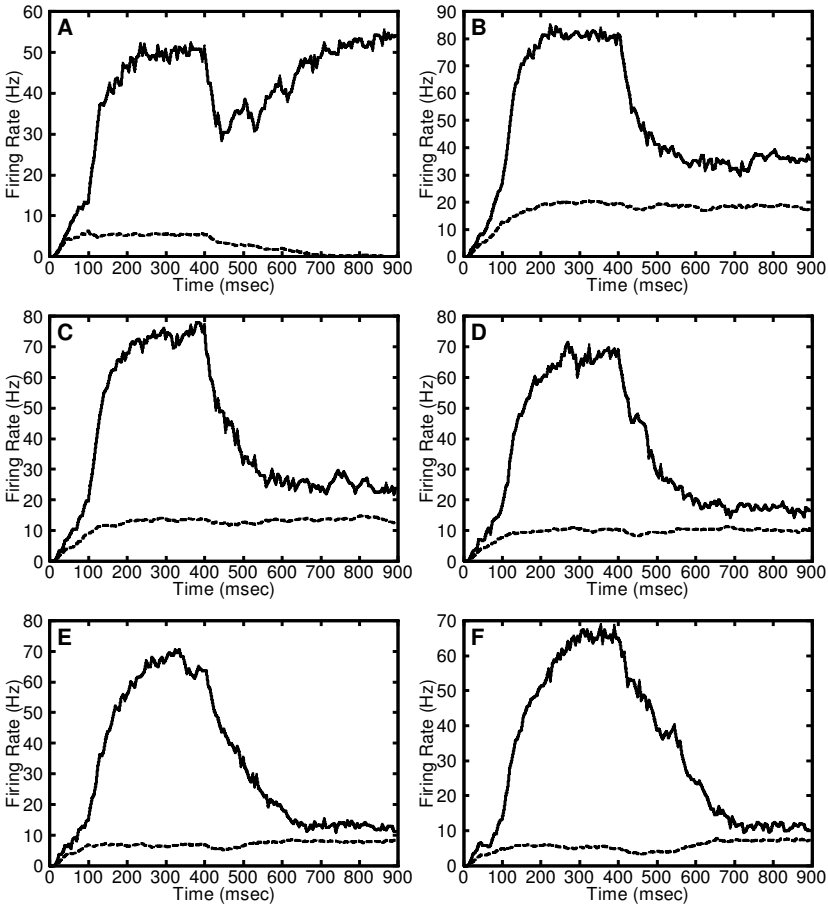

Figure 2: Firing rates computed with a time window of 30 msec are plotted for excitatory units for different geometries and reversal potential V_I. Units are divided between the 1 population (upper trace), active in the recalled memory, and the 0 population (lower trace), that was silent in the recalled memory. (A) $V_I = 0$ mV with respect to membrane equilibrium potential and inhibitory synapses are distributed along the dendritic body. In this condition, inhibition acts to some extent multiplicatively on the firing rate. Efficient retrieval of the memory is shown by sustained activity in the 1 population and complete activity suppression in the 0 population after the cue has been removed. (B–E): Inhibitory synapses are located on the edge of the dendritic cable. Reversal potential V_I is 0 mV (with respect to equilibrium) (B), -10 mV (C), -20 mV (D), -30 mV, (E) and -40 mV (F). Whatever the reversal potential, the two populations are never satisfactorily discriminated.

Table 1: Parameters Used for Integrate-and-Fire Simulations.

Quantity	Symbol	Value
Number of excitatory cells	N_E	800
Number of inhibitory cells	N_I	200
Corruption level	ρ	0.3
Random activity period	t_{init}	100 (msec)
Cue period	t_{cue}	300 (msec)
Retrieval period	t_{retr}	200 (msec)
Sampling time window	t_{win}	30 (msec)
Number of dendritic compartments	N_{cmp}	10
Dendritic compartment leakage conductance	G_0^d	6.28×10^{-12} (S)
Somatic compartment leakage conductance	G_0^s	5×10^{-9} (S)
Dendritic-dendritic axial conductance	G_0^{dd}	2.25×10^{-7} (S)
Excitatory somatic capacitance	$C_{soma,E}$	$0.5 - 4 \times 10^{-10}$ (F)
Inhibitory somatic capacitance	$C_{soma,I}$	5×10^{-12} (F)
Cue current	I_{cue}	0.25 (nA)
Firing threshold potential (excitatory)	Θ_E	32 (mV)
Firing threshold potential (inhibitory)	Θ_I	25 (mV)
After-spike hyperpolarization potential	V_{ahp}	−15 (mV)
Excitatory-excitatory connectivity level	C_{EE}	1
Excitatory-inhibitory connectivity level	C_{EI}	0.25
Inhibitory-excitatory connectivity level	C_{IE}	0.25
Inhibitory-inhibitory connectivity level	C_{II}	0.5
"Unitary" excitatory-excitatory synaptic conductance (see equation 4.3)	g_{EE}	$5 \times 10^{-8}(S)$*
Excitatory-inhibitory synaptic conductance	g_{EE}	$4 \times 10^{-9}(S)$
Inhibitory-excitatory synaptic conductance	g_{EE}	$2 \times 10^{-8}(S)$
Inhibitory-inhibitory synaptic conductance	g_{EE}	$9 \times 10^{-10}(S)$
Excitatory-inhibitory synaptic time constant	τ_{EI}	1 (msec)
Excitatory synaptic equilibrium (reversal) potential	V_E	65 (mV)
Inhibitory synaptic equilibrium (reversal) potential	V_I	0 (mV)
Excitatory-excitatory synaptic time constant	τ_{EE}	5–40 (msec)
Excitatory-inhibitory synaptic time constant	τ_{EI}	1 (msec)
Inhibitory-excitatory synaptic time constant	τ_{IE}	1 (msec)
Inhibitory-inhibitory synaptic time constant	τ_{II}	1 (msec)

Note: Ranges are indicated for quantities that varied within runs. Potential values are referred to membrane resting potential. * Scaled when the synaptic time constant is varied, to preserve the total charge transmitted during a synaptic event (see text). The value given is the one used for $\tau_{EE} = 20$ (msec). Simulations in Figure 2 are an exception as concerns inhibitory reversal potential (see the figure caption) and t_{retr}, which is set at 500 (msec).

on average a slightly stronger excitatory input, because the memory being tested contributes a positive term in the random walk construction of the connection strengths. Since p is not too large, even a single term makes a difference (Amit & Brunel, 1997).

After 100 msec, the random current is replaced with a *cue* current, injected in a fraction $a + \varrho(1 - a)$ of the units active in the pattern being tested and in a fraction $a(1 - \varrho)$ of the units inactive in the pattern. In this way, the cue is again a binary word with sparseness $a = 0.1$, and ϱ is the the average correlation between pattern and cue, which in the runs shown in the figures was set at $\varrho = 0.3$.

The cue current lasts for 300 msec. The average firing rate for the 1 units is much higher than for the 0 ones. When the cue current is removed, the 1 units sag briefly but then recover and stay steadily active, while activity in the others decays at zero firing or at a very low level. The memory pattern has therefore been successfully retrieved. To test the specific effect produced by the type of inhibition, we performed the stepwise manipulation shown in Figure 2. First (see panel B), all inhibitory connections to excitatory cells were moved to the end of the dendritic tree, colocalized with excitatory inputs. This made them somewhat less "multiplicative," and also weaker. The result is that inhibition becomes unable to suppress the firing of excitatory units, which should be quiescent, and the network fails to retrieve correctly (the residual difference between 1 and 0 units being due to the finite-p effect). To make inhibition stronger again while maintaining its subtractive character, the equilibrium potential of inhibitory synapses was lowered in panels C through F in steps of 10 mV. The result is that inhibition tends to suppress activity across excitatory units, without ever allowing the retrieval state to reemerge after removing the cue. This manipulation then indicates that altering the form of inhibition makes the network cross its capacity limit. Since even the first form, with the inputs spread along the dendritic tree, is far from being purely multiplicative, this capacity limit is well below the upper limit predicted by nondynamical calculations.

The simulations were repeated varying the neural and synaptic parameters—the excitatory synaptic time constant (changing at the same time the synaptic conductance to keep the strength of the connection invaried) and the somatic capacitance—in order to vary the firing rate. The inhibitory synaptic time constant was kept smaller than the excitatory time constant in order to speed up the stabilizing effect of recurrent inhibition.

To assess the quality of retrieval, we have taken the same information-theoretical measure used when recording from behaving animals (as opposed to immobilized ones, for example) (Rolls, Treves, & Tovee, 1997; Treves, Skaggs, & Barnes, 1996). The retrieval protocol is repeated for up to 30 trials for each stored memory. Ten randomly selected excitatory units are "recorded," that is, sampled for the number of spikes they fire in a time window of 30 msec. The window slides with a step of 5 msec spanning the entire simulated time course. The firing rate vector thus constructed at any

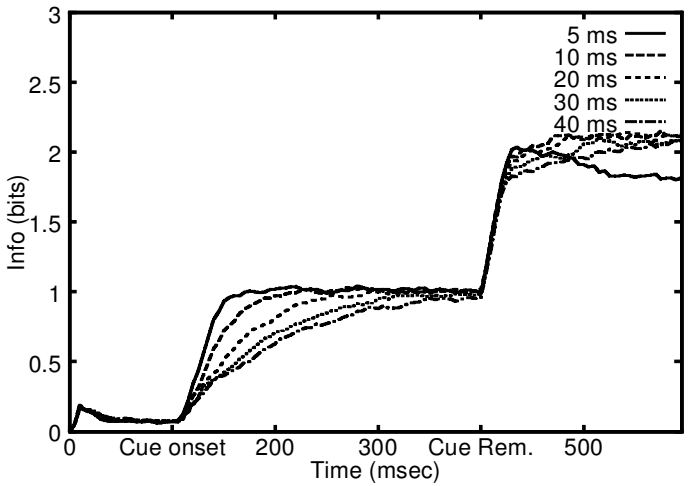

Figure 3: Information time course for different values of synaptic time constant. The transient corresponding to cue onset is well fitted by an exponential function. The rise is faster with shorter synaptic time constants.

time step of each trial is then decoded. This is done (Rolls et al., 1997) by matching it with the $p = 10$ mean firing rate vectors produced at the same time step when testing the retrieval of each of the memories and finding the closest match. The result of decoding all the trials is a probability table $P(s'|s)$ containing the likelihood that when testing for memory s, the activity of the sample of units was decoded as matching the average vector from pattern s'. The mutual information between the actual and decoded pattern,

$$I(s, s') = \sum_s \frac{1}{p} \sum_{s'} P(s'|s) \log_2 \frac{P(s'|s)}{P(s')}, \tag{4.4}$$

was calculated and then corrected for limited sampling (Treves & Panzeri, 1995; Panzeri & Treves, 1996). To reduce fluctuations, results were averaged at each time step over a number of samples of recorded units from the same run. The resulting quantity is a sensitive measure of how well the activity of the network in the time window can be used to discriminate which cue was presented and, unlike simpler measures (such as the correlation of the firing vector with the underlying memory pattern), can be used with identical procedures in simulations and recording experiments.

In Figure 3 we show the time course of the information for different values of the excitatory time constant. The mutual information stays close to zero during the random activity period (the small baseline is a remnant of the finite size error after the correction), and when the cue is presented, it

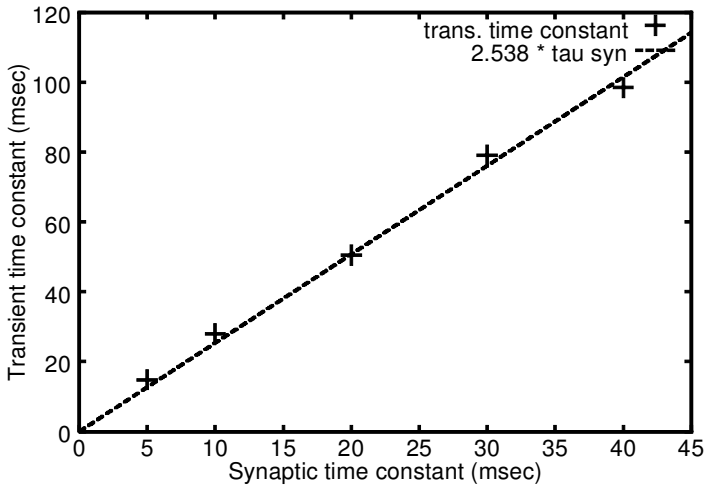

Figure 4: Transient time constant plotted against excitatory synaptic time constants. The firing rates were the same in each case, since conductance values were rescaled in order to equalize the charge entering into the cell through the synapse. The best linear fit line is shown. The slope of the fitted line is 2.538.

rises steadily to an equilibrium value, which depends on the correlation between the cue and the pattern, with a time course well fitted by a saturating exponential. This appears to be consistent with the linearized analysis for transients (Treves, 1993) and indicates that the transient modes that are activated in this condition belong to a single family; they share the same real part of the time constant. The time constant from the exponential fit is in a close-to-linear relationship with the synaptic (inactivation) time constant, as shown in Figure 4, with a best-fit proportionality coefficient of 2.538.

Varying the firing rate does not appear to have a comparable effect on the transient time constant. Figure 5 plots the transient time constant relative to different values of somatic capacitance, corresponding to firing rates ranging from ∼15 to ∼100 Hz.

When the cue is removed, the information rises again very rapidly to a higher equilibrium value, as the network is no longer constrained by the noisy cue, indicating that the network is acting as an "error corrector" during this later phase. The second transient is very rapid indeed, and it is in fact masked by an artifact induced by the finite size of the time window used to measure information (the artifact is that during the time window, what the measure reflects is actually a weighted sum of the lower value before cue removal and the higher value that is reached in a very short time). In fact, if one shrinks the sample window size, this linear raise shortens

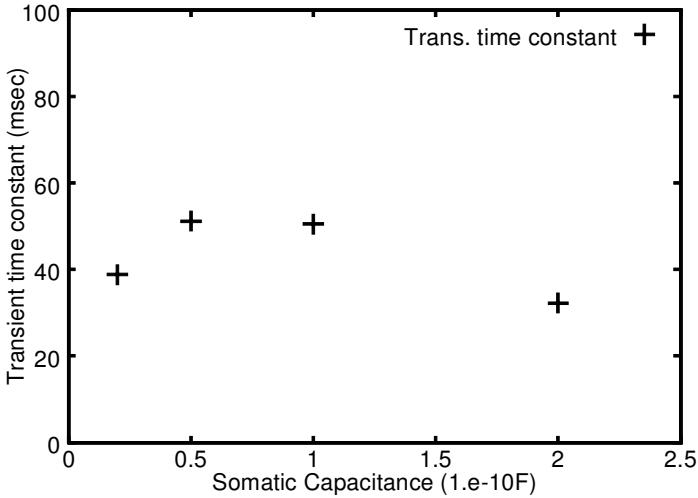

Figure 5: Transient time constant plotted for different values of somatic capacitance. Firing rates during the cue phase ranged correspondingly from 15 to 100 Hz. No clear dependence of the information time course is apparent when firing rates are varied in this way.

correspondingly (not shown). Although the actual time structure of this transient is still to be clarified, it seems clear that it follows a very different mode in this path to equilibrium. The final approach to the retrieval attractor is thus essentially immediate.

Finally, in Figure 6 we show the information behavior of the network when the excitatory collaterals are made informationless, or memoryless, by giving them all the same strength. A finite, small amount of information is seen in the cue phase only, at a much smaller level than for the structured network, and it falls to zero as the cue is removed. This demonstrates that selective activity, and in particular the capability of this network to retrieve memory patterns, depends crucially on the information encoded on its collaterals.

5 Implications for Recurrent Processing in the Brain

The more effective control that shunting inhibition may exert on runaway recurrent excitation, compared with subtractive inhibition, is an intuitive principle that has informed direct experimental studies (Miles & Wong, 1987). What has been shown here is how shunting inhibition may help avoid a specific conflict between stability and extensive memory storage

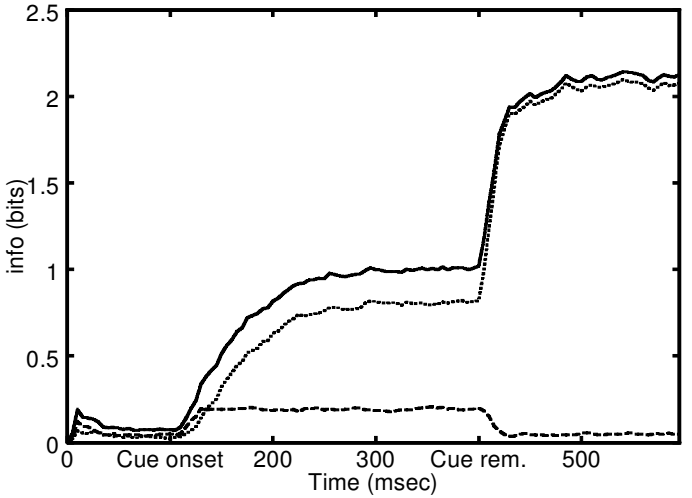

Figure 6: Information time course plotted for the structureless network compared with the time course for the network structured as in previous figures. During the cue phase, information reaches just a fraction of the steady-state value in the structured case. After the cue is removed, information decays to zero, reflecting the absence of self-sustained activity.

that would otherwise prevent the applicability of the abstract concept of a recurrent autoassociator to actual recurrent networks in the brain.

An attempt to demonstrate the large conductance changes that may underlie shunting inhibition (Douglas & Martin, 1991) has not confirmed the expectation; however, it is unclear to what extent the model used (the striate cortex of anesthetized cats) is relevant to the conditions we considered of massively reverberating excitation.

Having ensured the possibility of stable, asynchronous firing attractor states, simulations of a model network with spiking units and synaptic conductances have been used to confirm and extend earlier analytical results on the time required for memory retrieval mediated by recurrent processing to occur. The time course of the initial approach to the attractor state is, as in the analytical treatment, a saturating exponential, or a mixture of exponentially relaxing transient modes with similar (real part of the) time constant. This retrieval time constant is a linear function of the time constant for the inactivation of excitatory synaptic conductances and depends only mildly on prevailing firing rates or on neuronal time scales (as determined, for example, by membrane capacitance).

In practice, the contribution of recurrent processing, in this particular instance of an autoassociator, can be dominant within a few tens of milleseconds (with the parameters of Figure 3, within 2.5 times of the synaptic time

constant, which can be thought of as being in the 10 msec range; Colquhoun et al., 1992). This leads to the conclusion that at least local recurrent processing can be fast and that it is wrong to exclude its relevance in cases in which neuronal activity is found to acquire its selectivity within a few tens of milleseconds of its onset (Thorpe & Imbert, 1989; Treves, Rolls, & Tovee, 1996).

This result lends credibility to the hypothesis that recurrent autoassociation may be a ubiquitous function of local recurrent circuits throughout neocortex, as well as possibly the main function of recurrent connections in the hippocampal CA3 region (Treves & Rolls, 1991, 1994). At the same time, it raises the possibility of a direct manipulation of the time for such a function to be executed by acting on the inactivation kinetics of synaptic AMPA channels.

Acknowledgments

We are grateful to Mayank Mehta, Lorenzo Cangiano, and Martin Simmen, who participated in early phases of this project, and to Edmund Rolls, Stefano Panzeri, Carl van Vreeswijk, and Mike Hasselmo for extensive discussions. Partial support came from EC HCM contract CHRXCT930245 and CNR contribution 96.00287.CT02.

References

Abbott, L. F. (1991). Realistic synaptic input for model neural networks. *Network*, *2*, 245–258.

Abeles, M. (1991). *Corticonics—Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

Amit, D. J. (1995). The Hebbian paradigm reintegrated: Local reverberation as internal representation. *Behav. Brain Sci.*, *18*, 617–657.

Amit, D. J., & Brunel, N. (1997). Global spontaneous activity and local structured (learned) delay period activity in cortex. *Cerebral Cortex*, *7*, 237–252.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Ann. Phys. (N.Y.)*, *173*, 30–67.

Andersen, P., Eccles, J. C., & Loyning, Y. (1964). Location of post synaptic inhibitory synapses on hippocampal pyramids. *J. Neurophysiol.*, *27*, 592–607.

Battaglia, F. P., & Treves, A. (1996). Information dynamics in associative memories with spiking neurons. *Society for Neuroscience Abstracts*, *22*, 1124.

Colquhoun, D., Jonas, P., & Sakmann, B. (1992). Action of brief pulses of glutamate on AMPA/kainate receptors in patches from different neurones of rat hippocampal slices. *J. Physiol.*, *458*, 261–287.

Connors, B. W., Malenka, R. C., & Silva, L. R. (1988). Two inhibitory post-synaptic potentials, and $GABA_A$ and $GABA_B$ receptor-mediated responses in neocortex of rat and cat. *J. Physiol.*, *406*, 443-468.

Deppisch, J., Bauer, H.-U., Schillen, T., König, P., Pawelzik, K., & Geisel, T. (1993).

Alternating oscillators and stochastic states in a network of spiking neurons. *Network, 4,* 243–257.
Domany, E., Kinzel, W., & Meir, R. (1989). Layered neural networks. *J. Phys. A, 22,* 2081–2102.
Douglas, R. J., & Martin, K. A. C. (1991). A functional microcircuit for cat visual cortex. *J. Physiol (London), 440,* 735–769.
Eccles, J. C. (1957). *The physiology of nerve cells.* Baltimore: Johns Hopkins University Press.
Eccles, J. C. (1964). *The physiology of synapses.* New York: Academic Press.
Golomb, D., Rubin, N., & Sompolinsky, H. (1990). Willshaw model: Associative memory with sparse coding and low firing rates. *Phys. Rev. A., 41,* 1843–1854.
Hablitz, J. J., & Thalmann, R. H. (1987). Conductance changes underlying a late synaptic hyper-polarization in hippocampal CA3 neurons. *J. Neurophysiol., 58,* 160–179.
Hansel, D., Mato, G., & Meunier, C. (1995). Synchrony in excitatory neural networks. *Neur. Comp., 7,* 307–337.
Hebb, D. O. (1949). *The organization of behaviour.* New York: Wiley.
Hestrin, S., Nicoll, R. A., Perkel, D. J., & Sah, P. (1990). Analysis of excitatory synaptic action in pyramidal cells using whole-cells recording from rat hippocampal slices. *J. Physiol., 422,* 203–225.
Hopfield, J. J. (1982). Neural networks and physical systems with emerging collective computational abilities. *Proc. Natl. Acad. Sci. USA, 79,* 2554–2558.
Kohonen, T. (1977). *Associative memory.* Berlin: Springer-Verlag.
Lapique, L. (1907). Recherches qualitatives sur l'excitation électrique de nerfs traités comme une polarisation. *J. Physiol. Pathol. Gen., 9,* 620–635.
Little, W. A. (1974). The existence of persistent states in the brain. *Math. Biosci., 19,* 101–120.
MacGregor, R. J. (1987). *Neural and brain modeling.* San Diego: Academic Press.
Marr, D. (1971). Simple memory: A theory for the archicortex. *Phil. Trans. Roy. Soc. (London) B, 262,* 24–81.
McBain, C., & Dingledine, R. (1992). Dual-component miniature excitatory synaptic currents in rat hippocampal CA3 pyramidal neurons. *J. Neurophysiol., 68,* 16–27.
Miles, R., & Wong, R. K. S. (1987). Inhibitory control of local excitatory circuits in the guinea-pig hippocampus. *J. Physiol., 338,* 611–629.
Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network, 7,* 87–107.
Parisi, G. (1986). A memory which forgets. *J. Phys. A, 19,* L617.
Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by population of neurons in the primate temporal visual cortex. *Exp. Brain. Res., 114,* 149–162.
Simmen, M. W., Treves, A., & Rolls, E. T. (1996). On the dynamics of a network of spiking neurons. In F. H. Eekman & J. M. Bower (Eds.), *Computations and neuronal systems: Proceedings of CNS95.* Boston: Kluwer.
Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist models. In R. Pfeifer, Z. Schreter, & F. Fogelman-Soulié (Eds.), *Connectionism in perspective* (pp. 63–92). Amsterdam: Elsevier.

Treves, A. (1990). Graded-response neurons and information encodings in auto-associative memories. *Phys. Rev. A, 42,* 2418–2430.
Treves, A. (1993). Mean-field analysis of neuronal spike dynamics. *Network, 4,* 259–284.
Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neur. Comp., 7,* 399–407.
Treves, A., & Rolls, E. T. (1991). What determines the capacity of auto-associative memories in the brain? *Network, 2,* 371–397.
Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus, 4,* 374–391.
Treves, A., Rolls, E. T., & Tovee, M. J. (1996). In V. Torre & F. Conti (Eds.), *Neurobiology: Proceedings of the International School of Biophysics, XXIII Course, May 1995* (pp. 371–382). New York: Plenum Press.
Treves, A., Skaggs, W. E., & Barnes, C. A. (1996). How much of the hippocampus can be explained by functional constraints? *Hippocampus, 6,* 666–674.
Tsodyks, M. V., & Feigelman, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett., 46,* 101.
Tsodyks, M. V., Mitcov, I., & Sompolinsky, H. (1993). Patterns of synchrony in inhomogeneous networks of oscillators with pulse interactions. *Phys. Rev. Lett., 71,* 1280–1283.
van Vreeswijk, C. A., Abbott, L. F. & Ermentrout, G. B. (1994). When inhibition not excitation synchronizes neural firing. *J. Comput. Neurosci., 1,* 313–321.
Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature, 222,* 960–962.
Wilson, H. R., & Cowan J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J., 12,* 1.

Received January 3, 1997; accepted June 25, 1997.