Computing and stability in cortical networks

Peter E. Latham and Sheila Nirenberg

Dept. of Neurobiology, UCLA, Los Angeles, CA 90095-1763

pel@ucla.edu, sheilan@ucla.edu

http://culture.neurobio.ucla.edu/~pel

July 16, 2003

Abstract

Cortical neurons are predominantly excitatory and highly interconnected. In spite of this, the cortex is remarkably stable: normal brains do not exhibit the kind of runaway excitation one might expect of such a system. How does the cortex maintain stability in the face of this massive excitatory feedback? More importantly, how does it do so during computations, which necessarily involve elevated firing rates? Here we address these questions in the context of attractor networks – networks that exhibit multiple stable states, or memories. We find that such networks can be stabilized at the relatively low firing rates observed *in vivo* if two conditions are met: 1) the background state, where all neurons are firing at low rates, is inhibition-dominated, and 2) the fraction of neurons involved in a memory is above some threshold, so that there is sufficient coupling between the memory neurons and the background. This allows "dynamical stabilization" of the attractors, meaning feedback from the pool of background neurons stabilizes what would otherwise be an unstable state. We suggest that dynamical stabilization may be a strategy used for a broad range of computations, not just those involving attractors.

1 Introduction

Attractors networks – networks that exhibit multiple stable states – have served as a key theoretical model for several important computations, including associative memory (Hopfield, 1982; Hopfield, 1984), working memory (Amit & Brunel, 1997a; Brunel & Wang, 2001), and the vestibular-ocular reflex (Seung, 1996), and determining whether these models apply to real biological networks is an active area of experimental research (Miyashita & Hayashi, 2000; Aksay *et al.*, 2001; Ojemann *et al.*, 2002; Naya *et al.*, 2003). A definitive determination, however, has been difficult, mainly because attractors cannot be observed directly; instead, inferences must be made about their existence by comparing experimental data with model prediction.

To make these comparisons, it is necessary to have realistic models. Construction of such models, however, has proved difficult, because of what we refer to as the "stability problem". The stability problem arises primarily because cortical networks are highly recurrent – a typical neuron in the cortex receives input from 5,000-10,000 others, most of which are nearby (Braitenberg & Schüz, 1991). While this high connectivity undoubtedly provides immense computational power, it also has a downside: it can lead to instabilities in the form of runaway excitation. For example, even mild electrical stimulation applied periodically can eventually leads to seizures (McIntyre *et al.*, 2002), and epilepsy – a sign of intrinsic instability – occurs in 0.5-1% of the human population (Bell & Sander, 2001; Hauser, 1997).

The severity of the stability problem lies in the fact that recurrent connections in attractor networks have to be strong enough to allow activity in the absence of input, but not so strong that the activity can occur spontaneously. Moreover, in areas where attractor networks are thought to exist, such as prefrontal cortex (Fuster & Alexander, 1971; Wilson *et al.*, 1993; Freedman *et al.*, 2001) and inferior temporal cortex (Fuster & Jervey, 1982; Miyashita & Chang, 1988), the firing rates associated with attractor states are not much higher than those associated with background activity – the two differ by only a few Hz, with typical background rates ranging from 1-10 Hz and attractor rates ranging from 10-20 Hz (Fuster & Alexander, 1971; Miyashita & Chang, 1988; Nakamura & Kubota, 1995). This small difference makes a hard stability problem even harder, as there is almost no barrier preventing spontaneous jumps to attractors. Indeed, in previous attempts to build realistic models of attractor networks (Amit & Brunel, 1997a; Wang, 1999; Brunel, 2000; Brunel & Wang, 2001), fine-tuning of parameters was required to support attractors at the low rates seen *in vivo*. This was mainly because firing rates in those models were effectively set by single neuron saturation – by the fact that neurons simply cannot fire for extended periods above a maximum rate.

Here we propose a solution to the fine-tuning problem, one that allows attractors to exist at low rates over a broad range of parameters. The basic idea is to use natural interactions between the attractors and the background to limit firing rates, so that rates are set by network rather than single neuron properties. Limiting firing rates in this way may be a general computational strategy, not one used just by attractor networks. Thus, quantifying this mechanism in the context of attractor networks may serve as a general model for how cortical circuits carry out a broad range of computations while avoiding instabilities.

2 Reduced Model System

To understand qualitatively the properties of attractor networks, we analyze a model system in which the neurons are described by their firing rates. For simplicity, we start with a network that exhibits two stable equilibria: a background state, where all neurons fire at low rates, and a memory state, where a sub-population of neurons fires at an elevated rate. Our results, however, apply to multi-attractor networks – networks that exhibit multiple memories – and we will consider such networks below. The goal of the analysis in this section is to obtain a qualitative understanding of attractor networks; in particular, how they can fire at low rates without destabilizing the background. We will then use this qualitative understanding to guide network simulations of multi-attractor networks (containing up to 50 attractors), and use those simulations to verify the results obtained from the reduced model.

We construct our reduced model attractor network in two steps. In the first we build a randomly connected network of excitatory and inhibitory neurons that fire at a few Hz, on average. In the second we pick out a sub-population of the excitatory neurons (which we refer to as memory neurons) and strengthen the connections among them (Fig. 1). If the parameters are chosen properly, this strengthening will produce a network in which the

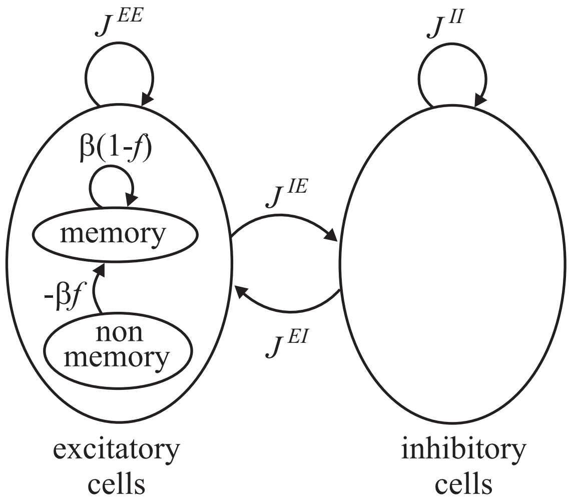

Figure 1: Schematic of network architecture. Excitatory and inhibitory neurons are synaptically coupled via the connectivity matrix J. The inhibitory neurons are homogeneous; the excitatory neurons break into two populations, memory and non-memory. The coupling among memory neurons is higher than among non-memory neurons by a factor of $\beta(1-f)$; the coupling from non-memory to memory neurons is lower by a factor of βf. The memory neurons make up a fraction f of the excitatory population, so the decrease in coupling from non-memory to memory neurons ensures that the total synaptic drive to both memory and non-memory neurons is the same.

memory neurons can fire either at an elevated rate or at the background rate, resulting in a bistable network.

To analyze the stability and computational properties of this network, we assume that, in equilibrium, the firing rate of a neuron is a function only of the firing rates of the neurons pre-synaptic to it. Then, near equilibrium, we expect Wilson and Cowan-like equations (Wilson & Cowan, 1972) in which the firing rates obey first order dynamics. Augmenting the standard Wilson and Cowan equations to allow increased connectivity among a subpopulation of neurons (see Fig. 1 and Appendix A), we have

$$\tau_E \frac{d\nu_{Ei}}{dt} + \nu_{Ei} = \phi_E \left(J^{EE}\nu_E - J^{EI}\nu_I + \frac{\beta}{N_E f(1-f)} \sum_j \xi_i(\xi_j - f)\nu_{Ej} \right) \tag{1a}$$

$$\tau_I \frac{d\nu_I}{dt} + \nu_I = \phi_I \left(J^{IE}\nu_E - J^{II}\nu_I \right) . \tag{1b}$$

Here τ_E and τ_I are the excitatory and inhibitory time constants, ν_{Ei} is the firing rate of the i^{th} excitatory neuron ($i = 1, ..., N_E$), ν_E and ν_I are the average firing rates of the excitatory and inhibitory neurons, respectively, the Js are the average coupling coefficients among the excitatory and inhibitory populations, ϕ_E and ϕ_I are the average excitatory and inhibitory gain functions, respectively, β is the effective strength of the coupling among the memory neurons, N_E is the number of excitatory neurons, and ξ is a random binary vector: $\xi_i = 1$ with probability f and 0 with probability $1 - f$. The factor $\xi_j - f$ ensures post-synaptic normalization: on average, the total synaptic strength to both memory and non-memory neurons is the same. The ξ-dependent term in Eq. (1a) is a standard one for constructing attractor networks (Hopfield, 1982; Tsodyks & Feigel'man, 1988; Buhmann, 1989).

The gain functions, ϕ_E and ϕ_I, play a key role in this analysis. These functions, which hide all the single neuron properties, have a natural interpretation: each one corresponds to the average firing rate of a population of neurons as a function of the average firing rates of neurons pre-synaptic to it. They thus have standard shapes when plotted versus ν_E: they look like f-I curves – firing rate versus injected current (McCormick *et al.*, 1985) – that have been smoothed at low firing rates (Brunel & Sergi, 1998; Tiesinga *et al.*, 2000; Fourcaud & Brunel, 2002; Brunel & Latham, 2003). A generic gain function is shown in Fig. 2. This curve does not correspond to any particular neuron or class of neurons – it is just illustrative. There are two important aspects to its shape: 1) it has a convex region, 2) the transition from convex to concave occurs at firing rates well above 10-20 Hz on the output side (that is, the transition occurs when $\phi \gg$ 10-20 Hz). These properties are typical of both model neurons (Brunel & Sergi, 1998; Tiesinga *et al.*, 2000; Fourcaud & Brunel, 2002; Brunel & Latham, 2003) and real neurons (McCormick *et al.*, 1985; Chance *et al.*, 2002), and in the next two sections we will see how they connect to the problem of robust, low firing rate attractors.

Although Eqs. (1a) and (1b) have a fairly broad range of applicability, they are not all-encompassing. In particular, they leave out the possibility of instability via synchronous oscillations (Abbott & van Vreeswijk, 1993; Gerstner & van Hemmen, 1993; Hansel & Mato, 2001), and they ignore the effects of higher order moments of the firing rate (Amit & Brunel, 1997a; Latham, 2002). We thus verify all predictions using large-scale simulations of synaptically coupled neurons.

As indicated in Fig. 1, the network described by Eq. (1) contains a preferentially connected sub-population. We are interested in determining under what conditions the network supports two states: a "memory" state in which this sub-population fires at elevated rate compared to the background, and a background state in which the sub-population fires at

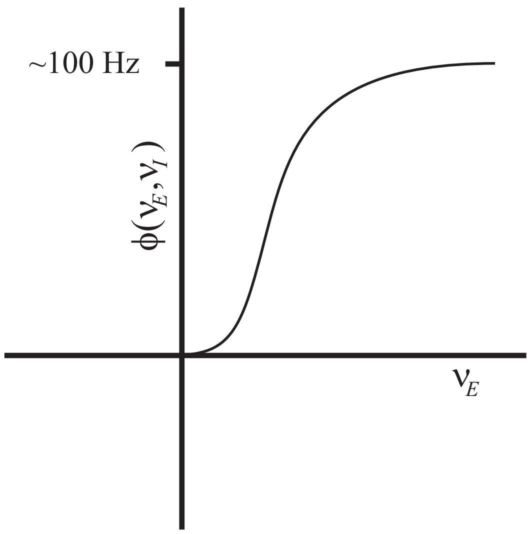

Figure 2: Generic gain function versus average excitatory firing rate, ν_E. There are three distinct regimes: First, at small ν_E, the mean current induced by pre-synaptic firing is too low to cause a post-synaptic neuron to fire, so post-synaptic activity is due to fluctuations in the current. This is the noise-dominated regime, and here ϕ is convex. Second, for slightly larger ν_E, ϕ becomes approximately linear or slightly concave. Finally, for large enough ν_E, the firing rate saturates – at around 100 Hz for typical cortical pyramidal cells (McCormick *et al.*, 1985). This gain function is a schematic, and does not correspond to any particular neuron model.

the same rate as the background. Since it is the difference between the attractor and background firing rates that is important, we define a variable, m, that is proportional to this difference; for convenience, we let the proportionality constant be $1/(1-f)$,

$$m \equiv \frac{1}{1-f}\left[\frac{1}{N_E f}\sum_i \xi_i \nu_{Ei} - \frac{1}{N_E}\sum_i \nu_{Ei}\right] . \tag{2}$$

The first expression inside the angle brackets is the average firing rate of the sub-population; the second is the average firing rate of the whole population, ν_E. Thus, the average firing rate of the sub-population is $\nu_E + (1-f)m$, and the background state corresponds to $m = 0$.

Intuitively, we expect that network dynamics should be governed by three variables: the firing rate associated with the memory state, m, the average excitatory firing rate, ν_E, and the average inhibitory rate, ν_I. In fact, if we average Eq. (1) over index, i, we can derive differential equations for these variables. As shown in Appendix A, the averaged equations are

$$\tau_E \frac{dm}{dt} + m = \phi_E(J^{EE}\nu_E - J^{EI}\nu_I + \beta m) - \phi_E(J^{EE}\nu_E - J^{EI}\nu_I) \tag{3a}$$

$$\tau_E \frac{d\nu_E}{dt} + \nu_E = (1-f)\phi_E(J^{EE}\nu_E - J^{EI}\nu_I) + f\phi_E(J^{EE}\nu_E - J^{EI}\nu_I + \beta m) \tag{3b}$$

$$\tau_I \frac{d\nu_I}{dt} + \nu_I = \phi_I(J^{IE}\nu_E - J^{II}\nu_I)\,. \tag{3c}$$

To simplify our analysis we adopt the the effective response function approach of Mascaro and Amit (Mascaro & Amit, 1999), which can be implemented by taking the limit of fast inhibition, $\tau_I \to 0$. The main caveat to this limit is that we may overestimate the stability of equilibria: equilibria can be stable when $\tau_I = 0$ but unstable when τ_I is above some threshold (Wilson & Cowan, 1972; van Vreeswijk & Sompolinsky, 1996; Latham *et al.*, 2000a; Hansel & Mato, 2001). With $\tau_I = 0$, we can solve Eq. (3c) for ν_I in terms of ν_E. This allows us to write $\nu_I = \nu_I(\nu_E)$, where the function $\nu_I(\nu_E)$ is determined implicitly from Eq. (3c) with $\tau_I = 0$. Replacing ν_I with $\nu_I(\nu_E)$ in Eqs. (3a) and (3b), we find that all the ν_E-dependence on the right hand side of these equations can be lumped into the single expression $J^{EE}\nu_E - J^{EI}\nu_I(\nu_E)$. We denote this $-\gamma(\nu_E)$, so that

$$\gamma(\nu_E) \equiv -J^{EE}\nu_E + J^{EI}\nu_I(\nu_E)\,. \tag{4}$$

With ν_I eliminated, we are left with just two equations,

$$\begin{aligned} \tau_E \frac{dm}{dt} + m &= \phi_E(-\gamma(\nu_E) + \beta m) - \phi_E(-\gamma(\nu_E)) \equiv \Delta\phi_E(\nu_E, m) && (5a)\\ \tau_E \frac{d\nu_E}{dt} + \nu_E &= \phi_E(-\gamma(\nu_E)) + f\Delta\phi_E(\nu_E, m)\,. && (5b) \end{aligned}$$

These equations are similar to ones derived previously by Brunel and colleagues (Amit & Brunel, 1997a; Brunel, 2000; Brunel & Wang, 2001).

2.1 The sparse Coding Limit. The question we are interested in is: under what conditions do Eqs. (5a) and (5b) admit two stable solutions, one with $m = 0$ (the background state) and one with $m \neq 0$ (the memory state)? Before we answer this question in general, let us consider the sparse coding limit, $f \to 0$, as this limit is relatively simple and it allows us to make contact with previous work.

When $f = 0$, Eqs. (5a) and (5b) decouple. In this regime we can solve Eq. (5b) for the excitatory firing rate, then solve Eq. (5a) for m. Let us assume that Eq. (5b) has a stable solution – meaning the network admits a stable background state – and focus on Eq. (5a). This equation is best solved graphically, by plotting $\Delta\phi_E(\nu_E, m)$ versus m and looking for intersections of this plot with the 45° line; those intersections correspond to equilibria (see Fig. 3). Stability can be read off the plots by looking at the slope: an equilibrium with slope less than 1 is stable and one with slope greater than 1 is unstable.

The number and stability of the equilibria are determined by β. If β is small – weak coupling among the neurons in the sub-population – there is only one equilibrium, at $m = 0$, and it is stable (Fig. 3a, dotted line). This makes sense, as weak coupling shouldn't have much effect on network behavior. As β increases so does the slope of $\Delta\phi_E(\nu_E, m)$, and eventually a new pair of equilibria appear (Fig. 3a, solid line). Of these, the one at higher firing rate (larger m) is stable, since its slope is less than 1, and the one at lower, but nonzero, rate is unstable, since its slope is greater than 1. Finally, for large enough β, the unstable equilibrium slides past zero, at which point the equilibrium at $m = 0$, and thus the background, becomes unstable (Fig. 3a, dashed line).

The intermediate β regime, where the network can support both a memory and a stable background, is of the most interest to us. It is in this regime that the network can actually compute, in the sense that input to the network controls which state it's in (memory or background). The low and high β regimes are not so interesting, however – at least if the goal is to construct a network that can store memories. If β is too low, the network is stuck in the background state; if it is too high, the network can easily jump spontaneously from the background to the memory state.

As Fig. 3a shows, it's not hard to build a network that can exhibit two states – so long as one is willing to allow firing rates near saturation, meaning at a healthy fraction of 100 Hz. It is much more difficult to build a network in which the memory state exhibits low firing rates – in the 10-20 Hz range, as is observed experimentally (Fuster & Alexander, 1971; Miyashita & Chang, 1988; Nakamura & Kubota, 1995). This is because low firing rates require too many bends in $\Delta\phi_E(\nu_E, m)$ in too small a firing rate range, where "too small" is relative to the saturation firing rate of $\sim$100 Hz (Figs. 3b and 3c).

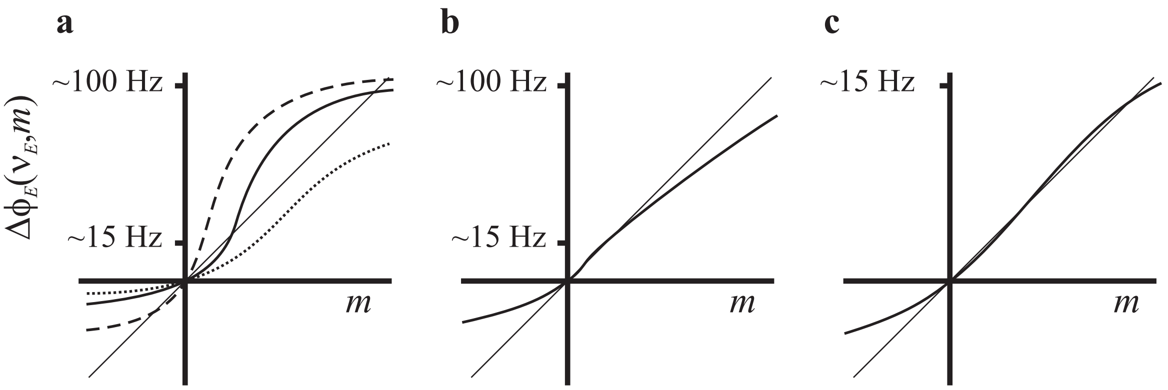

Figure 3: Gain functions and m-equilibria. **a**. $\Delta\phi_E(\nu_E, m)$ versus m for different values of β. Points where the curves cross the 45° line are equilibria. In the sparse coding limit, $f \to 0$, an equilibrium is stable if the slope of the curve is less than one and unstable otherwise. Dotted line: weak coupling (small β). The only equilibrium is at the background, and it is stable. Black line: intermediate coupling. Two more equilibria have appeared. The one with intermediate m is unstable, the one with large m is stable. Dashed line: large coupling. The background has become destabilized and the memory is easily activated without input. Since $\Delta\phi_E(\nu_E, m)$ is essentially an average single neuron f-I curve, it saturates at the maximum firing rate of single neurons, $\sim$100 Hz. Thus, the upper equilibrium occurs at high rate. **b**. To reduce the firing rate of the upper equilibrium, one must operate in a very narrow parameter regime: small changes in network parameters would either rotate the curve counterclockwise, which would destabilize the background and/or move the equilibrium to high rate, or rotate it clockwise, which would eliminate the memory state. **c**. Blowup of the region below 15 Hz in panel b.

These qualitative arguments were quantified in networks of leaky integrate-and-fire neurons using both analytic techniques and numerical simulations (Brunel, 2000; Brunel & Wang, 2001). For attractors that fired at about 15 Hz, the coupling, β, among the subpopulations had to be tuned to within 1% to ensure that both the background and memories were stable. At slightly higher firing rates, 25-40 Hz, the tuning was somewhat more forgiving, 3-5%. It should be pointed out, however, that these networks were more robust to other parameters: multiple attractors were supported at reasonably low rates (20-40 Hz) when external input varied over a 40% range and the strengths of different receptor types (AMPA, NMDA, and GABA) were varied over a 5-15% range.

2.2 Beyond the Sparse Coding Limit. What these results indicate is that parameters need to be finely tuned for attractor networks to exist at low rates, at least in the $f \to 0$ limit.

When f is finite, however, the picture changes, since the memories (m) and background (ν_E) couple (see Eq. (5)). This means that the slope of $\Delta\phi_E(\nu_E, m)$ with respect to m is no longer the sole factor determining stability; instead, stability depends on the interaction between m and ν_E. Consequently, an equilibrium in which the slope of $\Delta\phi_E(\nu_E, m)$ is greater than 1 can be stable. For example, the equilibrium on the gray curve in Fig. 3a that occurs at about 15 Hz, which is unstable in the $f \to 0$ limit, could become stable when $f > 0$. If this were to happen, attractors could exist robustly at low rate.

We refer to the regime in which the attractors are stable even though the slope of $\Delta\phi_E(\nu_E, m)$ is greater than 1 as the *dynamically stabilized* regime. This is because an equilibrium that would be unstable when one considers only the time evolution of the memory is dynamically stabilized by feedback from the background. Networks that operate in this regime have been investigated by Sompolinsky and colleagues (Rubin & Sompolinsky, 1989; Golomb *et al.*, 1990). Although the firing rates of the attractors in those networks were low, the networks were not realistic in an important respect: they did not exhibit a background state in which all neurons fired at about the same rate; instead, the only stable states were ones in which a sub-population of the neurons were active. Our goal here is to overcome this problem, and build a network that both operates in the dynamically stabilized regime, and thus at low rates, *and* supports a low firing rate background. To do this, we must evaluate the stability of the equilibria in m-ν_E space.

The network equilibria are found by setting dm/dt and $d\nu_E/dt$ in Eqs. (5a) and (5b) and solving the resulting algebraic equations. With the time derivatives set to zero, the solutions to Eqs. (5a) and (5b) are curves in m-ν_E space. These curves are referred to as nullclines, and their intersections correspond to equilibria. Constructing nullclines from the properties of ϕ_E is straightforward (Latham *et al.*, 2000a) but tedious. We thus skip the details and simply plot them; once we've made the plots, it's relatively easy to see that they have the correct qualitative shape.

One possible set of nullclines is shown in Fig. 4a, with the black curve corresponding to the ν_E-nullcline (the solution to Eq. (5b) with $d\nu_E/dt = 0$) and the gray curves corresponding to the m-nullcline (the solution to Eq. (5a) with $dm/dt = 0$; note that there are two pieces, a smooth curve and a vertical line). The shape of the ν_E-nullcline is relatively easy to

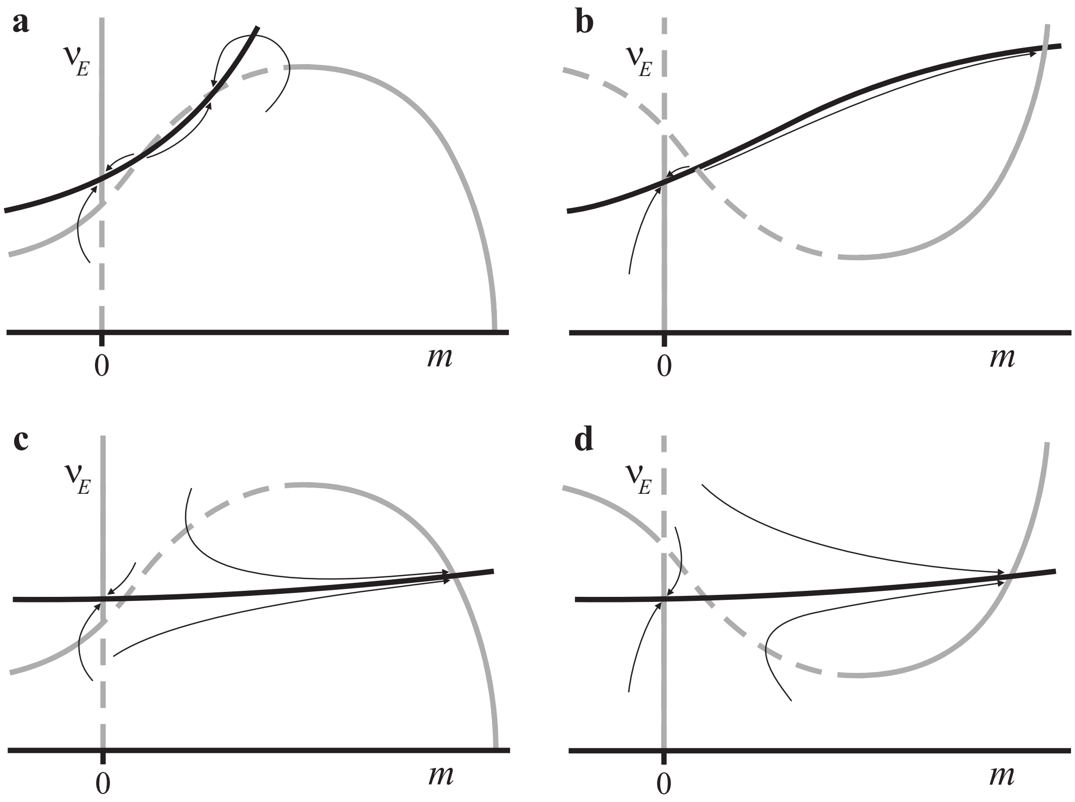

Figure 4: Nullclines in various parameter regimes. The gray curve, including the vertical line, is the nullcline for Eq. (5a); the black curve is the nullcline for Eq. (5b). The solid branches of the gray curves are stable at fixed ν_E, the dashed branches are unstable at fixed ν_E. The black arrows indicate typical trajectories; they are derived from Eq. (5). **a**. $\Delta\phi_E(\nu_E, m)$ is a decreasing function of ν_E; f is reasonably large. **b**. $\Delta\phi_E(\nu_E, m)$ is an increasing function of ν_E; f is reasonably large. **c**. $\Delta\phi_E(\nu_E, m)$ is a decreasing function of ν_E; $f \ll 1$. **d**. $\Delta\phi_E(\nu_E, m)$ is an increasing function of ν_E; $f \ll 1$.

understand: it is an increasing function of m, reflecting the fact that as m gets larger there is more excitatory drive to the network. This can also be seen from Eq. (5b), where ν_E is coupled to m through the term $f\Delta\phi_E(\nu_E, m)$, and $\Delta\phi_E(\nu_E, m)$ is an increasing function of m.

The m-nullcline is a little more complicated, primarily because it consists of two pieces rather than one. To understand its shape we must re-examine Fig. 3a. This figure shows three plots of $\Delta\phi_E(\nu_E, m)$ versus m. These plots correspond to three values of β; however, because ν_E as well as β affects $\Delta\phi_E(\nu_E, m)$, they could just as easily have corresponded to three values of ν_E. In fact, had we fixed β and varied ν_E, we would have obtained a set of curves qualitatively similar to the ones shown in Fig. 3a. In other words, depending on the value of ν_E, we would have seen three distinct regimes: 1) one equilibrium at $m = 0$ (dotted line in Fig. 3a); 2) one equilibrium at $m = 0$ and two at $m > 0$ (solid line); and 3) one equilibrium at $m = 0$, one at $m < 0$ and one at $m > 0$ (dashed line). These three regimes are reflected in the m-nullcline in Fig. 4a: when ν_E is large there is a single equilibrium at $m = 0$ (regime 1); when ν_E is intermediate there is an additional pair of equilibria at $m > 0$ (regime 2); and when ν_E is small one of the equilibria in that pair becomes negative (regime 3).

The order in which the regimes appear in Fig. 4a – one equilibrium at zero when ν_E is large, two positive ones when ν_E is slightly smaller, and a negative one when ν_E is smaller still – corresponds to a particular dependence of $\Delta\phi_E(\nu_E, m)$ on ν_E. Examining Fig. 3a, we see that this progression corresponds to a dependence in which $\Delta\phi_E(\nu_E, m)$ decreases as ν_E increases. Thus, for Fig. 4a to correctly describe the m-nullcline, the network parameters must be such that $\Delta\phi_E(\nu_E, m)$ is a *decreasing* function of ν_E. Below we derive explicit conditions under which this happens. First, however, we examine its consequences.

The black and gray nullclines in Fig. 4 intersect in three places, and thus exhibit three equilibria. One is at m=0 and two are at $m > 0$. The equilibrium at $m = 0$ corresponds to the background state; the other two are candidates for memory states. To determine which are stable, we have plotted a few typical trajectories (black arrows), which are derived from Eq. (5). These trajectories tell us that the stable equilibria occur at $m = 0$ and at the rightmost intersection. Importantly, the stable equilibrium at $m > 0$ occurs on the *unstable* branch of the m-nullcline, in the dynamically stabilized regime. (We can tell that this branch is unstable because at fixed ν_E the trajectories point away from it; flow is to the right below the m-nullcline and to the left above it. The ν_E-nullcline, on the other hand, consists of one stable branch, since flow is toward it at fixed m. This reflects the fact that the background is always stable at fixed m, independent of the firing rate of the memory state.)

The unstable branches, which are drawn with dashed lines, correspond to points where the slope of $\Delta\phi_E(\nu_E, m)$ is greater than one. That the memory state lives on the unstable branch of the m-nullcline is important, because the unstable branch corresponds to the intermediate equilibrium shown in Fig. 3a and can thus occur at low firing rate. All previous work in realistic networks that we know of (Amit & Brunel, 1997a; Brunel, 2000; Brunel &

Wang, 2001) put the memory state on the *stable* branch of the m-nullcline, the part with slope less than one. The stable branch corresponds to the upper equilibrium in Fig. 3a, and so tends to occur at high firing rate – at some reasonable fraction of the saturation firing rate for single neurons.

It may seem counterintuitive to have a stable equilibrium on the unstable branch, but it's well known that this can happen (Rinzel & Ermentrout, 1989). The only caveat is that there is no guarantee of stability: the equilibrium may become unstable via a Hopf bifurcation (Marsden & McCracken, 1976), leading to oscillations. In fact, oscillations were occasionally observed – mainly near or above the stability boundary in Fig. 7. The fact that the network did not oscillate in most of the parameter regime explored was, we believe, because the background is strongly stable (there is strong attraction to the ν_E-nullcline in Fig. 4).

What this analysis shows is that two conditions are necessary for building an attractor network in which the memory states occur on the unstable branch of the m-nullcline, and thus fire at low rates: $\Delta\phi_E(\nu_E, m)$ must be a decreasing function of ν_E, and the fraction, f, of neurons involved in a memory must be large enough to give the ν_E-nullcline significant curvature. What happens if either of these conditions is violated? Figure 4b shows the nullclines when $\Delta\phi_E(\nu_E, m)$ is an increasing function of ν_E; in this regime the m-nullcline turns upside down. A low-firing rate equilibrium still exists, but it is unstable, as indicated by the black arrows. The stable memory state in this regime is at high firing rate, near single neuron saturation – too high to be consistent with experiment. Figures 4c and 4d show nullclines when $f \ll 1$. Again, the only stable memory states are near saturation, and thus at high rate.

Is the condition that $\Delta\phi_E(\nu_E, m)$ be a decreasing function of ν_E satisfied for realistic networks? In other words, is it reasonable to expect $\partial\Delta\phi_E(\nu_E, m)/\partial\nu_E < 0$? To answer this, we use Eq. (5a) to write

$$\frac{\partial\Delta\phi_E(\nu_E, m)}{\partial\nu_E} = -\gamma'(\nu_E)[\phi'_E(-\gamma(\nu_E) + \beta m) - \phi'_E(-\gamma(\nu_E))] \tag{6}$$

where a prime after a function denotes a derivative with respect to its argument. The term in brackets on the right hand side of Eq. (6) is typically positive for a memory lying on the unstable branch of the m-nullcline; this is because the unstable branch corresponds to the intermediate equilibrium in Fig. 3a, where ϕ_E is generally convex. Thus, the sign of $\partial\Delta\phi_E(\nu_E, m)/\partial\nu_E$ is determined solely by the sign of $\gamma'(\nu_E)$. For typical cortical networks, it is likely that $\gamma'(\nu_E) > 0$. That's because cortical networks operate in the high gain regime, in which one excitatory action potential is capable of causing more than one excitatory action potential somewhere else in the network (Abeles, 1991; Matsumura *et al.*, 1996). The only way to stabilize such a system is to provide strong feedback from excitatory to inhibitory neurons, so that the inhibitory response to small increases in excitation dominates over the excitation (van Vreeswijk & Sompolinsky, 1996; Amit & Brunel, 1997b; Brunel & Wang, 2001). In terms of our network variables, this means that $d(J^{EI}\nu_I(\nu_E))/d\nu_E$ must be greater than $d(J^{EE}\nu_E)/d\nu_E$, which implies, via Eq. (4), that $\gamma'(\nu_E) > 0$. Thus, the condition $\partial\Delta\phi_E(\nu_E, m)/\partial\nu_E < 0$ is naturally satisfied in cortical networks.

Finally, we point out that a necessary condition for a low firing rate memory *and* a stable background state is that the ν_E-nullcline intersect the $m = 0$ line *above* the exchange-of-stability point (the point where the two branches of the m-nullcline intersect), and then intersect twice more on the unstable branch of the m-nullcline. A sufficient condition for this to happen is that the slope of the ν_E-nullcline is less than the slope of the m-nullcline at $m = 0$. If that condition is satisfied, then β can be increased until the ν_E-nullcline is sufficiently far above the exchange-of-stability point to guarantee two intersections on the unstable branch.

The slopes of the two nullclines, which can be determined by implicitly differentiating Eqs. (5a) and (5b), are given by

$$\left.\frac{d\nu_E}{dm}\right|_{m-\text{nullcline}} = \frac{1}{\gamma'}\,\frac{\beta\phi_E'(-\gamma+\beta m)-1}{\phi_E'(-\gamma+\beta m)-\phi_E'(-\gamma)} \tag{7a}$$

$$\left.\frac{d\nu_E}{dm}\right|_{\nu_E-\text{nullcline}} = \frac{\beta f\phi_E'(-\gamma+\beta m)}{1+\gamma'[(1-f)\phi_E'(-\gamma)+f\phi_E'(-\gamma+\beta m)]}\,. \tag{7b}$$

To calculate the slope of the m-nullcline at $m = 0$, it is necessary to take the $m \to 0$ limit of Eq. (7a); that will ensure that we do not pick up the vertical piece of the m-nullcline, which has infinite slope. Using the fact that $\beta\phi_E'(-\gamma + \beta m) = 1$ at the exchange-of-stability point

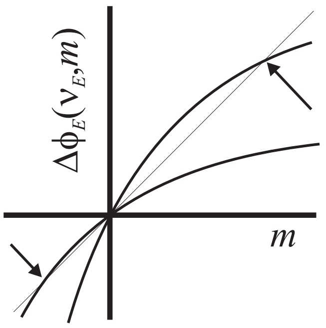

Figure 5: Concave gain functions. When the gain functions have no convex region, they intersect the 45° line (marked with arrows) with slope less than one, indicating that the network operates on the stable branch of the m-nullcline. Thus, the only way for the network to operate in the dynamically stabilized regime is for the gain functions to be convex for some range of m (see Fig. 3).

and Taylor expanding the numerator and denominator in Eq. (7a) around $m = 0$, we see that the slope of the m-nullcline is equal to β/γ' at $m = 0$. A simple calculation then tells us that the slope of the ν_E-nullcline at $m = 0$ is a factor of $f/(1 + 1/\gamma'\phi'_E)$ smaller. This ensures that, for some range of β, the nullclines will exhibit the set of equilibria shown in Fig. 4c.

We can now use Fig. 4 to understand why the two features listed in the beginning of Sec. 2 (convexity over some range and transition to concavity at firing rates well above 10-20 Hz) are necessary if attractor networks are to exhibit robust, low firing rate equilibria in the dynamically stabilized regime. First, if gain functions are not convex over some finite range of m, then they will intersect the 45° line with a slope that is less than 1 (disregarding the trivial intersection at $m = 0$), which eliminates the possibility of operating in the dynamically stabilized regime (see Fig. 5). Second, if the transition from a convex to a concave gain function occurs at an output rate that is less than, say, 20 Hz, then it would be possible for the equilibria in Figs. 4b-d to occur at firing rates less than 20 Hz, which in turn would imply that a stable, low firing rate equilibrium can exist on the stable branch of the m-nullcline, and thus not in the dynamically stabilized regime.

The above analysis tells us what is possible in principle; to see what is possible in practice, and to determine whether attractor networks can operate at low rates in a reasonably large parameter range, we now turn to simulations.

3 Simulations

To verify the reduced model described above, and to determine the range of parameters that supports multiple attractors, we performed simulations with a large network of synaptically coupled, spiking neurons. Network connectivity was based on the firing rate model given in Eq. (1), with two enhancements to make it more realistic. First, we used random, sparse background connectivity rather than uniform, all-all connectivity. Second, we allowed multiple attractors – multiple memories – rather than just a single memory. To implement multiple memories, we included in the connectivity matrix a term proportional to $\beta \sum_{\mu=1}^{p} \xi_i^{\mu}(\xi_j^{\mu} - f)$ where p is the number of memories and the ξ^{μ} are uncorrelated, random binary vectors with a fraction f of their components equal to one (Hopfield, 1982; Tsodyks & Feigel'man, 1988; Buhmann, 1989). This term is a natural extension of the one given in Eq. (1a). A detailed description of the network is provided in Appendix B.

We are interested not only in whether the network can exhibit memories at low rates, but also if it can do so without fine-tuning parameters. Ideally, we would like to explore the whole parameter space. However, there are 17 parameters (see Table I, Appendix B), making this prohibitively time-consuming. What we did instead was choose a network with reasonable parameters (synaptic and membrane time constants, PSP size, etc.; see Table I, Appendix B), and then vary two parameters over a broad range. The ones we varied were $V_{PSP_{EE}}$, the excitatory-to-excitatory EPSP size, and β, the increase in EPSP size among the neurons contained in each of the p memories. At a range of points in this two-dimensional space we checked how many memories could be embedded out of the p memories we attempted to embed, and whether any memories were spontaneously activated.

A run – a simulation at a particular set of parameters – lasted 12 seconds. The first 5 seconds consisted of background activity, in which the neurons fired at low average rate. At 5 seconds all neurons in one of the memories received a 100 ms barrage of EPSPs. After 2 seconds the same neurons received a second 100 ms barrage, this time of IPSPs. The network then ran for another 5 seconds at background.

A successful run – one in which the desired memory was activated and de-activated at the right times, and no other memories were spontaneously activated – is shown in Fig. 6a. Note that the firing rate during a memory is relatively low, about 14 Hz. Thus, for at least one set of parameters, an attractor can exist at low rate.

Not all runs were successful, of course. Two things can go wrong: the desired memory might not stay active for the whole two seconds (Fig. 6b), or a spurious memory might

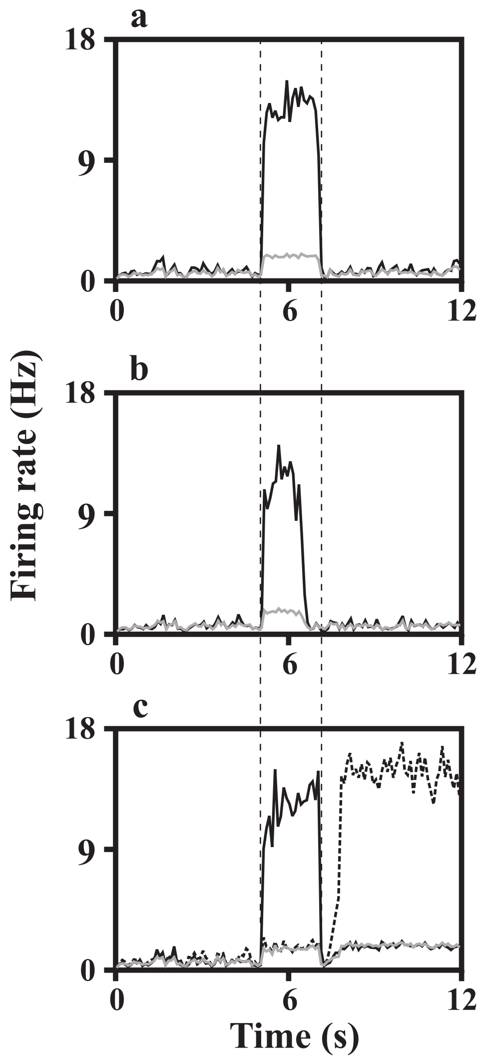

Figure 6: Simulations in which a set of memory neurons (black line) was activated at t=5 seconds (first vertical dashed line) and de-activated at t=7 seconds (second vertical dashed line). The gray line in all plots is the background firing rate. **a**. The activated memory is successfully embedded for the full two seconds. **b**. The activated memory lasts for only 1.5 seconds, indicating that the attractor is not stable. **c**. A spurious memory (dashed line) is spontaneously activated, indicating that the background is not stable. Parameters are given in Table I, Appendix B, with $V_{PSP_{EE}} = 0.48$ mV, $\beta = 0.21$ mV, and $f = 0.1$. Onset times for the memories were 50-100 ms, consistent with what is observed *in vivo* (Wilson *et al.*, 1993; Tomita *et al.*, 1999; Naya *et al.*, 2003).

become spontaneously active (Fig. 6c). To determine whether a particular set of parameters can support multiple memories without any of them becoming unstable, we activated and de-activated, one at a time, each of the p memories. A particular memory was considered to be successfully embedded if it stayed active for the requisite 2 seconds, and a particular set of parameters was considered to be stable if no spurious memories were activated in any of the p runs.

The results of simulations with $p = 50$ are summarized in Fig. 7. Figure 7a shows the number of memories that were successfully embedded versus $V_{PSP_{EE}}$ and β. The black line in this figure is the stability boundary: the background state is stable in the region below it,

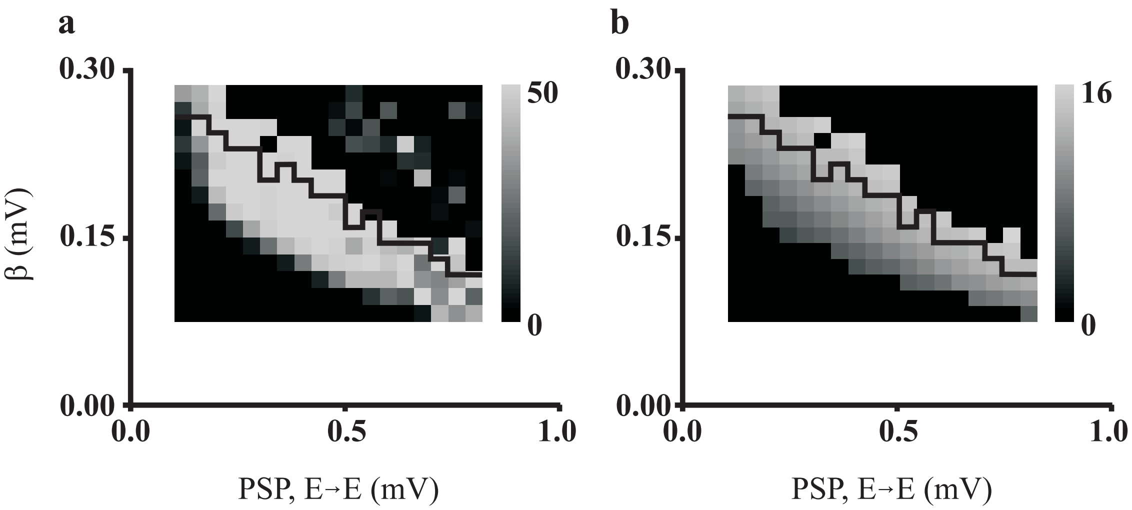

Figure 7: Summary of simulations. **a**. Number of memories successfully embedded out of 50, the number attempted. **b**. Average firing rate of memory neurons. Black line in both plots is the stability boundary: below it the background is stable, meaning no spurious memories were activated; above it the background is unstable. For EPSPs ranging from 0.2 to 0.5 mV, the network supports multiple attractors and a stable background for values of β varying over a 15-25% range. Scale bar is shown to the right. Parameters are given in Table I, Appendix B, with $f = 0.1$.

meaning no spurious memories were activated (Fig. 6a); it is unstable in the region above it, meaning at least one spurious memory was activated (Fig. 6c). Figure 7b shows the average firing rate of the memory neurons, for those memories that were successfully embedded. As predicted by the reduced model, the firing rate is uniformly low, never exceeding 15 Hz in the stable regime. The background rate (not shown) was ~0.1-0.2 Hz; lower than what is seen *in vivo*. For these network parameters, 50 memories was close to capacity: increasing p to 75 drastically reduced the size of the stable regime.

Consistent with the analysis of the reduced model, the parameter regime that supports multiple memories and a stable background is large: for EPSPs ranging from 0.2 to 0.5 mV, the region with multiple (> 45) memories extended about 0.04 mV below the stability boundary. If we consider the dynamic range of β to lie between zero and the stability boundary, then this corresponds to a parameter range for β of 15-25%.

Although the low firing rates of the memory neurons and the large parameter regime

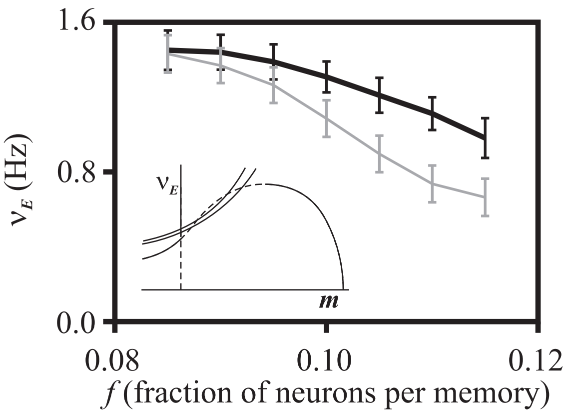

Figure 8: Average firing rate during the activation of a memory, ν_E, as a function of the fraction of neurons involved in a memory, f, for two parameter sets. Black curve: $V_{PSP_{EE}} = 0.40$ mV, $\beta = 0.18$ mV; gray curve: $V_{PSP_{EE}} = 0.20$ mV, $\beta = 0.21$ mV; other parameters are given in Table I, Appendix B. Averages are over all 50 memories; error bars are standard deviations. As indicated in the inset, the prediction of our model is that the firing rate, ν_E, should drop as f increases and drives the ν_E-nullcline up (the upper ν_E-nullcline corresponds to larger f; see text). This prediction is verified by the decreasing firing rate versus f.

that supports multiple memories is consistent with the reduced model introduced above, it is not a direct confirmation of it. What we would like to know is whether the equilibria we observed in the simulations really do lie on the unstable branch of the m-nullcline, as in Fig. 4a, or whether they in fact lie on the stable one, as in Figs. 4b-d. One way to find out is to manipulate the nullclines, and thus the equilibrium firing rates, and then check to see if the firing rates change as predicted by the reduced model.

A convenient manipulation is one in which the ν_E-nullcline is raised or lowered while the m-nullcline remains fixed. This is because raising the ν_E-nullcline lowers the average firing rate during the activation of a memory *only* when the equilibrium is on the unstable branch of the m-nullcline (Fig. 4a; inset in Fig. 8); if the equilibrium is on the stable branch, raising the ν_E-nullcline increases the average firing rate (Figs. 4b-d). Examining Eq. (5), we see that the ν_E-nullcline can be raised without affecting the m-nullcline by increasing f, the fraction of neurons involved in a memory. The prediction from the reduced model, then, is that the background firing rate during the activation of a memory should decrease as f increases. This prediction is verified in Fig. 8, thus providing strong evidence that the reduced model really does explain the simulations, and that the simulations operate in a regime corresponding to the nullclines shown in Fig. 4a. Note also that the range of f that supports stable memories is large, $\sim 30\%$, providing further evidence for the robustness of the network.

4 Discussion

The question we addressed in this paper was: how can cortical networks, which are highly interconnected and dominated by excitatory neurons, carry out computations and yet maintain stability? We answered this question in the context of attractor networks – networks that support multiple stable states, or "memories". We chose attractor networks for two reasons: First, they are thought to underlie several computations in the brain, including associative memory (Hopfield, 1982; Hopfield, 1984), working memory (Amit & Brunel, 1997a; Brunel & Wang, 2001), and the vestibular-ocular reflex (Seung, 1996); consequently, they are an active area of experimental research (Miyashita & Hayashi, 2000; Aksay *et al.*, 2001; Ojemann *et al.*, 2002; Naya *et al.*, 2003). Second, the stability problem in attractor networks is especially severe, so if we can solve this problem for attractor networks we should be able to solve it for networks implementing other kinds of computations.

In previous models of realistic attractors networks, the strategy was to operate on the stable branch of the m-nullcline (Amit & Brunel, 1997a; Wang, 1999; Brunel, 2000; Brunel & Wang, 2001), where firing rates are set essentially by the saturation rates of single neurons (Fig. 4c). Those rates are $\sim$100 Hz, making the experimentally observed 10-20 Hz (Fuster & Alexander, 1971; Miyashita & Chang, 1988; Nakamura & Kubota, 1995) hard to reach without fine-tuning parameters. Here we showed, using a reduced, two-variable model and simulations with large networks of spiking neurons, that attractor networks can operate on the unstable branch of the m-nullcline, and thus at low rates.

There are two key components to this model. The first is strong coupling between excitatory and inhibitory cells. This coupling, which is essential in networks with strong recurrent excitatory connections like those found in the cortex, means that any upward fluctuation in excitatory firing rate is matched by a larger upward fluctuation in inhibitory rate. (Here "larger" has a technical definition: it means that $\gamma'(\nu_E) > 0$; see Eqs. (4) and (6).) This makes the background state effectively inhibitory, a fact that is somewhat counterintuitive but extremely valuable, as it allows the background to act as a stabilizing pool. The second component is strong coupling between the attractors and the background state. Since the background is effectively inhibitory, this coupling can dynamically stabilize what would otherwise be an unstable state, and thus allow operation on the unstable branch of the m-nullcline.

The approach we used – in which the network operates on the unstable branch of the m-nullcline – is fundamentally different from approaches in which networks operate on the stable branch. In particular, the latter require fine-tuning of network parameters to keep neurons from firing near saturation (see Fig. 3c); the former does not. Importantly, the fine tuning problem associated with operation on the stable branch cannot be eliminated simply by including effects like synaptic and spike-frequency adaption, or by adding inhibition. This is because the fine-tuning problem is associated with the structure of the nullclines in Fig. 4 and the fact that neurons saturate at about 100 Hz (even with spike-frequency adaptation taken into account), neither of which are changed by these manipulations.

To verify the predictions of the reduced model, we performed large-scale simulations with networks of spiking neurons. We found that ~50 overlapping memories could be embedded in a network of 10,000 neurons. As predicted, the firing rates of the neurons in the attractor were low, between 10 and 15 Hz on average – approximately what is seen *in vivo*. More importantly, the network was stable over a broad range of parameters: background EPSP amplitude could vary by ~100%, the connection strength among the neurons involved in a memory could vary by ~25%, and the fraction of neurons involved in a memory could vary by ~30%, all without degrading network operation (Figs. 7 and 8).

4.1 Implications For Memory Storage. An important outcome of our model is that it predicts that the fraction of neurons involved in a memory, f, must be above some threshold. Otherwise, the feedback between the memory and the background would be too weak to stabilize activity at low rates, as can be seen by comparing Figs. 4a and 4c. As shown by (Gardner, 1988; Tsodyks & Feigel'man, 1988; Buhmann, 1989; Treves *et al.*, 1996), the maximum number of memories that can be embedded in a network is $\sim 0.2K/|f \log f|$ where K is the number of connections per neuron. This relation, combined with our result that f must be above some minimum, means that there is a limit to the number of memories that can be stored in any one network. This precludes the possibility of increasing the number of memories by adding neurons to a network while keeping the number of neurons in a memory fixed: this manipulation would decrease f and thus increase the maximum number of allowed memories, but, as f becomes too small, it would ultimately make the network unable to operate at low rates.

If we take $f = 0.1$ (Fuster & Jervey, 1982) and $K = 5000$ (Braitenberg & Schüz, 1991), then the maximum number of memories that can be stored in any strongly interconnected network in the brain is ~4000. This has major implications for how we store memories. Humans, for example, can remember significantly more than 4000 items, even within a particular category (words, for example). Thus, our model implies that memory must be highly distributed: local networks store at most thousands of items each, and during recall, those items must be dynamically linked to form a complete memory.

4.2 Biological Plausibility of our Model. Our simulations contained garden variety neurons and synapses – simple point neurons and fast synapses. This allowed us to cleanly separate behavior due to collective interactions from behavior due to single neuron properties. However, it leaves open the possibility that more realistic neurons and synapses might change our findings, especially the size of the parameter range in which the network is stable. Probably the most important parameters for stability are the excitatory and inhibitory synaptic time constants, with long excitatory time constants being stabilizing and long inhibitory time constants destabilizing (Tegner *et al.*, 2002). In our networks, we used the same time constant for both (3 ms). More realistic would be a longer inhibitory time constant (Salin & Prince, 1996; Xiang *et al.*, 1998) and a mix of long (NMDA) and short (AMPA) excitatory time constants (Andrasfalvy & Magee, 2001). Fortunately, such a mix has an overall stabilizing effect (Wang, 1999). Thus, we expect that with more realistic synapses, the parameter regime that supports a large number of memories should, in fact, be larger than the one we found.

Because slow excitatory synapses reduce fluctuations, they may increase the number of memories that can be embedded. This is important, since the number of memories in our network, 50, was smaller than what we expect of realistic networks, and much smaller than the theoretical maximum of 2000 for our network (derived using the above formula, $0.2K/|f \log f|$, with $K = 2500$ and $f = 0.1$).

The background firing rate in our simulations was ~0.1-0.2 Hz, lower than the 1-10 Hz observed *in vivo* (Fuster & Alexander, 1971; Miyashita & Chang, 1988; Nakamura & Kubota, 1995). These low rates can be traced to the fact that our network consisted of two pools of neurons, one near threshold and one well below threshold, with the latter pool firing at a very low rate. The below-threshold pool was needed to ensure that the average gain function, ϕ_E, was convex at the background firing rate, which is necessary for memories to exist (see Fig. 3). With other methods for achieving convex gain functions, such as coherent external input or nonlinearities in the current-to-firing-rate transformation (especially the nonlinearity provided by persistent neurons (Egorov *et al.*, 2002)), it is likely that a larger background rate could be supported. A nonlinear current-to-firing-rate transformation is especially helpful: when we included even a slight nonlinearity, the background rate increased to about 1 Hz and the capacity of the network increased to 100 memories (data not shown). Alternatively, it is possible that our network was consistent with real cortex, and the firing rates *in vivo* are over-estimated. This could happen because experimental recordings are, by necessity, biased toward neurons that fire at detectable rates, whereas we averaged firing rates over all neurons in the network, some of which rarely fired. This possibility is strengthened by the recent finding that, based on energy considerations, average firing rates in the cortex are likely to be less than 1 Hz (Lennie, 2003).

4.3 Summary. These results demonstrate that dynamical stabilization can be used to embed multiple, overlapping, low firing rate attractors over a broad range of network parameters. This opens the door to detailed comparison between theory and experiment, and should help resolve the question of whether attractor networks really do exist in the brain. In addition, dynamical stabilization can be used for all kinds of networks – not just those involving attractors – and may turn out to be a general computational principle.

Acknowledgments

The authors would like to thank Nicolas Brunel and Alex Pouget for insightful comments on the manuscript. This work was supported by NIMH Grant R01 MH62447.

Appendix A: Derivation of reduced model equations

In large neuronal networks in which neurons are firing asynchronously, it is reasonable to model neurons by their firing rates (Amit & Brunel, 1997a). While this approach is unlikely to capture the full temporal network dynamics (Treves, 1993), it is useful for studying equilibria. Moreover, near an equilibrium we expect the firing rates to obey first order dynamics. Thus, a firing rate model that uses linear summation followed by a nonlinearity would be described by the equations

$$\tau_E \frac{d\nu_{Ei}}{dt} + \nu_{Ei} = \phi_{Ei}\left(\sum_j J_{ij}^{EE}\nu_{Ej} - \sum_j J_{ij}^{EI}\nu_{Ij}\right) \tag{A.1a}$$

$$\tau_I \frac{d\nu_{Ii}}{dt} + \nu_{Ii} = \phi_{Ii}\left(\sum_j J_{ij}^{IE}\nu_{Ej} - \sum_j J_{ij}^{II}\nu_{Ij}\right) \tag{A.1b}$$

where ν_{Ei} and ν_{Ii} are the firing rates of individual excitatory and inhibitory neurons, respectively, ϕ_{Ei} and ϕ_{Ii} are their gain functions, and J_{ij} determines the connection strength from neuron j to neuron i.

The behavior of the network described by Eq. (A.1) depends critically on connectivity. If the connectivity is purely random, then the network can be described qualitatively by the Wilson and Cowan model (Wilson & Cowan, 1972). However, if the connectivity among a sub-population of excitatory neurons is strengthened, as it is in this study, we need to augment the Wilson and Cowan model. Ignoring firing rate fluctuations, an approximation that is valid qualitatively (Amit & Brunel, 1997a; Latham, 2002), we can construct heuristically such an augmented model by letting

$$J_{ij}^{EE} \rightarrow N_E^{-1}J^{EE} + \beta[N_E f(1-f)]^{-1}\xi_i(\xi_j - f) \tag{A.2a}$$

$$J_{ij}^{LM} \rightarrow N_M^{-1}J^{LM}, \qquad LM = EI, IE, II\,. \tag{A.2b}$$

In these expression, N_E and N_I are the number of excitatory and inhibitory neurons, respectively, and ξ is a random binary vector: ξ_i is 1 with probability f and 0 with probability $1-f$. If we apply the replacements given in Eq. (A.2) to Eq. (A.1), average Eq. (A.1b) over index, i, define $\phi_I(x) \equiv N_I^{-1}\sum_i \phi_{Ii}(x)$, and let $\phi_{Ei} \rightarrow \phi_E$, then Eq. (A.1) turns into Eq. (1). Replacing ϕ_{Ei} with ϕ_E was done for convenience only: it simplifies the resulting equations without detracting from the main result.

To derive Eqs. (3a) and (3b) from Eq. (1), we must average ϕ_E and $\xi_i \phi_E$ over index. To perform these averages, we first use the definition of m, Eq. (2), to simplify the argument of ϕ_E; this definition implies that $\phi_E = \phi_E(J^{IE}\nu_E - J^{II}\nu_I + \beta m \xi_i)$. We then note that for any function $F(\xi_i)$,

$$\begin{aligned}
N_E^{-1} \sum_i F(\xi_i) &= N_E^{-1} \sum_i (1-\xi_i) F(\xi_i) + N_E^{-1} \sum_i \xi_i F(\xi_i) \\
&= N_E^{-1} \sum_i (1-\xi_i) F(0) + N_E^{-1} \sum_i \xi_i F(1) \\
&= (1-f)F(0) + fF(1) \, .
\end{aligned} \tag{A.3}$$

The second line follows because ξ_i is either 0 or 1; the third (which is strictly valid only in the $N_E \to \infty$ limit) because ξ_i averages to f. Similarly,

$$(fN_E)^{-1} \sum_i \xi_i F(\xi_i) = (fN_E)^{-1} \sum_i \xi_i F(1) = F(1) \, . \tag{A.4}$$

Equations (A.3) and (A.4), along with the definition of m given in Eq. (2), can be used to derive Eqs. (3a) and (3b).

Appendix B: Simulation details

The network we simulate consists of N_E excitatory and N_I inhibitory quadratic integrate-and-fire neurons (Ermentrout & Kopell, 1986; Ermentrout, 1996; Gutkin & Ermentrout, 1998; Brunel & Latham, 2003). The membrane potential of the i^{th} neuron, V_i, and the conductance change at neuron i in response to a spike at neuron j, s_{ij}, evolve according to

$$\tau \frac{dV_i}{dt} = \frac{(V_i - V_r)(V_i - V_t)}{V_t - V_r} + V_{0i} - (V_i - \mathcal{E}_E) \sum_{j \in E} J_{ij} s_{ij}(t) - (V_i - \mathcal{E}_I) \sum_{j \in I} J_{ij} s_{ij}(t) \tag{B.1a}$$

$$\frac{ds_{ij}}{dt} = -\frac{s_{ij}}{\tau_s} + \sum_l \delta(t - t_j^l) \, . \tag{B.1b}$$

Here τ is the cell time constant, V_r and V_t are the nominal resting and threshold voltages, V_{0i} is the product of the applied current and the membrane resistance (in units of voltage), J_{ij}

is the (dimensionless) connection strength from cell j to cell i, $\mathcal{E}_E$ and $\mathcal{E}_I$ are the excitatory and inhibitory reversal potentials, respectively, the notation $j \in M$ means sum over only those cells of type M, $\delta(\cdot)$ is the Dirac δ-function, and t_j^l is the l^{th} spike emitted by neuron j.

To mimic the heterogeneity seen in cortical neurons, we let V_0, which determines the distance from resting membrane potential to threshold, have a range of values. This range is captured by the distributions

$$
\begin{aligned}
P_E(V_0) &= 0.75\,\frac{\exp\left[-(V_0-1.5)^2/2(0.5)^2\right]}{[2\pi(0.5)^2]^{1/2}} + 0.25\,\frac{\exp\left[-(V_0-3.75)^2/2(1.0)^2\right]}{[2\pi(1.0)^2]^{1/2}} \\
P_I(V_0) &= \frac{1}{4.5} \times \begin{cases} 1 & 0.5 < V_0 < 5.0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

where $P_E(V_0)$ and $P_I(V_0)$ are the distributions for excitatory and inhibitory neurons, respectively. In the absence of synaptic drive, the distance between resting membrane potential and threshold is $(V_t - V_r)[1 - 4V_0/(V_t - V_r)]^{1/2}$ (see Eq. (B.1a)). Since we use $V_t - V_r = 15$ mV (Table I), $V_0 = 3.75$ corresponds to a resting membrane potential that is equal to threshold while $V_0 = 1.5$ corresponds to a resting membrane potential about 12 mV below threshold. Neurons for which $V_0 > 3.75$ are endogenously active – they fire repeatedly without input.

The distribution $P_E(V_0)$ tells us that the excitatory neurons consist of two pools: one with resting membrane potential well below threshold, and one with resting membrane potential near threshold. About half the neurons in the latter pool are above threshold, and thus endogenously active. In realistic networks, this endogenous activity could be due to external input, intrinsic single neurons properties (Latham *et al.*, 2000b), or a combination of the two. While endogenously active cells greatly facilitate the existence of a low-firing rate background state (Latham *et al.*, 2000a), they are not absolutely necessary for it (Hansel & Mato, 2001).

A spike is emitted from neuron j whenever V_j reaches $+\infty$, at which point it is reset to $-\infty$. To attain the values $\pm\infty$ in our numerical integration, we make the change of variables $V = (V_r + V_t)/2 + (V_t - V_r)\tan\theta$ and integrate θ instead of V.

The connectivity matrix, J_{ij}, consists of two parts. One is random, corresponding to background connectivity before learning; the other is structured, corresponding to memories. In addition, we impose sparse connectivity by making the connection probability less than 1. We thus write

$$J_{ij} = g\Big(c_{ij}(W_{ij} + A_{ij})\Big)$$

where c_{ij} is 1 with probability c_{connect} and zero otherwise, W_{ij} corresponds to the background connectivity, A_{ij} corresponds to the structured connectivity, and g is a clipping function chosen so that $0 \leq g(x) \leq g_{\max}$; for convenience we choose $g(x)$ to be threshold linear and truncated at $g_{\max}$: $g(x) = \max(x, 0) - \max(x - g_{\max}, 0)$. The clipping function ensures that a particular connection strength neither violates Dale's law by falling below zero nor exceeds physiological levels by becoming too large.

The random part of the connectivity matrix, W, is chosen to produce realistic post-synaptic potentials. Specifically, we let

$$W_{ij} = w_{ij} \frac{V_{PSP_{LM}}}{V_{\scriptscriptstyle M}}$$

where neuron i is of type L, neuron j is of type M $(L, M = E, I)$, w_{ij} is a random variable uniformly distributed between $1 - \sqrt{3}\Delta$ and $1 + \sqrt{3}\Delta$ (so that its variance is Δ^2), and

$$V_{\scriptscriptstyle M} \equiv \frac{\mathcal{E}_{\scriptscriptstyle M} - V_r}{(\tau/\tau_s)\exp[\ln(\tau/\tau_s)/(\tau/\tau_s - 1)]} \, .$$

With this choice for the connection matrix, a neuron of type L will exhibit post-synaptic potentials on the order of $V_{PSP_{LM}}$ when a neuron of type M fires, assuming that $V_0 = 0$ for the neuron of type L (Latham *et al.*, 2000a).

The structured part of the connectivity matrix is the natural multi-memory extension of the matrix given in Eq. (1a), except with a slightly different normalization,

$$A_{ij} = \frac{\beta/V_{\scriptscriptstyle E}}{N_{\scriptscriptstyle E} f(1-f)} \sum_{\mu=1}^{p} \xi_i^\mu (\xi_j^\mu - f)$$

where p is the number of memories and the ξ^μ are uncorrelated, random binary vectors: ξ_i^μ is 1 with probability f and 0 with probability $1 - f$. Only excitatory neurons participate in the memories. The factor $1/V_{\scriptscriptstyle E}$ is included so that β has units of voltage.

The parameters in the model are listed in Table I. All were fixed throughout the simulation except for β and $V_{PSP_{EE}}$, which were varied to explore robustness (see Fig. 7), and f, which was 0.1 except in Fig. 8, where it ranged from 0.085 to 0.115.

Table I. Parameters used in the simulations.

Excitatory neurons	8,000
Inhibitory neurons	2,000
c_{connect}	0.25
Δ	0.25
τ	10 ms
τ_s	3 ms
V_r	-65 mV
V_t	-50 mV
$\mathcal{E}_E$	0 mV
$\mathcal{E}_I$	-80 mV
V_{EPSP} (E $\rightarrow$ E)	0.2-0.8 mV
V_{EPSP} (E $\rightarrow$ I)	1.0 mV
V_{IPSP} (I $\rightarrow$ E, I $\rightarrow$ I)	-1.5 mV
$g_{\max}$	2.5 mV
β	0.08-0.28 mV
p	50
f	0.085-0.115
Time step	0.5 ms

References

Abbott, L.F., & van Vreeswijk, C. Asynchronous states in networks of pulse-coupled oscillators. *Phys. Rev. E* **48**:1483–1490, 1993.

Abeles, M. *Corticonics: neural circuits of the cerebral cortex.* Cambridge, UK: Cambridge University Press, 1991.

Aksay, E., Gamkrelidzek, G., Seungk, H.S., Baker, R., & Tank, D.W. In vivo intracellular recording and perturbation of persistent activity in a neural integrator. *Nature Neurosci.* **1**:184–193, 2001.

Amit, D.J., & Brunel, N. Dynamics of a recurrent network of spiking neurons before and following learning. *Network* **8**:373–404, 1997a.

Amit, D.J., & Brunel, N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* **7**:237–252, 1997b.

Andrasfalvy, B.K., & Magee, J.C. Distance-dependent increase in AMPA receptor number in the dendrites of adult hippocampal CA1 pyramidal neurons. *J. Neurosci.* **21**:9151–9159, 2001.

Bell, G.S., & Sander, J.W. The epidemiology of epilepsy: the size of the problem. *Seizure* **10**:306–314, 2001.

Braitenberg, V., & Schüz, A. *Anatomy of the Cortex.* Berlin: Spring-Verlag, 1991.

Brunel, N. Persistent activity and the single-cell frequency-current curve in a cortical network model. *Network: Computation in Neural Systems* **11**:261–280, 2000.

Brunel, N., & Latham, P.E. Firing rate of the noisy quadratic integrate-and-fire neuron. *Neural Comput.* **15**:2281–2306, 2003.

Brunel, N., & Sergi, S. Firing frequency of leaky intergrate-and-fire neurons with synaptic current dynamics. *J. Theor. Biol.* **195**:87–95, 1998.

Brunel, N., & Wang, X.J. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J. Comput. Neurosci.* **11**:63–85, 2001.

Buhmann, J. Oscillations and low firing rates in associative memory neural networks. *Phys. Rev. A* **40**:4145–4148, 1989.

Chance, F.S., Abbott, L.F., & Reyes, A.D. Gain modulation from background synaptic input. *Neuron* **35**:773–782, 2002.

Egorov, A.V., Hamam, B.N., Fransen, E., Hasselmo, M.E., & Alonso, A.A. Graded persistent activity in entorhinal cortex neurons. *Nature* **420**:173–178, 2002.

Ermentrout, B. Type I membranes, phase resetting curves, and synchrony. *Neural Comput.* **8**:979–1001, 1996.

Ermentrout, B., & Kopell, N. Parabolic bursting in an excitable system coupled with a slow oscillation. *SIAM J. Appl. Math.* **46**:233–253, 1986.

Fourcaud, N., & Brunel, N. Dynamics of the firing probability of noisy integrate-and-fire neurons. *Neural Comput.* **14**:2057–2110, 2002.

Freedman, D.J., Riesenhuber, M., Poggio, T., & Miller, E.K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**:312–316, 2001.

Fuster, J.M., & Alexander, G.E. Neuron activity related to short-term memory. *Science* **173**:652–654, 1971.

Fuster, J.M., & Jervey, J.P. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *J. Neurosci.* **2**:361–375, 1982.

Gardner, E. The space of interactions in neural network models. *J. Phys. A: Math. Gen.* **21**:257–270, 1988.

Gerstner, W., & van Hemmen, L. Coherence and incoherence in a globally coupled ensemble of pulse-emitting units. *Phys. Rev. Lett.* **71**:312–315, 1993.

Golomb, D., Rubin, N., & Sompolinsky, H. Willshaw model: Associative memory with sparse coding and low firing rates. *Phys. Rev. A* **41**:1843–1854, 1990.

Gutkin, B.S., & Ermentrout, B. Dynamics of membrane excitability determine interspike interval variability: a link between spike generation mechanisms and cortical spike train statistics. *Neural Comput.* **10**:1047–1065, 1998.

Hansel, D., & Mato, G. Existence and stability of persistent states in large neuronal networks. *Phys. Rev. Lett.* **86**:4175–4178, 2001.

Hauser, W.A. *Incidence and prevelance.* In: Epilepsy. A Comprehensive Textbook, Vol. 1, pp 47-58. Engel, J. and Pedley, T.A., eds. Philadelphia, PA: Lippincott-Raven, 1997.

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

Hopfield, J. J. Neurons with graded responses have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci.* **81**:3088–3092, 1984.

Latham, P.E. *Associative memory in realistic neuronal networks.* Advances in Neural Information Processing Systems **14**: MIT Press, Cambridge MA, 2002.

Latham, P.E., Richmond, B.J., Nelson, P.G., & Nirenberg, S.N. Intrinsic dynamics in neuronal networks. I. Theory. *J. Neurophysiol.* **83**:808–827, 2000a.

Latham, P.E., Richmond, B.J., Nirenberg, S.N., & Nelson, P.G. Intrinsic dynamics in neuronal networks. II. Experiment. *J. Neurophysiol.* **83**:828–835, 2000b.

Lennie, P. The cost of cortical computation. *Curr. Biol.* **13**:493–497, 2003.

Marsden, J., & McCracken, M. *The Hopf bifurcation and its applications.* Lecture notes in applied mathematical sciences, Vol. 18. Berlin-Heidelberg-New York: Springer-Verlag, 1976.

Mascaro, M., & Amit, D.J. Effective neural response function for collective population states. *Network* **10**:351–373, 1999.

Matsumura, M., Chen, D., Sawaguchi, T., Kubota, K., & Fetz, E.E. Synaptic interactions between primate precentral cortex neurons revealed by spike-triggered averageing of intracellular membrane potentials in vivo. *J. Neurosci.* **16**:7757–7767, 1996.

McCormick, D.A., Connors, B.W., Lighthall, J.W., & Prince, D.A. Comparitive electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol.* **54**:782–806, 1985.

McIntyre, D.C., Poulter, M.O., & Gilby, K. Kindling: some old and some new. *Epilepsy Res.* **50**:79–92, 2002.

Miyashita, Y., & Chang, H.S. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* **331**:68–70, 1988.

Miyashita, Y., & Hayashi, T. Neural representation of visual objects: encoding and top-down activation. *Curr. Opin. Neurobiol.* **10**:187–194, 2000.

Nakamura, K., & Kubota, K. Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. *J. Neurophysiol.* **74**:162–178, 1995.

Naya, Y., Yoshida, M., & Miyashita, Y. Forward processing of long-term associative memory in monkey inferotemporal cortex. *J. Neurosci.* **23**:2861–2871, 2003.

Ojemann, G.A., Schoenfield-McNeill, J., & Corina, D.P. Anatomic subdivisions in human temporal cortical neuronal activity related to recent verbal memory. *Nature Neurosci.* **5**:64–71, 2002.

Rinzel, J., & Ermentrout, G.B. *Models for excitable cells and networks.* In: Methods in neuronal modeling: from synapses to networks, pp 137-173. Koch, C. and Segev, I., eds. Cambridge, MA: MIT Press, 1989.

Rubin, N., & Sompolinsky, H. Neural networks with low local firing rates. *Europhys. Lett.* **10**:465–470, 1989.

Salin, P. A., & Prince, D. A. Spontaneous $GABA_A$ receptor-mediated inhibitory currents in adult rat somatosensory cortex. *J. Neurophysiol.* **75**:1573–1588, 1996.

Seung, H.S. How the brain keeps the eyes still. *Proc. Natl. Acad. Sci.* **93**:13339–13344, 1996.

Tegner, J., Compte, A., & Wang, X.J. The dynamical stability of reverberatory neural circuits. *Biol. Cybern.* **87**:471–481, 2002.

Tiesinga, P.H., Jose, J.V., & Sejnowski, T.J. Comparison of current-driven and conductance-driven neocortical model neurons with Hodgkin-Huxley voltage-gated channels. *Phys. Rev. E* **62**:8413–8419, 2000.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., & Y., Miyashita. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* **401**:699–703, 1999.

Treves, A. Mean-field analysis of neuronal spike dynamics. *Network* **4**:259–284, 1993.

Treves, A., Skaggs, W.E, & Barnes, C.A. How much of the hippocampus can be explained by functional constraints? *Hippocampus* **6**:666–674, 1996.

Tsodyks, M.V., & Feigel'man, M.V. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.* **6**:101–105, 1988.

van Vreeswijk, C., & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**:1724–1726, 1996.

Wang, X.J. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.* **19**:9587–9603, 1999.

Wilson, F.A., Scalaidhe, S.P., & Goldman-Rakic, P.S. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* **260**:1955–1958, 1993.

Wilson, H.R., & Cowan, J.D. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* **12**:1–24, 1972.

Xiang, Z., Huguenard, J.R., & Prince, D.A. $GABA_A$ receptor-mediated currents in interneurons and pyramidal cells of rat visual cortex. *J. Physiol.* **506**:715–730, 1998.