Associative memory in realistic neuronal networks

P.E. Latham*
Department of Neurobiology
University of California at Los Angeles
Los Angeles, CA 90095
pel@ucla.edu

Abstract

Almost two decades ago, Hopfield [1] showed that networks of highly reduced model neurons can exhibit multiple attracting fixed points, thus providing a substrate for associative memory. It is still not clear, however, whether realistic neuronal networks can support multiple attractors. The main difficulty is that neuronal networks in vivo exhibit a stable background state at low firing rate, typically a few Hz. Embedding attractor is easy; doing so without destabilizing the background is not. Previous work [2, 3] focused on the sparse coding limit, in which a vanishingly small number of neurons are involved in any memory. Here we investigate the case in which the number of neurons involved in a memory scales with the number of neurons in the network. In contrast to the sparse coding limit, we find that multiple attractors can co-exist robustly with a stable background state. Mean field theory is used to understand how the behavior of the network scales with its parameters, and simulations with analog neurons are presented.

One of the most important features of the nervous system is its ability to perform associative memory. It is generally believed that associative memory is implemented using attractor networks - experimental studies point in that direction [4-7], and there are virtually no competing theoretical models. Perhaps surprisingly, however, it is still an open theoretical question whether attractors can exist in realistic neuronal networks. The “realistic” feature that is probably hardest to capture is the steady firing at low rates - the background state - that is observed throughout the intact nervous system [8-13]. The reason it is difficult to build an attractor network that is stable at low firing rates, at least in the sparse coding limit, is as follows [2, 3]:

Attractor networks are constructed by strengthening recurrent connections among sub-populations of neurons. The strengthening must be large enough that neurons within a sub-population can sustain a high firing rate state, but not so large that the sub-population can be spontaneously active. This implies that the neuronal gain functions - the firing rate of the post-synaptic neurons as a function of the average

*http://culture.neurobio.ucla.edu/∼pel
firing rate of the pre-synaptic neurons — must be sigmoidal: small at low firing rate to provide stability, high at intermediate firing rate to provide a threshold (at an unstable equilibrium), and low again at high firing rate to provide saturation and a stable attractor. In other words, a requirement for the co-existence of a stable background state and multiple attractors is that the gain function of the excitatory neurons be superlinear at the observed background rates of a few Hz [2, 3]. However — and this is where the problem lies — above a few Hz most realistic gain function are nearly linear or sublinear (see, for example, Fig. B1 of [14]).

The superlinearity requirement rests on the implicit assumption that the activity of the sub-population involved in a memory does not affect the other neurons in the network. While this assumption is valid in the sparse coding limit, it breaks down in realistic networks containing both excitatory and inhibitory neurons. In such networks, activity among excitatory cells results in inhibitory feedback. This feedback, if powerful enough, can stabilize attractors even without a saturating nonlinearity, essentially by stabilizing the equilibrium (above considered unstable) on the steep part of the gain function. The price one pays, though, is that a reasonable fraction of the neurons must be involved in each of the memories, which takes us away from the sparse coding limit and thus reduces network capacity [15].

1 The model

A relatively good description of neuronal networks is provided by synaptically coupled, conductance-based neurons. However, because communication is via action potentials, such networks are difficult to analyze. An alternative is to model neurons by their firing rates. While this is unlikely to capture the full temporal network dynamics [16], it is useful for studying equilibria. In such simplified models, the equilibrium firing rate of a neuron is a function of the firing rates of all the other neurons in the network. Letting \( v_E \) and \( v_I \) denote the firing rates of the excitatory and inhibitory neurons, respectively, and assuming that synaptic input sums linearly, the equilibrium equations may be written

\[
\begin{align*}
    v_E &= \phi_E \left( \sum_j A_{ij}^E v_E j, \sum_j A_{ij}^I v_I j \right), \\
    v_I &= \phi_I \left( \sum_j A_{ij}^E v_E j, \sum_j A_{ij}^I v_I j \right). 
\end{align*}
\]

Here \( \phi_E \) and \( \phi_I \) are the excitatory and inhibitory gain functions and \( A_{ij} \) determines the connection strength from neuron \( j \) to neuron \( i \). The gain functions can, in principle, be derived from conductance-based model equations [17].

Our goal here is to determine under what conditions Eq. (1) allows both attractors and a stable state at lower firing rate. To accomplish this we will use mean field theory. While this theory could be applied to the full set of equations, to reduce complexity we make a number of simplifications. First, we let the inhibitory neurons be completely homogeneous (\( \phi_I \) independent of \( i \) and connectivity to and from inhibitory neurons all-to-all and uniform). In that case, Eq. (1b) becomes simply \( v_I = \phi(v_E, v_I) \) where \( v_E \) and \( v_I \) are the average firing rates of the excitatory and inhibitory neurons. Solving for \( v_I \) and inserting the resulting expression into Eq. (1a) results in the expression \( v_E = \phi_E \left( \sum_j A_{ij}^E v_E j, A_{ij}^I v_I \right) \) where \( A_{ij}^{EI} \equiv \sum_j A_{ij}^{Ej} \).
Second, we let \( \phi_{ib} \) have the form \( \phi_{ib}(u, v) = \phi_{ib}(x_i + bu - cv) \) where \( x_i \) is a Gaussian random variable, and similarly for \( \phi_{i} \) (except with different constants \( b \) and \( c \) and no dependence on \( t \)). Finally, we assume that \( \phi_i \) is threshold linear and the network operates in a regime in which the inhibitory firing rate is above zero. With these simplifications, and a trivial redefinition of constants, Eq. (1a) becomes

\[
\nu_i = \beta p^{1/2} \phi \left( x_i - (a + 1)\nu + \sum_j A_{ij} \nu_j \right) .
\]

We have dropped the sub and superscript \( E \), since Eq. (2) refers exclusively to excitatory neurons, defined \( \nu \) to be the average firing rate, \( \nu \equiv N^{-1} \sum_i \nu_i \), and rescaled parameters. We let the function \( \phi \) be \( O(1) \), so \( \beta \) can be interpreted as the gain. The parameter \( p \) is the number of memories. The reduction from Eq. (1) to Eq. (2) was done solely to simplify the analysis; the techniques we will use apply equally well to the general case, Eq. (1).

Note that the gain function in Eq. (2) decreases with increasing average firing rate, since its argument is \(- (1 + a) \nu \) and \( a \) is positive. This negative dependence on \( \nu \) arises because we are working in the large coupling regime in which excitation and inhibition are balanced [18, 19]. The negative coupling to firing rate has important consequences for stability, as we will see below.

We let the connectivity matrix have the form

\[
A_{ij} = \frac{1}{(g) N} C_{ij} g(W_{ij} + J_{ij}) .
\]

Here \( N \) is the number of excitatory neurons; \( C_{ij} \), which regulates the degree of connectivity, is \( 1/c \) with probability \( c \) and and \( 0 \) with probability \( 1 - c \) (except \( C_{ii} = 0 \), meaning no autapses); \( g(z) \) is an \( O(1) \) clipping function that keeps weights from falling below zero or getting too large; \( (g) \) is the mean value of \( g(z) \), defined in Eq. (4) below; \( W_{ij} \), which corresponds to background connectivity, is a random matrix whose elements are Gaussian distributed with mean 1 and variance \( \delta w^2 \), and \( J_{ij} \) produces the attractors. We will follow the Hopfield prescription and write \( J_{ij} \) as

\[
J_{ij} = \frac{\epsilon}{\sqrt{p}} \sum_{\mu=1}^{p} \eta_{\mu i} \eta_{\mu j}.
\]

where \( \epsilon \) is the coupling strength among neurons involved in the memories, and the patterns \( \eta_{\mu i} \) determine which neurons participate in each memory. The \( \eta_{\mu i} \) are a set of uncorrelated vectors with zero mean and unit variance. In simulations we use \( \eta_{\mu i} = [(1 - f)/f]^{1/2} \) with probability \( f \) and \(- [(1 - f)/f]^{1/2} \) with probability \( 1 - f \), so a fraction \( f \) of the neurons are involved in each memory. Other choices are unlikely to significantly change our results.

2 Mean field equations

The main difficulty in deriving the mean field equations from Eq. (2) is separating the signal from the noise. Our first step in this endeavor is to analyze the noise
associated with the clipped weights. To do this we break \( C_{ij}g(W_{ij} + J_{ij}) \) into two pieces: \( C_{ij}g(W_{ij} + J_{ij}) = \langle g \rangle + \langle g' \rangle J_{ij} + \delta C_{ij} \) where

\[
\delta C_{ij} \equiv C_{ij}g(W_{ij} + J_{ij}) - \left( \langle g \rangle + \langle g' \rangle J_{ij} \right).
\]

The angle brackets around \( g \) represent an average over the distributions of \( W_{ij} \) and \( J_{ij} \), and a prime denotes a derivative. In the large \( p \) limit, \( \delta C_{ij} \) can be treated as a random matrix whose main role is to increase the effective noise [20]. The mean of \( \delta C_{ij} \) is zero and its variance normalized to \( \langle g \rangle^2 / \epsilon \), which we denote \( \sigma^2 \), is given by

\[
\sigma^2 \equiv \frac{\epsilon}{\langle g \rangle^2} \text{Var}[\delta C_{ij}] = \frac{\langle g^2 \rangle - \epsilon(1 + \langle g' \rangle^2 \langle J^2_{ij} \rangle)}{\langle g \rangle^2}.
\]

For large \( p \), the elements of \( J_{ij} \) are Gaussian with zero mean and variance \( \epsilon^2 \), so the averages involving \( g \) can be written

\[
\langle g^k \rangle = \int dz \frac{\exp[-z^2/2(\delta u^2 + \epsilon^2)]}{\sqrt{2\pi(\delta u^2 + \epsilon^2)^{1/2}} - g^k(1 + z)
\]

where \( k \) can be either an exponent or a prime and the “1” in \( g(1 + z) \) corresponds to the mean of \( W_{ij} \). In our simulations we use the clipping function \( g(z) = z \) if \( z \) is between 0 and 2, \( 0 \) if \( z \leq 0 \) and 2 if \( z \geq 2 \).

Our main assumptions in the development of a mean field theory are that \( \sum_{j \neq i} \delta C_{ij} \nu_j \) is a Gaussian random variable, and that \( \delta C_{ij} \) and \( \nu_j \) are independent. Consequently,

\[
\text{Var} \left[ \frac{1}{\langle g \rangle N} \sum_{j \neq i} \delta C_{ij} \nu_j \right] = \frac{\sigma^2}{cN} \langle \nu^2 \rangle
\]

where \( \langle \nu^2 \rangle \equiv N^{-1} \sum_i \nu_i^2 \) is the second moment of the firing rate. Letting \( \tilde{\theta}_i \) be a zero mean Gaussian random variable with variance \( \theta^2 \equiv \sigma^2 \langle \nu^2 \rangle / cN \), we can use the above assumptions along with the definition of \( J_{ij} \), Eq. (3), to write Eq. (20) as

\[
\nu_i = \beta p^{1/2} \phi \left( x_i - \alpha \nu + p^{-1/2} \epsilon \eta_i \sum_{j \neq i} \eta_{ij} \nu_j + \tilde{\theta}_i \right).
\]

We have defined the clipped memory strength, \( \epsilon_c \), as \( \epsilon_c \equiv \epsilon \langle g' \rangle / \langle g \rangle \). While it is not totally obvious from the above equations, it can be shown that both \( \sigma^2 \) and \( \epsilon_c \) become independent of \( \epsilon \) for large \( \epsilon \). This makes network behavior robust to changes in \( \epsilon \), the strength of the memories, so long as \( \epsilon \) is large.

Derivation of the mean field equations from Eq. (5) follow standard methods [21, 22]. For definiteness we take \( \phi(x) \) to be threshold linear: \( \phi(x) = \max(0, x) \). For the case of one active memory, the mean field equations may then be written in the form
\[ w = \frac{\beta \epsilon_c}{1-r} \Delta F_1(w, z) \]  
\[ 1 = \frac{\beta^2 \epsilon_c^2}{(1-r)^2} \left[ \frac{\sigma^2}{\epsilon_c^2} + \frac{1}{(1-q)^2} \right] [F_2(z) + f \Delta F_2(w, z)] \]  
\[ + \frac{\beta^2 \theta_0^2 \epsilon_c^2}{(1-r)^2} [F_1(z) + f \Delta F_1(w, z)]^2 \] 
\[ r = \frac{\alpha \beta \epsilon_c}{1-q} \] 
\[ q = \frac{\beta \epsilon_c}{1+\alpha \beta \epsilon_c} [F_0(z) + f \Delta F_0(w, z)] \]

where \( \alpha \equiv p/N \) is the load parameter, \( x_0 \) and \( \theta_0^2/p \) are the mean and variance of of \( x_i \) (see Eq. (2)), and, recall, \( f \) is the fraction of neurons that participate in each memory. The functions \( F_k \) and \( \Delta F_k \) are defined by

\[ F_k(z) \equiv \int_{-\infty}^{\infty} \frac{d\xi}{(2\pi)^{1/2}} (z + \xi)^k \exp(-\xi^2/2) \]
\[ \Delta F_k(w, z) \equiv F_k(w + z) - F_k(z). \]

For large negative \( z \), \( F_k(z) \) vanishes as \( \exp(-z^2/2) \), while for large positive \( z \), \( F_k(z) \to z^k/k! \).

The average firing rate, \( \nu \), and strength of the memory, \( m \equiv N^{-1} \sum_{i,j} \eta_{ij} \nu_j \) (taken without loss of generality to be the overlap with pattern 1), are given in terms of \( z \) and \( w \) as

\[ \nu = \frac{x_0}{\Delta F_1(w, z) + f \Delta F_1(w, z) / F_1(z) + f \Delta F_1(w, z)} \]
\[ m = \frac{\nu}{(1-f) \Delta F_1(w, z)} \]

3 Results

The mean field equations can be understood by examining Eqs. (6a) and (6b). The first of these, Eq. (6a), is a rescaled form of the equation for the overlap, \( m \). (From the definition of \( \Delta F_1 \) given above, it can be seen that \( m \) is proportional to \( w \) for small \( w \)). This equation always has a solution at \( w = 0 \) (and thus \( m = 0 \)), which corresponds to a background state with no memories active. If \( \beta \epsilon_c \) is large enough, there is a second solution with \( w \) (and thus \( m \)) greater than zero. This second solution corresponds to a memory. The other relevant equation, Eq. (6b), describes the behavior of the mean firing rate. This equation looks complicated only because the noise – the variation in firing rate from neuron to neuron – must be determined self-consistently.

The solutions to Eqs. (6a) and (6b) are plotted in Fig. 1 in the \( z-w \) plane. The solid lines, including the horizontal line at \( w = 0 \), represents the solution to Eq. (6a), the