Current Biology 22, 2095–2103, November 20, 2012 ©2012 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2012.08.058

Article

Neuronal Circuits Underlying Persistent Representations Despite Time Varying Activity

Shaul Druckmann[1] and Dmitri B. Chklovskii[1,*]
[1]Janelia Farm Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20176, USA

Summary

Background: Our brains are capable of remarkably stable stimulus representations despite time-varying neural activity. For instance, during delay periods in working memory tasks, while stimuli are represented in working memory, neurons in the prefrontal cortex, thought to support the memory representation, exhibit time-varying neuronal activity. Since neuronal activity encodes the stimulus, its time-varying dynamics appears to be paradoxical and incompatible with stable network stimulus representations. Indeed, this finding raises a fundamental question: can stable representations only be encoded with stable neural activity, or, its corollary, is every change in activity a sign of change in stimulus representation?

Results: Here we explain how different time-varying representations offered by individual neurons can be woven together to form a coherent, time-invariant, representation. Motivated by two ubiquitous features of the neocortex—redundancy of neural representation and sparse intracortical connections—we derive a network architecture that resolves the apparent contradiction between representation stability and changing neural activity. Unexpectedly, this network architecture exhibits many structural properties that have been measured in cortical sensory areas. In particular, we can account for few-neuron motifs, synapse weight distribution, and the relations between neuronal functional properties and connection probability.

Conclusions: We show that the intuition regarding network stimulus representation, typically derived from considering single neurons, may be misleading and that time-varying activity of distributed representation in cortical circuits does not necessarily imply that the network explicitly encodes time-varying properties.

Introduction

The representation afforded by neural circuits can be remarkably stable; we are able to maintain fixed representation of transient stimuli in working memory for long periods of time [1], and our actions can unfold slowly in a continuous, reliable manner over behavioral time scales of seconds [2]. Yet underlying this stable network stimulus representation is neuronal activity that can vary on much faster time scales down to a few to tens of milliseconds [3–9]. Thus, to maintain stable representations across behaviorally relevant time scales the brain must somehow weave together the wildly time varying activity of individual neurons across the network into a stable whole.

Perhaps the strongest demonstration of network stimulus representation stability can be found in working memory tasks, such as the delayed match to sample task [10]. In this paradigm, an animal is presented two transient stimuli, e.g., briefly flashed images, separated by a delay period, and must decide whether the images are the same or different. Clearly, in order to succeed the animal must maintain a stable representation of the first transient stimulus during the delay period. This has been attributed to neurons in the prefrontal cortex which exhibit elevated firing rates during the delay period [11–13].

While early work emphasized that neurons show sustained activity during delay periods [10, 11], more-recent work demonstrated that most neurons actually have complex, time-varying dynamics during the delay period [14–17]. What is the significance of this variability? One explanation is that the variability in activity is due to explicit representation of different properties of the environment that are in fact time varying, such as the amount of time passed since stimulus presentation [14, 18]. Accordingly, changes in neuronal activity are a consequence of changes in these additional parameters.

This interpretation raises a fundamental question: is every change in neuronal activity evidence of a change in some (known or unknown) property of the network stimulus representation? Since the activity of each neuron in a network represents some property of the environment (or some internal state), it seems that the answer should be yes and that indeed changes in activity should lead to changes in network stimulus representation.

Here, we show in contrast that the answer to this question is actually *no*; time-varying neural activity is perfectly consistent with a fixed, time-invariant network stimulus representation. Intuitively speaking, if multiple neurons represent overlapping properties of the stimulus, a change in the contribution of one neuron due to its time-varying activity can be compensated for by an appropriate change in the other neurons. Thus, though the contribution of each neuron to the network stimulus representation changes due to its varying activity, the network can still maintain a stable network stimulus representation.

A central result of this paper is a derivation of the network architecture that implements such compensation automatically, thus explicitly demonstrating that stable representation of a multidimensional input stimulus is perfectly compatible with time varying activity. The neuronal dynamics of such network can be seen as a generalization of the line attractor [19] to multidimensional stimuli. Unexpectedly, our network architecture possesses many of the structural properties measured in cortical networks.

Results

Persistent Network Stimulus Representations by Feature Vector Recombination

In order to show that the network architecture we propose is sufficient to achieve persistent representations despite time varying activity and that no complicated dynamics or representation are required, we begin by considering the simplest forms of dynamics and representation: linear rate models

*Correspondence: mitya@janelia.hhmi.org

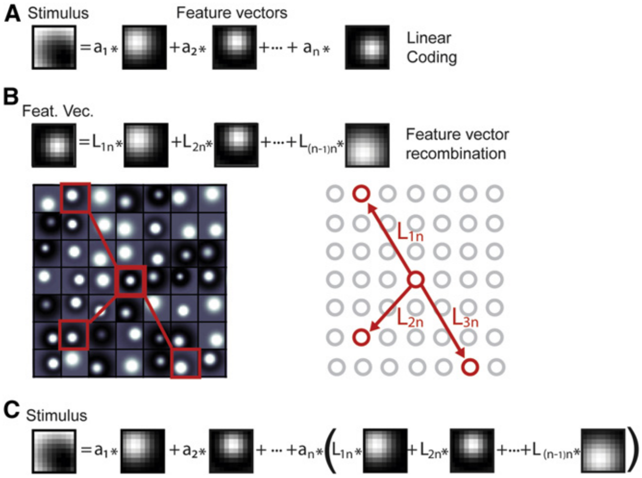

Figure 1. FEVER Principle

(A) A stimulus can be linearly encoded by a sum over many neurons of the feature vector times the activity of that neuron.

(B) In a similar fashion, each feature vector can be expressed as a weighted sum of other feature vectors. In a FEVER network, the weights of each neuron's outgoing synaptic connections are given by the coefficients of its feature vector representation in terms of other neurons' feature vectors.

(C) In the FEVER network, as the activity of a neuron decays (in this case the last neuron), its contribution to the network stimulus representation is recovered by elicited activity in other neurons.

See also Figure S1.

and linear population coding. In the linear decoding approach [20] to the network activity [21], the activity of each of *n* neurons represents a certain feature of the stimulus, which can be viewed as a feature vector, $\vec{d_i}$ (i = 1..n), in the space of stimuli.

In the linear coding framework, the stimulus represented by the network, $\vec{s}$, is found as a linear combination of each neuron's feature vector, $\vec{d_i}$, weighted by its activity, a_i (Figure 1A):

$$\vec{s} = \sum_i \vec{d_i} a_i. \qquad \text{(Equation 1)}$$

For instance, the frequency of a vibrotactile stimulus, which would be considered one direction in the space of stimuli, has been found to be encoded monotonically (and nearly linearly) in the activity of a subset of prefrontal neurons [15]. In sensory areas, feature vectors correspond to the direction in stimulus space to which a neuron responds most strongly and can be experimentally determined, e.g., by reverse correlation [22].

In order for the network to support working memory, the representation should be constant in time, i.e., its temporal derivative must be zero:

$$0 = \frac{d\vec{s}}{dt} = \frac{d}{dt}\sum_i \vec{d_i} a_i = \sum_i \vec{d_i}\frac{da_i}{dt}. \qquad \text{(Equation 2)}$$

The implication of this equation for the activity of individual neurons depends on the relationship between the different feature vectors, $\vec{d_1}$, $\vec{d_2}$, etc. First, any changes in a neuron's activity will change its contribution to the network stimulus representation. If the feature vectors, $\vec{d_i}$, are linearly independent, no other neuron in the network, or combination of neurons, can compensate for this change (see the Experimental Procedures). Thus, the only way to satisfy Equation 2 for all stimuli $\vec{s}$ is to hold the activity of each neuron constant in time. This observation explains the prevalent notion that in order to support a time-invariant network stimulus representation the activity must be time invariant, as is indeed the case in some systems [23].

In contrast, in a redundant representation, where the vectors $\vec{d_i}$ are linearly dependent, the network stimulus representation can remain constant despite changing activity (see the Supplemental Experimental Procedures). Since the number of cortical neurons greatly exceeds the number of thalamic neurons projecting to them [24–26], cortical representation is almost certainly redundant. In a redundant representation even when the contribution of each neuron to the network stimulus representation changes due to its varying activity, the network as a whole can maintain a stable representation.

Here, we propose a network architecture that maintains stable representation despite time-varying activity. In this architecture, each neuron's lateral connections drive the other neurons in a way that reintroduces into the network stimulus representation the component that is lost by the neuron's own change in activity. We stress that this is achieved not by keeping the activity of each neuron constant, but by keeping the representation in the network *as a whole* constant, despite individual neurons' fluctuating activity. Mathematically, this is possible if the representation matrix, $D = \{\vec{d_1}, \vec{d_2}\ldots\}$, has a null space, i.e., a set of activity vectors $\vec{a}$, such that $D\vec{a} = 0$. Representation will remain constant if the dynamics are confined to the null space. Our architecture ensures that this is the case.

Let us give a mathematical derivation of such architecture assuming linear firing rate dynamics:

$$\tau\frac{da_i}{dt} = -a_i + \sum_j L_{ij}a_j, \qquad \text{(Equation 3)}$$

where a_i is the activity of the i^{th} neuron, τ its time constant, and L denotes the lateral connectivity matrix, adopting the convention that the element L_{ij} represents a weight of the synapses from neuron j to neuron i. By substituting Equation 3, the dynamics equation, into Equation 2, the constancy-of-representation condition equation, we obtain

$$0 = \sum_i \vec{d_i}\left(-a_i + \sum_j L_{ij}a_j\right) = -\sum_i \vec{d_i}a_i + \sum_j a_j \sum_i \vec{d_i}L_{ij} = \sum_i\left(-\vec{d_i} + \sum_j \vec{d_j}L_{ji}\right)a_i. \qquad \text{(Equation 4)}$$

If this is to hold for all activity patterns, $\vec{a}$, we obtain (Figure 1B)

$$\vec{d_i} = \sum_j \vec{d_j}L_{ji},\ \textit{for all}\ 1 \le i \le n. \qquad \text{(Equation 5)}$$

Expressing this mathematical relationship in words, if the sum of each neuron's outgoing synaptic weights times the feature vector of the corresponding postsynaptic neurons is equal to the neuron's feature vector, the compensation is automatically accomplished (Figures 1B and 1C). We term this principle *f*eature *ve*ctor *r*ecombination (FEVER) and a network that obeys this principle a FEVER network. Mathematically, the right eigenvectors of L must either have unit

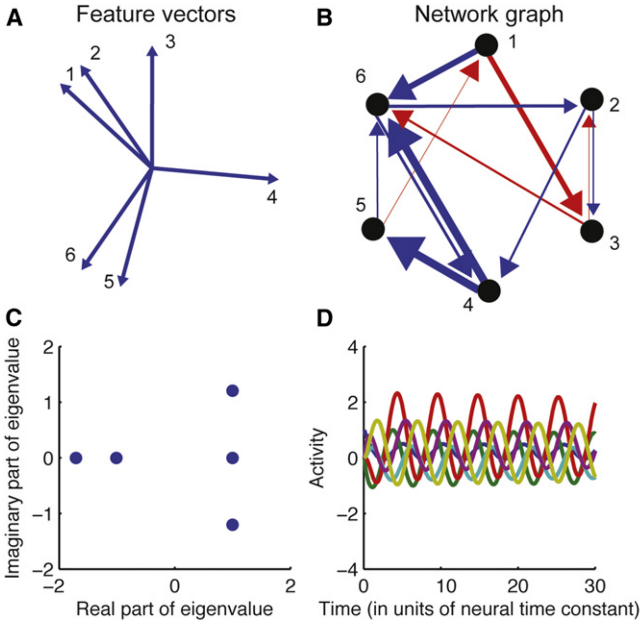

Figure 2. Oscillatory Activity in a Simple FEVER Network

(A) Simple set of feature vectors from which the FEVER network will be constructed.

(B) Graph of the FEVER network.

(C) Eigenvalues of the FEVER network. Note that two eigenvalues are located at real part 1, imaginary part 0. Since they occur at the same position, only one can be seen in the plot.

(D) Oscillatory network activity of the FEVER network.

See also Figure S2.

eigenvalue, resulting in integrating modes [19], or they must reside in the null space of the representation matrix D (see the Supplemental Experimental Procedures). In summary, the calculations above describe the structure of networks that are capable of maintaining persistent network stimulus representation despite individual neurons' variable activity.

To illustrate how representation can stay constant while activity varies, even oscillates, we consider a simple example of a circuit representing a two-dimensional stimulus. In this case, the preferred direction of each neuron is a direction on a plane (Figure 2A) corresponding to the two dimensions of the stimulus. The FEVER principle requires synaptic connections such that each neuron's preferred direction can be expressed as a weighted sum over the preferred directions of the other neurons, where the vector weights are the strengths of outgoing synaptic connections (Figure 2B). For the example of the neuron coding for the "up" direction in space (neuron 3), this can be accomplished by summing the receptive fields of any two neurons in an appropriately scaled way so that the horizontal contributions will cancel out, leaving only the vertical contribution, which is equal to the neuron's own feature vector as required. Accordingly, there are two "modes" by which the up direction can be encoded: either through the activity of neuron 3 or through the combined activity of the postsynaptic neurons.

The dynamics of such network can be analyzed by considering its eigenspectrum (Figure 2C). As expected for a FEVER network representing a two-dimensional stimulus, there are two eigenvalues at unity, corresponding to the two dimensions of the stimulus that are stably encoded. After stimulus presentation at time zero, the network has sustained complex activity for over 30 neural time constants (Figure 2D), as would be expected from complex eigenvalues with real values close (but not equal to) unity, yet stimulus representation remains perfectly stable (in other FEVER networks, long transients can be the result of nonnormal dynamics [27, 28] as well; Figure S2).

Dynamics of Biologically Realistic FEVER Networks

The above network is only one of many possible solutions satisfying the FEVER principle. Since in Equation 5 the number of equalities (number of entries in each vector d_i) is smaller than the number of unknowns (number of rows of L), the number of solutions is infinite. For example, there is a trivial solution in which the connectivity matrix is set to the identity matrix, corresponding to a network that has no real lateral connections and only autapses. We do not consider this trivial solution any further since it is not truly a network model (the neurons not being connected to each other) and due to its incompatibility with the known fact of cortical neurons having numerous lateral connections and few autapses [25, 29].

Here we propose a specific choice of connectivity in FEVER networks. Since creating and maintaining a synaptic contact requires energy and takes up limited volume in the densely packed cortex [25, 30–32], a natural choice would be to find the pattern of synaptic connectivity with the least volume (or number) of synaptic contacts that still manages to maintain persistent network stimulus representations. We term this network the sparse FEVER network.

Stable FEVER networks must contain both excitatory and inhibitory synapses (see the Supplemental Experimental Procedures for proof). Yet, according to Dale's law, each neuron can make either excitatory or inhibitory synapses but not both. We thus consider a network comprised of excitatory and inhibitory neurons. Since interneurons, as their name suggests, are local circuit elements, we assume that the stimulus representation is carried out by excitatory neurons only. Therefore, we generated such a FEVER network containing excitatory and inhibitory synapses (see the Supplemental Experimental Procedures).

We start by analyzing the dynamics of this FEVER network as a model for persistent activity in prefrontal working memory networks. Numerical simulations confirm that the stimulus represented by the network is indeed accurately maintained despite the fact that the activity of individual neurons is time varying (Figure 3). Therefore, consistent with the analytical calculations above, we demonstrate that FEVER networks can indeed maintain stable representations despite individual neurons activity being variable. We stress that the variability in neural activity is not due to an explicit representation of time varying parameters, but rather a property of the network's dynamics.

Neurons in the FEVER network exhibit similar responses to those found in cortical parametric working memory networks [14, 18] (Figure 3D). Namely, the trial-averaged responses are parametric functions of stimulus intensity and they exhibit diverse temporal responses, such as ramping up (Figure 3D, top) or down (Figure 3D, middle) of activity over time, as well as the traditional time-invariant (constant) responses (Figure 3D, bottom). More formally, any activity occurring in the null space of the representation matrix will not affect the network stimulus representation yet will influence neuronal activity. Thus, neurons may exhibit a wide range of complex dynamics.

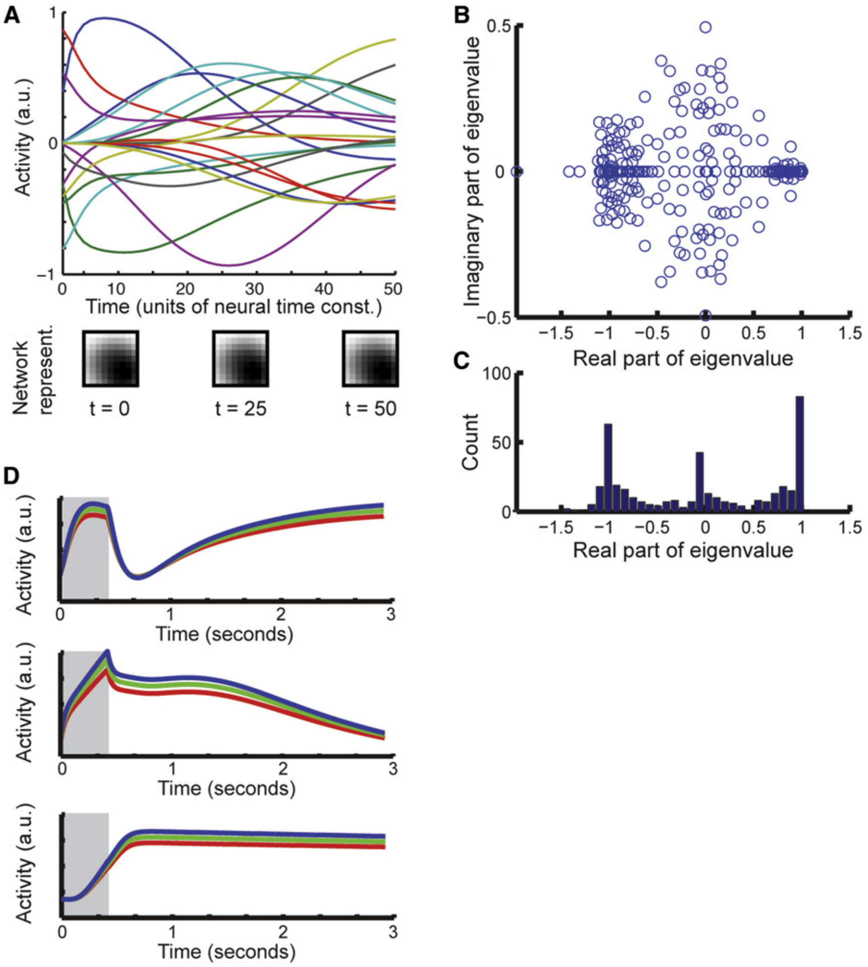

Figure 3. Network Activity

(A) Activity of a subset of neurons out of the full firing rate network in response to the application of a stimulus at time zero as a function of time. Different neurons are shown in different colors. Grayscale boxes (bottom) show stimulus represented by the network at time equal to horizontal location.

(B) Scatter plot of eigenvalues of connectivity matrix in the complex plane. Note the different scale in the x and y axes.

(C) Histogram of real part of eigenvalue. Note the large number of eigenvalues at exactly one. The number of these eigenvalues corresponds to the dimensionality of the stimulus space (81).

(D) Trial-averaged responses of FEVER network to parametric working memory stimuli. Averaged activity of neurons exhibiting different forms of parametric working memory is shown. For each of the neurons, activity is shown for a weak (red), medium strength (green), and strong (blue) stimulus. The activity of the neurons varies as a function of the strength of the stimulus, just as in the cortical data. Top: Ramping up activity over time as in [14]. Middle: Ramping down activity over time. Bottom: Traditional time-invariant delayed activity.

See also Figure S4.

Up until now we described the FEVER network in the context of a delayed match to sample task where a stimulus is presented for a brief time and then removed. In the context of ongoing stimulus presentation the network acts as a high-dimensional integrator (see the Supplemental Experimental Procedures).

How realistic is the FEVER architecture? One strategy to test this hypothesis, albeit a very technically challenging one, is to densely reconstruct prefrontal cortical networks, a "connectomics" approach [33–35]. Then, the eigenvalues of the experimentally determined network can be compared with the highly nontypical eigenspectrum of the predicted FEVER network (Figure 3B,C specifically, the large number of unitary eigenvalues and remaining eigenvalues spread across the spectrum). Another strategy is to look for statistics indicative of a FEVER architecture in measurements available for sensory cortical areas, which we discuss next.

FEVER Networks as a Model of Partial Persistence in Cortical Networks

Our original question, whether changes in neural activity necessarily indicate changes in representation, is relevant not only to prefrontal cortex networks. Persistent representations for time periods beyond the single-neuron time constant, though typically most strongly associated with working memory networks, are also observed in sensory cortical areas following removal of stimuli [36, 37] and even in primary visual cortex, V1 [38, 39]. Intriguingly, a hierarchy of representation persistency has been suggested [36], with primary sensory areas showing brief persistency, secondary sensory areas longer persistency and prefrontal cortices very long persistency.

FEVER networks offer an elegant explanation of the cortical hierarchy of partial persistence since they provide a unified mechanism for extending the persistence of representation beyond the single neuron time constant for different lengths of time. This can be accomplished by a similar FEVER principle, introducing a scaling constant, α ($0 < \alpha < 1$):

$$\alpha \vec{d_i} = \sum_j \vec{d_j} L_{ji}. \qquad \text{(Equation 6)}$$

In response to a transient input, network activity and the associated network stimulus representation do decay, but over a longer time constant than that offered by a single neuron. By varying the scaling factor the effective time constant of network stimulus representation persistence can be varied between the neural time constant and infinity (see the Supplemental Experimental Procedures). In response to a continuous input, such network acts as a leaky integrator. Mathematically, the dynamics must still reside in the null space of D, but now the integrating modes have less than unit eigenvalue, hence the leaky integration [19].

Next, we examine how biologically realistic the structure of FEVER networks is by constructing a FEVER model of early sensory cortex (V1) where much detailed structural information is known and partial persistence has been experimentally verified [38–40]. The FEVER principle is a specific relation between a neuron's coding, or functional, properties and the network architecture. Since the full, high-order, relation is

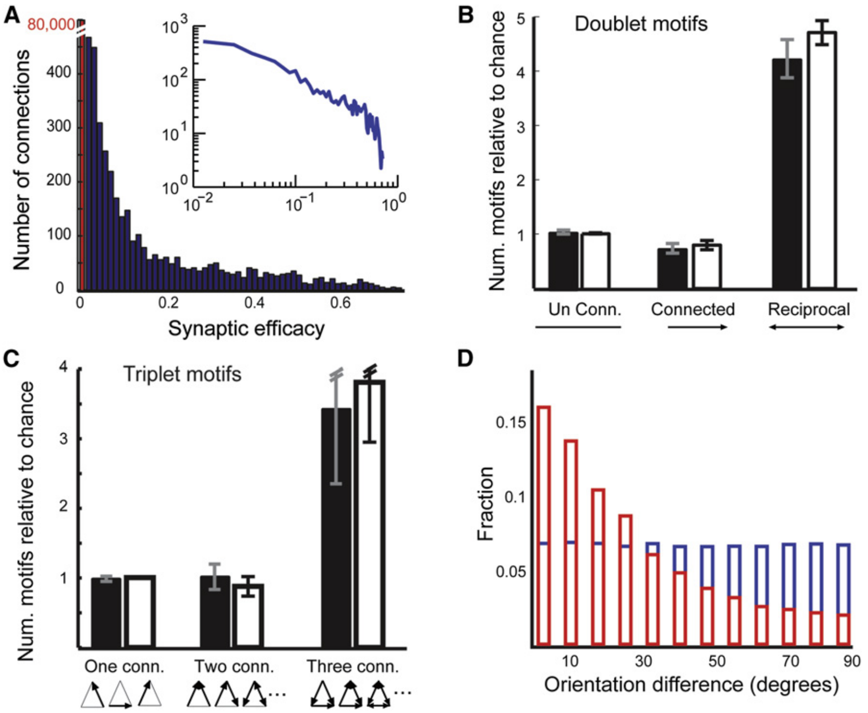

Figure 4. Structure of the Sparse FEVER Network

(A) Histogram of excitatory synaptic weights in blue. Only?5% of possible connections are nonzero valued. The red bar indicates the number of zero-valued connections and is on a different scale marked by the red number. The inset shows log-log. For a plot of the connectivity matrix, see Figure S1.

(B) Number of doublet motifs relative to random degree-matched network for cortical (black bars, gray error bars) and FEVER (white bars, black error bars) networks. Error bars indicate the standard deviation.

(C) Number of triplet motifs pooled according to number of neurons out of the triplet with nonzero connections (for a full plot without pooling, see Figure S3). Error bars indicate the standard deviation. Note that the top error bars are cut to save space.

(D) Distribution of synaptic connections for pairs of neurons according to the difference in their orientation in the learned feature vector FEVER network. Red bars indicate the actual connections, while blue bars indicate potential connections.

See also Figure S3.

difficult to test, we consider whether lower order connectivity predictions of the FEVER can be experimentally tested. Surprisingly, several such predictions, such as the distribution of synaptic weights, the frequency of two and three neuron motifs, and the dependence of connection probability on feature vector properties, agree with the experimentally determined statistical properties of cortical networks.

We consider two sets of feature vectors. First, we consider a more generic option, assuming a difference of Gaussians, excitatory center—inhibitory surround connectivity (Mexican hat, Figure 1). Second, we consider a set of feature vectors trained to efficiently encode patches of natural images [26]. We find that the properties of the sparse FEVER networks constructed from each feature vector model are similar indicating their robustness to the specific choice of the model. In addition, we verified that these properties are robust to reasonable variations of tunable parameters, such as sparseness of lateral connections (within 20%).

The distribution of synaptic weights in the sparse graph, Figure 4A, shows a strong bias to zero valued connections and a heavier than Gaussian tail as does the cortical data [40]. We note that this distribution of synaptic weights is related to the sparseness constraint (l_1 norm, see the Experimental Procedures) on the lateral synaptic connectivity, and is thus not entirely unexpected.

As in cortical networks, the frequency with which a number of motifs occur differs greatly from that expected in a degree-matched random network [41, 42]. Reciprocal two-neuron connections (A to B and B to A) are overrepresented in both the sparse FEVER network and cortical networks. They occur in the sparse FEVER network nearly five times as often as would be expected by chance (Figure 4B, see the Experimental Procedures), compared with four times in the cortical network data [29, 41, 43, 44].

Intuitively, nonrandom connections arise in sparse FEVER networks because each neuron must connect with the fewest number of other neurons that collectively represent its feature vector. For example, if for a given neuron A there exists a neuron B whose feature vector is strongly correlated with A, then a strong connection from A to B would satisfy both the FEVER rule and be favorable in terms of the sparsity constraint. Similarly, a strong connection from B to A would satisfy the FEVER rule for B and sparseness. Therefore, the reciprocal motif is likely to arise in the presence of strongly correlated feature vectors.

The frequency of occurrence of three-neuron motifs also closely resembles that of the cortical network, with highly connected motifs being significantly over abundant relative to chance in both the cortical [41] and FEVER networks (Figures 4C and S3). The same logic of the argument for the overabundance of doublet motifs in FEVER networks presented above applies also to the triplet motifs.

To further understand why motif structures arise in FEVER networks, we explored the relationship between pairwise connection probability and feature vector similarity. We found that connections in the sparse FEVER network are preferentially made between neurons with correlated feature vectors. Preferential connections for neurons with similar feature vectors have been shown in V1, where neurons with similar orientation preference are more likely to be interconnected [45]. Similarly, in the FEVER network with learned, oriented feature vectors, one finds preferential connections for neurons with similar orientation tuning (Figure 4D).

Turning to the dynamics of these networks, numerical simulation verifies that the activity decays according to the network time constant set by Equation 6 and that within this time frame the activity of individual neurons is time varying (Figure 5A). The activity of FEVER neurons is sparse. The distribution of spontaneous firing rates in the sparse FEVER network has a heavy tail (Figure 5B), similar to that observed in cortical neurons [46]. Previously, a broad distribution of firing rates among neurons was explained by postulating that some neurons in

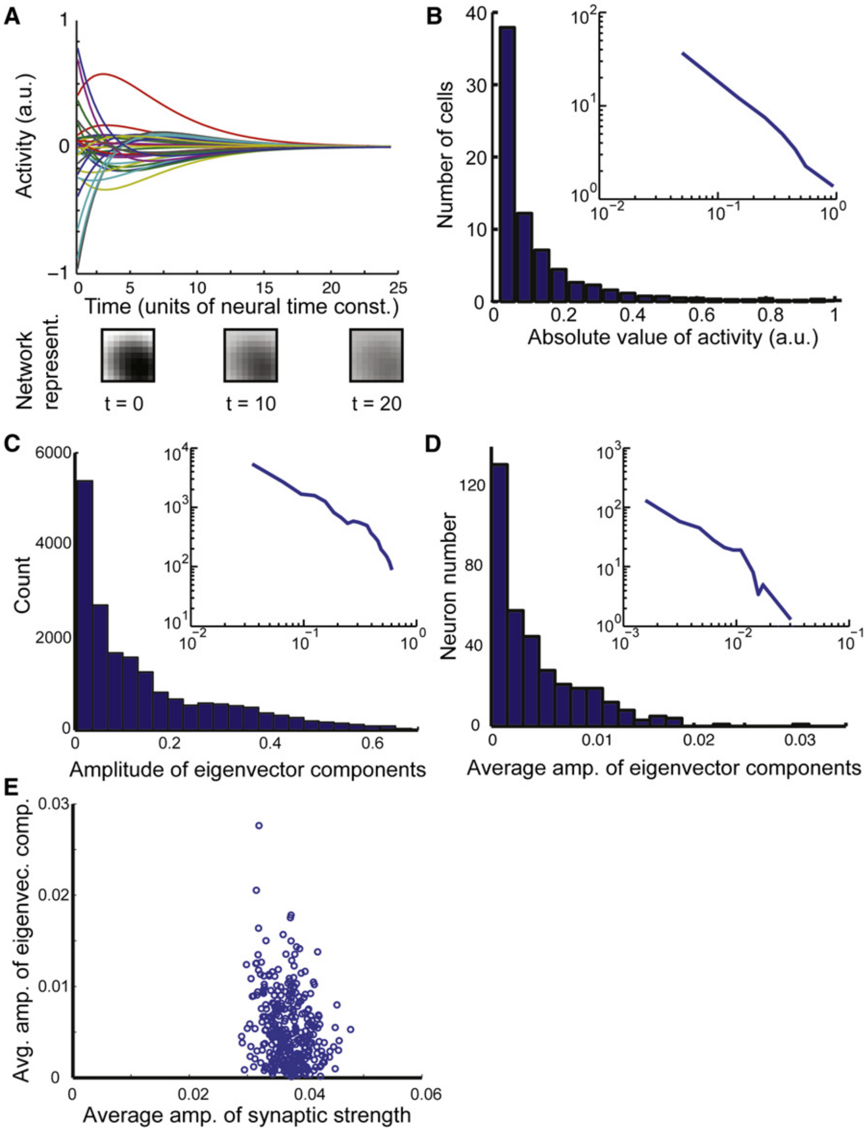

Figure 5. Dynamics of Sensory Cortex FEVER Network

(A) Activity of a subset of neurons out of the full firing rate network in response to the application of a stimulus at time zero as a function of time. Different neurons are shown in different colors. Grayscale boxes (bottom) show stimulus represented by the network at time equal to horizontal location.

(B) Histogram of number of neurons as a function of the strength of their activity. The inset shows log-log.

(C) Histogram of amplitude of eigenvector components with maximal eigenvalue. The inset shows log-log.

(D) Histogram of mean eigenvalue component of maximal eigenvalue eigenvectors, averaged for each neuron. The inset shows log-log.

(E) Scatter of mean presynaptic weight amplitude, averaged for each neuron (x axis) and mean eigenvalue component (data from D).

See also Figure S5.

the network have consistently stronger incoming synapses [47]. In contrast, in sparse FEVER networks the average synaptic strength per neuron is similar among neurons, yet a broad distribution of spontaneous firing rates still occurs.

The broad distribution of firing rates can be traced to the network eigenspectrum arising from the nonrandom structure of the FEVER network. Since activity corresponding to modes less than the maximal eigenvalue will decay, the distribution of spontaneous firing rates will be determined by the degenerate eigenvectors with a maximal eigenvalue, regardless of the original profile of activity. We find that the components of these eigenvectors have a heavy tailed distribution (Figure 5C). Moreover, the components corresponding to the same neuron averaged over maximal eigenvalue eigenvectors also have a heavy tailed distribution (Figure 5D). Assuming that spontaneous activity is described by a linear combination of these eigenvectors with random coefficients, one expects a broad distribution of spontaneous firing rates. Similar to networks considered in [47], the broad distribution of firing rates arises due to the difference in mean activity among neurons. However, unlike [47], this difference arises despite a Gaussian-like distribution of total presynaptic weights among neurons with no obvious relation between the variation in the presynaptic and the distribution of spontaneous firing rates ($R^2 = 0.05$, Figure 5E).

Applicability of the FEVER Rule to Nonlinear Dynamics

Although the derivation of the FEVER rule (Equation 5) relied on a linear rate model (Equation 3), the FEVER rule is also applicable for certain networks of spiking neurons. Consider for instance the important case of a network of ideal integrate-and-fire neurons. Denoting the subthreshold membrane potential of the i^{th} neuron by a_i, its dynamics are described by two terms. First, a post-spiking reset, $-f(a_i)$, where f is a Heavyside or threshold function, and second a weighted summation of spikes from presynaptic neurons $\sum_j L_{ij} f(a_j)$. The dynamics are therefore of the following form:

$$\frac{da_i}{dt} = -f(a_i) + \sum_j L_{ij} f(a_j). \qquad \text{(Equation 7)}$$

Given that the dynamics follow Equation 7, we can plug in the FEVER rule and calculate the change of a network stimulus representation with the network dynamics:

$$\begin{aligned}
\frac{d\vec{s}}{dt} &= \frac{d}{dt}\left(\sum_i \vec{d}_i a_i\right) = \sum_i \vec{d}_i \frac{d}{dt} a_i \\
&= \sum_i \vec{d}_i \left(-f(a_i) + \sum_j L_{ij} f(a_j)\right) \\
&= -\sum_i \vec{d}_i f(a_i) + \sum_i \vec{d}_i \sum_j L_{ij} f(a_j) \qquad \text{(Equation 8)} \\
&= -\sum_i \vec{d}_i f(a_i) + \sum_j f(a_j) \sum_i L_{ij} \vec{d}_i \\
&= -\sum_i \vec{d}_i f(a_i) + \sum_j \vec{d}_j f(a_j) = 0.
\end{aligned}$$

Thus, if we assume that the network stimulus representation is given by the subthreshold potential times the feature vector, the representation will remain persistent despite spiking activity and varying membrane potential (Figure S4). Although the subthreshold value of each neuron is not available to other neurons, it is composed of postsynaptic currents generated by spikes from other neurons. Therefore, it is indirectly available to read-out. We note that in these cases, the FEVER rule remains unmodified, and therefore all of the results relating to network structure will also hold for these nonlinear networks. The same calculation would hold true when the threshold function *f* is replaced by any other nonlinear function, extending the relevance to a broader class of spiking networks.

Establishment of FEVER Network through Hebbian Learning

Although generation of a network satisfying the FEVER rule may seem complicated, if feature vectors form a "tight frame" (see the Supplemental Experimental Procedures), such network architecture can be learned with Hebbian learning rules (Figure S5), which are the basis of other related network learning models [48, 49]. For two sensory dimensions, an example of such feature vectors are the set of periodic rotations of one feature vector.

To understand intuitively why a Hebbian learning rule will result in FEVER networks for such feature vectors, consider the FEVER principle—the sum of postsynaptic feature vectors weighted by the synaptic strength should be equal to the neurons own feature vector. For instance, let us examine the feature vector pointing up the vertical direction and its relation to the two feature vectors immediately adjacent to it in the clockwise and counterclockwise direction. Assuming that the sensory input is whitened [50, 51], the correlation of activity is equal to the correlation of feature vectors. The correlation between the vertical feature vector and the two adjacent feature vectors is identical in magnitude. Thus, a Hebbian rule will strengthen synapses equally between the vertical feature vectors and these two neurons. A sum across these two feature vectors with equal weights will cancel out the horizontal contribution and leave only a vertical contribution, which is equal to a scaled version of the original feature vectors; therefore the FEVER principle will be met. The same argument applies to the rest of the neurons' feature vectors in this example (the full mathematical details of the derivation can be found in the Supplemental Experimental Procedures section).

Finally, we point out that though the learning rule converges to the correct solution, deviations of the connectivity from these values will deteriorate the performance of the network, like in many other models of persistent activity, resulting in shorter integration times, an issue known as "synaptic fine-tuning" [19].

Discussion

In this paper, we show that not every change in neural activity necessarily indicates a change in the network stimulus representation. Furthermore, we derive the FEVER network architecture that ensures persistency (full or partial) of network stimulus representations of multi-dimensional stimuli, despite time-varying activity. In the case of transient input, the network stimulus representation can be stored indefinitely despite time-variable neuronal activity; in the case of continuous input, its (weighted) integral over time is represented by the network.

How biologically realistic is this explanation for the apparent contradiction between time-varying activity and stable network stimulus representations? We began by considering FEVER networks in the context of working memory in prefrontal cortex, where the contradiction was originally recognized and showed that FEVER networks are capable of maintaining stable representations indefinitely despite having time-varying activity like that seen in prefrontal network delay activity. Next, we considered finite-time-constant FEVER networks as a model for sustained network stimulus representation after removal of stimulus in sensory cortices [38–40] and found that many features of network architecture, such as the distribution of synaptic weights and few-neuron motifs, qualitatively match known experimental distributions. Considering both scenarios, we believe FEVER networks offer a simple, biologically plausible conceptual framework for understanding the stability of network stimulus representations despite time varying activity in individual neurons.

There are alternative explanations for the apparent contradiction between stable network stimulus representation and complex, time-varying activity in individual neurons. First, one cannot directly refute the possibility that in cortical networks, relevant information is encoded only in the small fraction of neurons that do maintain constant activity, making the time-variant nature of these networks irrelevant. Second, the complex activity could be directly tied to a complex, explicit representation of different factors beyond the stimulus, such as the amount of elapsed time, as has been previously suggested [14, 18, 52, 53]. Since the combined stimulus (original stimulus and time) is time varying, the activity that represents it will be time varying, as well. Our study does not refute the idea of explicit coding of time in working memory networks but rather shows that time-varying activity does not necessarily imply that the underlying network stimulus representation explicitly encodes time-varying properties.

Out of a variety of different network mechanisms suggested to account for working memory [19, 27, 49, 52–60], FEVER networks are most closely related to the linear outer product line attractor network [19] conceived as a model for representing eye position in the goldfish occulomotor system [61, 62]. Both networks operate by creating neutral stability (corresponding to eigenvalue one) along coding dimensions. In fact, the dynamics of FEVER networks can be seen as that of a "subspace attractor"—a generalization of the line attractor to multidimensional stimuli.

The central prediction of FEVER networks regarding network dynamics is that there are two distinct modes of network dynamics: coding and noncoding modes. The prediction is that activity during delay periods in working memory tasks will be limited to the noncoding modes, i.e., specific directions in activity space (corresponding to the null space of the representation matrix D), and will not have a significant component in coding modes, other directions in activity space. By correlating activity with different stimulus values, the coding directions in activity space may be revealed. In practice, multiunit recording samples population activity only partially, thus resulting in data sets in which these distinctions are blurred. Simulating a simplified scenario for this process (see the Supplemental Experimental Procedures), we test the fraction of neurons needed to observe this phenomenon (Figure S5).

In summary, in this paper we explicitly showed that not every change in neural activity will necessarily result in a change in the network stimulus representation. This realization could be relevant beyond working memory circuits and might be useful for explaining complex activity in other redundant neural circuits, such as motor cortex [63] and even the basal ganglia [64].

Experimental Procedures

Simulating Neural Activity

A rate model was simulated according to Equation 3. Numerical simulation was performed by a Runge-Kutta integration algorithm implemented in MATLAB. Results were compared to the analytical solution through eigenvector decomposition to ensure accuracy. Membrane time constant was set at 10 ms, and simulation was performed with a time step (dt) of 0.1 ms. Simulations were typically run for 5 s of simulated time.

Construction of Sparse FEVER Networks

The FEVER rule is expressed by Equation 5, where the vectors *d* are feature vectors and *L* is the matrix of lateral connectivity. To construct a sparse FEVER network directly, one can minimize the square deviation from Equation 5 along with a sparseness-inducing term:

$$C = \sum_i \left(\vec{d}_i - \sum_j \vec{d}_j L_{ji} \right)^2 + \lambda \sum_i \| \vec{L}_i \|_1 . \qquad \text{(Equation 9)}$$

This sparse approximation problem can be solved via standard numerical recipes, such as the SPAMS toolbox [65]. A more technical account of constructing such FEVER networks has appeared as conference proceedings [66]. However, in the networks constructed with Equation 9, each neuron makes both excitatory and inhibitory synapses, thus violating Dale's law. The details of generating networks obeying Dale's law are found in the Supplemental Experimental Procedures.

Motif Analysis

Motif analysis was performed as in [41]. Briefly stated, 1,000 degree-matched random networks were generated, and the occurrence of pair and triplet motifs was compared between the FEVER matrix and the control networks. We note that comparison of degree-matched networks takes into account contributions from sparseness, since two degree-matched networks are equally sparse. Since the fully connected triplet motifs showed the most interesting results but contained only a small number of samples, we pooled the motifs across one connected triplet motifs, two connected triplet motifs, and three connected triplet motifs. The ungrouped motifs can be seen in Figure S3.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures and five figures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2012.08.058.

Acknowledgments

We are grateful to M. Goldman, H. Greenside, A. Grinshpan, A. Koulakov, A. Lee, K. Svoboda, J. Magee, and anonymous reviewers for helpful comments on the manuscript.

Received: February 16, 2012
Revised: June 28, 2012
Accepted: August 31, 2012
Published online: October 18, 2012

References

1. Baddeley, A.D. (1986). Working Memory (Oxford: Oxford University Press).
2. Hogan, N., and Flash, T. (1987). Moving Gracefully - Quantitative Theories of Motor Coordination. Trends Neurosci. *10*, 170–174.
3. Koch, C. (1999). Biophysics of Computation: Information Processing in Single Neurons (New York: Oxford University Press).
4. Tchumatchenko, T., Malyshev, A., Wolf, F., and Volgushev, M. (2011). Ultrafast population encoding by cortical neurons. J. Neurosci. *31*, 12171–12179.
5. Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. Science *304*, 1926–1929.
6. Marom, S. (2010). Neural timescales or lack thereof. Prog. Neurobiol. *90*, 16–28.
7. Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. *117*, 500–544.
8. Haider, B., and McCormick, D.A. (2009). Rapid neocortical dynamics: cellular and network mechanisms. Neuron *62*, 171–189.
9. Churchland, P.S., and Sejnowski, T.J. (1992). The Computational Brain (Cambridge, Massachusetts: MIT Press).
10. Fuster, J.M. (1973). Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. J. Neurophysiol. *36*, 61–78.
11. Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654.
12. Kojima, S., and Goldman-Rakic, P.S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. *248*, 43–49.
13. Funahashi, S., Chafee, M.V., and Goldman-Rakic, P.S. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. Nature *365*, 753–756.
14. Brody, C.D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb. Cortex *13*, 1196–1207.
15. Romo, R., Brody, C.D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. Nature *399*, 470–473.
16. Batuev, A.S., Pirogov, A.A., and Orlov, A.A. (1979). Unit activity of the prefrontal cortex during delayed alternation performance in monkey. Acta Physiol. Acad. Sci. Hung. *53*, 345–353.
17. Baeg, E.H., Kim, Y.B., Huh, K., Mook-Jung, I., Kim, H.T., and Jung, M.W. (2003). Dynamics of population code for working memory in the prefrontal cortex. Neuron *40*, 177–188.
18. Machens, C.K., Romo, R., and Brody, C.D. (2010). Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. J. Neurosci. *30*, 350–360.
19. Seung, H.S. (1996). How the brain keeps the eyes still. Proc. Natl. Acad. Sci. USA *93*, 13339–13344.
20. Bialek, W., Rieke, F., de Ruyter van Steveninck, R.R., and Warland, D. (1991). Reading a neural code. Science *252*, 1854–1857.
21. Salinas, E., and Abbott, L.F. (1994). Vector reconstruction from firing rates. J. Comput. Neurosci. *1*, 89–107.
22. Victor, J.D. (2005). Analyzing receptive fields, classification images and functional images: challenges with opportunities for synergy. Nat. Neurosci. *8*, 1651–1656.
23. Major, G., and Tank, D. (2004). Persistent neural activity: prevalence and mechanisms. Curr. Opin. Neurobiol. *14*, 675–684.
24. Abeles, M. (1991). Corticonics: Neural Circuits of the Cerebral Cortex (Cambridge: Cambridge University Press).
25. Braitenberg, V., and Schüz, A. (1998). Cortex: Statistics and Geometry of Neuronal Connectivity (Berlin, Germany: Springer).
26. Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609.
27. Goldman, M.S. (2009). Memory without feedback in a neural network. Neuron *61*, 621–634.
28. Murphy, B.K., and Miller, K.D. (2009). Balanced amplification: a new mechanism of selective amplification of neural activity patterns. Neuron *61*, 635–648.
29. Markram, H., Lübke, J., Frotscher, M., Roth, A., and Sakmann, B. (1997). Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. J. Physiol. *500*, 409–440.
30. Laughlin, S.B., and Sejnowski, T.J. (2003). Communication in neuronal networks. Science *301*, 1870–1874.
31. Varshney, L.R., Sjöström, P.J., and Chklovskii, D.B. (2006). Optimal information storage in noisy synapses under resource constraints. Neuron *52*, 409–423.
32. Ramon y Cajal, S. (1899). Textura del Sistema Nervioso del Hombre y de los Vertebrados (Madrid, Spain: Nicolas Moya).
33. Briggman, K.L., Helmstaedter, M., and Denk, W. (2011). Wiring specificity in the direction-selectivity circuit of the retina. Nature *471*, 183–188.

34. Bock, D.D., Lee, W.C., Kerlin, A.M., Andermann, M.L., Hood, G., Wetzel, A.W., Yurgenson, S., Soucy, E.R., Kim, H.S., and Reid, R.C. (2011). Network anatomy and in vivo physiology of visual cortical neurons. Nature *471*, 177–182.

35. Chklovskii, D.B., Vitaladevuni, S., and Scheffer, L.K. (2010). Semi-automated reconstruction of neural circuits using electron microscopy. Curr. Opin. Neurobiol. *20*, 667–675.

36. Hernández, A., Nácher, V., Luna, R., Zainos, A., Lemus, L., Alvarez, M., Vázquez, Y., Camarillo, L., and Romo, R. (2010). Decoding a perceptual decision process across cortex. Neuron *66*, 300–314.

37. Pasternak, T., and Greenlee, M.W. (2005). Working memory in primate sensory systems. Nat. Rev. Neurosci. *6*, 97–107.

38. Benucci, A., Ringach, D.L., and Carandini, M. (2009). Coding of stimulus sequences by population responses in visual cortex. Nat. Neurosci. *12*, 1317–1324.

39. Duysens, J., Orban, G.A., Cremieux, J., and Maes, H. (1985). Visual cortical correlates of visible persistence. Vision Res. *25*, 171–178.

40. Nikolić, D., Häusler, S., Singer, W., and Maass, W. (2009). Distributed fading memory for stimulus properties in the primary visual cortex. PLoS Biol. *7*, e1000260.

41. Song, S., Sjöström, P.J., Reigl, M., Nelson, S., and Chklovskii, D.B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol. *3*, e68.

42. Newman, M.E.J. (2010). Networks: An Introduction (Oxford: Oxford University Press).

43. Perin, R., Berger, T.K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. Proc. Natl. Acad. Sci. USA *108*, 5419–5424.

44. Brown, S.P., and Hestrin, S. (2009). Intracortical circuits of pyramidal neurons reflect their long-range axonal targets. Nature *457*, 1133–1136.

45. Ko, H., Hofer, S.B., Pichler, B., Buchanan, K.A., Sjöström, P.J., and Mrsic-Flogel, T.D. (2011). Functional specificity of local synaptic connections in neocortical networks. Nature *473*, 87–91.

46. Hromádka, T., Deweese, M.R., and Zador, A.M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol. *6*, e16.

47. Koulakov, A.A., Hromádka, T., and Zador, A.M. (2009). Correlated connectivity and the distribution of firing rates in the neocortex. J. Neurosci. *29*, 3685–3694.

48. Arnold, D.B., and Robinson, D.A. (1992). A neural network model of the vestibulo-ocular reflex using a local synaptic learning rule. Philos. Trans. R. Soc. Lond. B Biol. Sci. *337*, 327–330.

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

50. Dan, Y., Atick, J.J., and Reid, R.C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J. Neurosci. *16*, 3351–3362.

51. Atick, J.J., and Redlich, A.N. (1990). Towards a theory of early visual processing. Neural Comput. *2*, 308–320.

52. Barak, O., Tsodyks, M., and Romo, R. (2010). Neuronal population coding of parametric working memory. J. Neurosci. *30*, 9424–9430.

53. Singh, R., and Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. J. Neurosci. *26*, 3667–3678.

54. Compte, A., Brunel, N., Goldman-Rakic, P.S., and Wang, X.J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex *10*, 910–923.

55. Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

56. Cannon, S.C., Robinson, D.A., and Shamma, S. (1983). A proposed neural network for the integrator of the oculomotor system. Biol. Cybern. *49*, 127–136.

57. Boerlin, M., and Denève, S. (2011). Spike-based population coding and working memory. PLoS Comput. Biol. *7*, e1001080.

58. Koulakov, A.A., Raghavachari, S., Kepecs, A., and Lisman, J.E. (2002). Model for a robust neural integrator. Nat. Neurosci. *5*, 775–782.

59. Amit, D.J., Brunel, N., and Tsodyks, M.V. (1994). Correlations of cortical Hebbian reverberations: theory versus experiment. J. Neurosci. *14*, 6435–6445.

60. Wang, X.J. (2001). Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. *24*, 455–463.

61. Aksay, E., Baker, R., Seung, H.S., and Tank, D.W. (2000). Anatomy and discharge properties of pre-motor neurons in the goldfish medulla that have eye-position signals during fixations. J. Neurophysiol. *84*, 1035–1049.

62. Robinson, D.A. (1970). Oculomotor unit behavior in the monkey. J. Neurophysiol. *33*, 393–403.

63. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Ryu, S.I., and Shenoy, K.V. (2010). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? Neuron *68*, 387–400.

64. Morris, G., Nevet, A., and Bergman, H. (2003). Anatomical funneling, sparse connectivity and redundancy reduction in the neural networks of the basal ganglia. J. Physiol. Paris *97*, 581–589.

65. Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. *11*, 19–60.

66. Druckmann, S., and Chklovskii, D.B. (2010). Over-complete representations on recurrent neural networks can support persistent percepts. In Advances in Neural Information Processing Systems 23, J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, eds. (Red Hook, NY: Curran Associates, Inc.), pp. 541–549.