Published in final edited form as:
Neuron. 2018 July 11; 99(1): 227–238.e4. doi:10.1016/j.neuron.2018.05.038.

Attractor dynamics in networks with learning rules inferred from *in vivo* data

Ulises Pereira[a] and **Nicolas Brunel**[a,b,c,d,1,2]

[a]Department of Statistics, The University of Chicago, Chicago, Illinois 60637, USA

[b]Department of Neurobiology, The University of Chicago, Chicago, Illinois 60637, USA

[c]Department of Neurobiology, Duke University, Durham, North Carolina 27710, USA

[d]Department of Physics, Duke University, Durham, North Carolina 27708, USA

Abstract

The attractor neural network scenario is a popular scenario for memory storage in association cortex, but there is still a large gap between models based on this scenario and experimental data. We study a recurrent network model in which both learning rules and distribution of stored patterns are inferred from distributions of visual responses for novel and familiar images in inferior temporal cortex (ITC). Unlike classical attractor neural network models, our model exhibits graded activity in retrieval states, with distributions of firing rates that are close to lognormal. Inferred learning rules are close to maximizing the number of stored patterns within a family of unsupervised Hebbian learning rules, suggesting learning rules in ITC are optimized to store a large number of attractor states. Finally, we show that there exists two types of retrieval states: one in which firing rates are constant in time, another in which firing rates fluctuate chaotically.

Introduction

Attractor networks have been proposed as models of learning and memory in the cerebral cortex (Hopfield, 1982; Amit, 1992, 1995; Brunel, 2005). In these models, synaptic connectivity in a recurrent neural network is set up in such a way that the network dynamics have multiple attractor states, each of which represents a particular item that is stored in memory. Each attractor state is a specific pattern of activity of the network, that is correlated with the state of the network when the particular item is presented through external inputs.

[1]Correspondence: nicolas.brunel@duke.edu.
[2]Lead Contact

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Contributions

U.P. and N.B. designed the research. U.P. and N.B. developed the mathematical theory. U.P. performed the analytical calculations and numerical simulations. U.P. analyzed the data. U.P. and N.B. wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

The attractor property means that the network converges to the stored pattern, even if the external inputs are correlated to, but not identical, to the pattern, a necessary requirement for an associative memory model. In many of these models, the appropriate synaptic connectivity is assumed to be generated thanks to a ‘Hebbian’ learning process, according to which synaptic efficacies are modified by the activity of pre and post-synaptic neurons (Hebb, 1949).

These models have been successful in reproducing qualitatively several landmark observations in delayed response tasks experiments in monkeys (Fuster et al., 1971; Miyashita, 1988; Funahashi et al., 1989; Goldman-Rakic, 1995) and rodents (Liu et al., 2014; Guo et al., 2014; Inagaki et al., 2017). In some of the monkey experiments, animals are trained to perform a task in which they have to remember for short times the identity or the location of a visual stimulus. These tasks share in common a presentation period during which the monkey is subjected to an external stimulus, and a delay period during which the monkey has to maintain in working memory the identity of the stimulus, which is needed to solve the task after the end of the delay period. One of the major findings of these experiments is the observation of selective persistent activity during the delay period in a subset of recorded neurons in many cortical areas, in particular in prefrontal cortex (Fuster et al., 1971; Funahashi et al., 1989; Romo et al., 1999), parietal cortex (Koch and Fuster, 1989), inferior temporal cortex (Fuster and Jervey, 1981; Miyashita, 1988; Nakamura and Kubota, 1995) and other areas of the temporal lobe (Nakamura and Kubota, 1995). In those neurons, the firing rate does not decay to baseline during the delay period, but it is rather maintained at higher than baseline levels. Furthermore, this increase in firing rate is selective, i.e. it occurs only for a subset of stimuli used in the experiment. Selective persistent activity is consistent with attractor dynamics in a recurrent neural network, whose synaptic connectivity is shaped by experience dependent synaptic plasticity (Amit, 1995; Wang, 2001; Brunel, 2005).

The attractor network scenario was originally instantiated in highly simplified fully connected networks of binary neurons (Amari, 1972; Hopfield, 1982). While theorists have since strived to incorporate more neurophysiological realism into associative memory models, using e.g. asymmetric and sparse connectivity (Derrida et al., 1987), sparse coding of memories (Tsodyks and Feigel’Man, 1988; Tsodyks, 1988), online learning (Mézard et al., 1986; Parisi, 1986; Amit and Fusi, 1994), spiking neurons (Gerstner and van Hemmen, 1992; Treves, 1993; Amit and Brunel, 1997; Brunel and Wang, 2001; Lansner, 2009), there is still a large gap between these models and experimental data. First, none of the existing models use patterns whose statistics is consistent with data. Most models use bimodal distributions of firing rates, with neurons either ‘activated’ by a stimulus or not, while there is no indication of such a bimodality in the data. Second, the connectivity matrices used in these models are essentially engineered (and sometimes highly fine-tuned) such as to produce attractor dynamics, but are totally unconstrained by data. Third, the attractor network scenario has been challenged by the observation of a high degree of irregularity and strong temporal variations in the firing rates of many neurons, which seem hard to reconcile with fixed point attractors (Druckmann and Chklovskii, 2012; Barak et al., 2013; Murray et al., 2017).

A recent study (Lim et al., 2015) provides us with the tools to potentially bridge these gaps. It used data from experiments in which neuronal activity is recorded in IT cortex in response to large sets of novel and familiar stimuli (Woloszyn and Sheinberg, 2012). The distribution of neuronal responses to novel stimuli allows the inference of the distribution of firing rates of neurons in stimuli that are being memorized. This distribution is close to a lognormal, at odds with bimodal distributions of firing rates used in the vast majority of theoretical studies (for a few exceptions, see Treves (1990*a*,*b*); Festa et al. (2014)). Comparison between the distributions of responses to novel and familiar stimuli allows the inference of the dependence of the learning rule on post-synaptic firing rates. The inferred learning rule is Hebbian, but shows two major differences with classic rules such as the covariance rule (Sejnowski, 1977): (1) The post-synaptic dependence of the rule is dominated by depression, such that the vast majority of external inputs leads to a net decrease in total synaptic inputs to a neuron with learning, leading to a sparser representation of external stimuli; (2) The dependence of the rule on post-synaptic firing rates is highly non-linear, as in the Bienenstock-Cooper-Munro rule (Bienenstock et al., 1982).

These results beg the question of whether associative memory can emerge in networks whose distributions of firing rates and learning rules are consistent with data. We therefore set out to study a recurrent network model in which distributions of external inputs, single neuron transfer function and learning rule are all inferred from ITC data (Lim et al., 2015). We show that: (1) learning rules inferred from visual responses in ITC lead to attractor dynamics, without any need for parameter adjustment or fine tuning; (2) Activity in the delay period is graded, with broad distributions of firing rates; (3) Learning rules inferred from data are close to maximizing the number of stored patterns, in a space of unsupervised Hebbian learning rules with sigmoidal dependence on pre and post-synaptic firing rates; (4) In a large parameter region, our model presents irregular temporal dynamics during retrieval states that strongly resembles the temporal variability observed during delay periods. In this region, retrieval states are chaotic attractors that maintain a positive overlap with the corresponding stored memory, and the network performs as a associative memory device with fluctuations internally generated by the chaotic dynamics.

Results

We model local cortical circuits in IT cortex by a recurrent network composed of ‘firing rate’ units (Hopfield, 1984). The network is composed of N neurons whose firing rates are described by analog variables r_i, where $i = 1, 2, \ldots, N$ represents the neuron index, as a simplified model for a local network in ITC (see Fig. 1 for a schematic depiction of the network). Firing rates obey standard rate equations (Grossberg, 1969; Hopfield, 1984)

$$\tau \dot{r}_i = -r_i + \phi\left(I_i + \sum_{i \neq j}^{N} J_{ij} r_j\right), \quad (1)$$

where τ is the time constant of firing rate dynamics, ϕ is the input-output single neuron transfer function (or f-I curve), I_i are the external inputs to neuron i, and J_{ij} is the strength of the synapse connecting neuron j to neuron i.

The connectivity matrix is sparse, and existing connections are shaped by external inputs (‘patterns’) through a non-linear unsupervised Hebbian synaptic plasticity rule. In this rule, external synaptic inputs ξ_i^μ to neuron i during presentation of pattern μ ($i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, p$) are generated randomly and independently from a Gaussian distribution (see Fig. 1A,B and Methods). The assumption of independence of the patterns is consistent with the data (see Fig. S1). The external inputs shape the connectivity matrix through the firing rates $\phi(\xi_i^\mu)$ generated by such inputs, and through two non-linear functions f and g that characterize the dependence of the learning rule on the post-synaptic rate (f) and pre-synaptic rate (g), respectively. When p patterns are learned by the network, the final connectivity after learning gets structured as

$$J_{ij} = \frac{Ac_{ij}}{cN} \sum_{k=1}^{p} f[\phi(\xi_i^k)]g[\phi(\xi_j^k)], \qquad (2)$$

where c_{ij} is a sparse random (Erdos-Renyi) structural connectivity matrix ($c_{ij} = 1$ with probability c, $c_{ij} = 0$ with probability $1 - c$, where $c \ll 1$). This synaptic connectivity matrix can be obtained by a learning rule that changes the synaptic connectivity matrix by a factor $\Delta J_{ij} \propto f[\phi(\xi_i^\mu)]g[\phi(\xi_j^\mu)]$ when a pattern μ is presented to the network, starting from an initial *tabula rasa* $J_{ij} = 0$, and neglecting the contributions of recurrent connections during learning. This rule is a generalization of Hebbian rules used in classic models such as the Hopfield model (Hopfield, 1982) or the Tsodyks-Feigel’man model (Tsodyks and Feigel’Man, 1988), with two important differences: patterns have a Gaussian distribution instead of binary; and the dependence of the rule on firing rates is non-linear instead of linear. In the following, the patterns that have shaped the connectivity matrix will be termed ‘familiar’ while all other random patterns presented to the network will be termed ‘novel’.

Inferring transfer function and learning rule from data

The model defined by Eqs. (1,2) depends on three functions ϕ, f and g that define the single neuron transfer function and synaptic learning rule, respectively. How to choose these functions? We used a method that was recently introduced by Lim et al. (2015) to infer the tranfer function (ϕ) and the post-synaptic dependence of the learning rule f from electrophysiological data recorded in ITC (Woloszyn and Sheinberg, 2012). The transfer function ϕ is obtained by finding the function that maps a standard Gaussian distribution to the empirical distribution of visual responses of neurons to a large set of novel stimuli (see Methods). The post-synaptic dependence of the learning rule f was obtained from the differences between the distribution of visual responses to familiar and novel stimuli, under the assumption that changes in such distributions are due to changes in synaptic connectivity in recurrent ITC circuits. Note that only the function f, and not g, can be inferred from data - this is due to the fact that the mean inputs to a neuron are proportional to $[\phi(\xi_i^k)]$ while the function g only appears in an integral (see Methods, Eq. (20)). Therefore, the knowledge of how the mean inputs change with learning as a function of its firing rate allows us to infer f but not g. As an additional step to the procedure described by Lim et al. (2015), we fitted the resulting functions ϕ and f using sigmoidal functions (see Methods and Fig. 2). These sigmoidal functions provided good fits to the data (see Fig.2A–C, that shows fits of three representative ITC neurons; and Fig. S2–4 for all neurons in the data set). This fitting procedure gave us for each neurons three parameters of the transfer function: the maximal firing rate r_m (median: r_m = 76.2Hz), a measure of the slope at the inflection point β_T (median: β_T = 0.82), and the threshold (current at the inflection point, median: h_0 = 2.46 - see Fig. 2D for a boxplot of these parameters). It also gives us for each neuron three parameters characterizing the function f: the threshold x_f (median: 26.6 Hz), slope at the inflection point β_f (median: 0.28 s) and saturation q_f (median: 0.83). Finally, the fitting procedure also gives us the learning rate A (median: 3.55).

A number of features of these fitted functions are noteworthy: First, the vast majority of the visual responses of neurons are in the supralinear part of the transfer function, and therefore far from saturation. This is consistent with many studies showing supra-linear transfer functions at low firing rates, both *in vitro* (Rauch et al., 2003) and *in vivo* (Anderson et al., 2000). Second, this has the consequence that the distribution of visual responses are strongly right-skewed, and in fact close to lognormal distributions, consistent with multiple observations *in vivo* (Hromadka et al., 2008; Roxin et al., 2011; Buzsaki and Mizuseki, 2014; Lim et al., 2015). Third, the function f is strongly non-linear, and the threshold between depression and potentiation occurs at a firing rate that is much higher than the mean rate, leading to depression of the mean synaptic inputs to a neuron for the vast majority of shown stimuli. Fourth, the average of the function f across the distribution of patterns is negative, which leads to a decrease of the average visual response with familiarity (Lim et al., 2015).

The only parameters that are left unconstrained by data are two parameters characterizing the function g. In most of the following, we will take those parameters to be identical to the corresponding parameters of the function f (i.e. $x_g = x_f$ and $\beta_g = \beta_f$; note that q_g is fixed by the condition that the average of the function g across the distribution of patterns is zero, see Methods). We will also explore the space of values of x_g and β_g (see below).

Dynamics of the network following presentation of a familiar stimulus

Having specified the model, we now turn to the dynamics of the network described by Eqs. (1,2), whose parameters are set to the median best-fit parameters according to the procedure described above. In particular, we ask whether the model exhibits attractor dynamics. To address this question, we used both numerical simulations of large networks (see methods) and mean field theory (MFT - see methods and methods S1). For the MFT, we assume that both the number of neurons and stored patterns are large (i.e. more specifically the limit $p, N \to \infty$), while the number of stored patterns p divided by the average number of synapses per neuron (Nc), $\alpha \equiv p/Nc$ remains of order one. We call α the *memory load* of the network. The results of the MFT only depend on N, c and p via this quantity (see methods and methods S1). From our MFT analysis, we obtain mathematical expressions for two ‘order parameters’ that describe how network states are correlated (or not) with stored patterns. We are specifically interested here in the situation when the network state is correlated with one of the stored patterns (e.g. following the presentation of this particular pattern).

The first order parameter describes the ‘overlap’ m between the current state of the network (described by the vector of firing rates r_i, for $i = 1, 2, \ldots, N$) and the pattern of interest (see Methods for the mathematical definition of m). When m is of order 1, this indicates that the corresponding pattern is retrieved from memory. Consequently, each pattern stored in memory can be retrieved by initializing the network dynamics with a configuration that is close to that particular pattern, and letting the network evolve towards its attractor state. In this case, giving a partial cue to the network leads the dynamics towards an attractor state correlated with the stored pattern, a signature of associative memory. The other order parameter M describes the interference due to the other stored patterns in the connectivity matrix; it is proportional to the average squared firing rates of the network (see Methods). Equations for the order parameters as a function of α, ϕ, f and g are given in Methods.

The results of the simulation of a particular realization of a network of $N = 50,000$ neurons with $c = 0.005$ (an average of 250 connections per neuron) storing $p = 30$ patterns ($\alpha = 0.12$), and the comparison with the results from MFT are shown in Fig. 3. In the simulations, the network was initialized in a state which was uncorrelated with all the stored patterns. For these parameters, the network converged to a ‘background’ state in which all neurons fire at low rates (average 7.98/s, standard deviation 2.92/s). Upon presentation of a novel stimulus (Fig. 3A), neurons were driven to stimulus-specific firing rates, with a distribution of firing rates that was close to a lognormal distribution (Fig. 3C), similar to experimental observations (Lim et al., 2015). The distribution is close to lognormal because the distribution of inputs to neurons is Gaussian, and the neuronal transfer function is close to being exponential at low rates (see Methods). After the end of the presentation of the stimulus, the network came back to its initial background state (Fig. 3A). Upon presentation of a familiar stimulus (Fig. 3D), the statistics of neuronal responses differed markedly from the response to novel stimuli: a few neurons responded at higher rates, but the majority of neurons responded at lower rates compared to a novel stimulus. The distribution of visual responses for familiar stimuli had consequently a lower mean compared to the distribution of responses for novel stimuli but a larger tail at high rates (compare Fig. 3C and F). These two features were consistent with data recorded in ITC by multiple groups (Li et al., 1993; Kobatake et al., 1998; Logothetis et al., 1995; Freedman et al., 2006; Woloszyn and Sheinberg, 2012).

After removal of a familiar stimulus, the network no longer came back to the initial background state, but rather converged to an attractor state that was strongly correlated with the shown stimulus (Fig. 3D), as shown by the strong overlap between the network state and the shown pattern (see blue curve in Fig. 3E). A small fraction of neurons exhibited persistent activity at high rates (4.3% of the neurons are above half maximal rate), but most neurons remained at low rates during the simulated delay period (Fig. 3F). The distribution of firing rates was again similar to a lognormal distribution at low rates, but the tail of the distribution was shaped by neuronal saturation and therefore exhibited a tiny peak close to maximal firing rates. Both overlap with presented pattern and distributions of firing rates could be computed by the MFT and were in close agreement with network simulations (Fig. 3E and F). When the heterogeneity on the neuronal saturation is included into our model by randomly selecting maximal firing rates for each neuron from a log-normal distribution that fits the empirical distribution of the best-fit maximal firing rates (see Fig. 2E), the peak at maximal firing rate disappears. Thus, in a heterogeneous network, distributions of firing rates during both presentation and delay periods become unimodal (Fig. 3F dashed lines).

Thus, our network behaved as an associative memory when constrained by ITC data, without any need for parameter variation or fine tuning. Furthermore, in addition to reproducing the distributions of visual responses for both novel and familiar stimuli seen experimentally, it also exhibited qualitatively some of the main features observed both during spontaneous and delay activity in IT cortex: broad distribution of firing rates in both spontaneous and delay period activity, and small fraction of neurons firing at elevated rates during persistent activity (Miyashita, 1988; Nakamura and Kubota, 1995).

Storage capacity, and its dependence on g

We now turn to the question of the storage capacity of the network, i.e. how many different patterns can be stored in the connectivity matrix. The calculation of the storage capacity of associative memory models such as the Hopfield model was one of the first successful applications of statistical physics to theoretical neuroscience (Amit et al., 1987). One of the main findings of such models is that the number of patterns that can be stored scales linearly with the number of plastic connections per neuron, i.e. the maximal value of α is of order 1. This maximal storage capacity α_c has been computed in many variants of the Hopfield model (see e.g. Amit (1992)). To compute the storage capacity of our network, we found numerically the largest value of α for which retrieval states (i.e. states with positive overlap with one of the stored patterns, $m > 0$) exist. Fig. 4A shows how the overlap in retrieval states m varies as a function of the storage load α, computed using both MFT (solid line) and simulations (symbols with errorbars) when parameters of the functions ϕ and f are taken to be the median best-fit parameters, and those of the function g (except q_g, that is set by the balance condition, Eq. 25) are taken to be identical to f. It shows that m gradually decreases with α, due to more ‘noise’ in the retrieval due to other stored patterns, until it drops abruptly to zero at a value of $\alpha_c = 0.56$. This value is remarkably close to the maximal capacity of the sparsely connected Hopfield model of binary neurons storing binary patterns, for which $\alpha_c = 0.64$ (Derrida et al., 1987).

We then explored how the capacity depends on the parameters of the function g, that describes the dependence of the learning rule on the presynaptic firing rate. Fig. 4B and C show that the capacity is close to being maximized when these parameters match those of the function f, i.e. $x_g = x_f$ and $\beta_g = \beta_f$. Fig. 4B shows that the capacity is non-zero only when the g is sufficiently non-linear, i.e. $\beta_g > 0.1$. It peaks around $\beta_g = \beta_f$, but remains high in the $\beta_g \to \infty$ limit when the function g becomes a step function. Fig. 4C shows that the capacity is non-zero only in a finite range of x_f, between 10 and 30/s. It shows again that capacity peaks when x_g is close to x_f.

Learning rules inferred from ITC data are close to maximizing memory storage

The storage capacity of the network with median parameters is in the same range or higher than the capacity of classic associative memory models of binary neurons - for instance, the Hopfield model has a capacity of $\alpha_c \sim 0.14$ (Amit et al., 1987), while its sparsely connected variant has a capacity of $\alpha_c \sim 0.64$ (Derrida et al., 1987). The next question we addressed is how this capacity depends on the parameters of this learning rule. We have already discussed above the dependence of the capacity on x_g and β_g. Here, we explore the dependence on the four remaining parameters characterizing the learning rule - A, x_f, β_f and q_f. Using MFT, we explored systematically the space of these four parameters, and plot in Fig. 5 all possible cuts of this four dimensional space, in which 2 of the 4 parameters are varied, while the other 2 are set to the median values. In all these plots, the maximal capacity α_c is plotted as a function of two parameters, using a gray scale (white indicate high capacity, black low capacity). The yellow dashed line indicates the line for which the function f is 'balanced' (i.e. its average across the distribution of patterns is zero). It marks the border between a depression-dominated region, for which learning leads to a decrease in average responses, and a potentiation-dominated region, for which learning leads to an increase of such responses. The red cross mark indicates the median parameters, while the dashed red rectangle indicates the interquartile range.

Fig. 5 shows that the median parameters are close to maximizing storage capacity. In fact, we found that the maximal capacity over this space is $\alpha_c \approx 0.85$ (see Fig. S6–7 and methods S1 for details). These figures show also that most (but not all) of the interquartile range lie in a high-capacity region. It also shows that some parameter variations lead to little changes in capacity, while others lead to a drastic drop. Decreasing the learning strength A from its optimal value leads to an abrupt drop in capacity, while increasing it leads to a much gentler decrease (see Fig. 5D–F). A similar effect is observed for the slope of f; decreasing the slope (i.e. making f more linear) leads to an abrupt decrease in capacity, while increasing it beyond the median value leads to very little change in capacity (see Fig. 5B–D). Thresholds x_f for which high capacities are obtained are much higher than the mean response to novel visual stimuli (Fig. 5A,B and D), leading to a sparsening of the representations of the patterns by the network. Finally, the optimal offset is close to the 'balanced' line, but slightly on the depression-dominated region, as the median parameter (Fig. 5A,C and F).

A chaotic phase with associative memory properties

Are fixed point attractors the only possible dynamical regime in this network? Firing rate models with asymmetric connectivity have been shown to exhibit strongly chaotic states (Sompolinsky et al., 1988; Tirozzi and Tsodyks, 1991). Varying parameters of the learning rule, we found parameter regions in which background and/or retrieval fixed point attractor states destabilize and the network settle into strongly chaotic states. Fig. 6A shows an example of such chaotic states, obtained for the median parameters as in Fig. 3, except for the learning rate which is three times its median best-fit value (A = 10.65). For such parameters, the background state is strongly chaotic. Presentation of a familiar stimulus leads to a transition to another chaotic state, in which all neurons fluctuate chaotically around stimulus-specific firing rates, such that the mean overlap with the corresponding pattern remains high (see Fig. 6 B). Remarkably, chaotic retrieval states remain strongly correlated with the corresponding patterns (see Fig. 6B), so that the network can still perform as an associative memory in spite of the chaotic fluctuations of network activity. Interestingly, the storage capacity for such parameters is larger than the capacity estimated from the static MFT (see Fig. 6C).

In such chaotic retrieval states, single neuron activity exhibit strong firing rate fluctuations which vary from trial to trial (see thin colored lines in Fig. 6D–F showing three randomly selected neurons), but trial-averaged firing rates show systematic temporal patterns. For instance, the activity of the neuron shown in Fig. 6D ramps up in the first second of the delay period, before this activity plateaus at a rate of about 40/s. The neuron shown in Fig. 6F shows a rapid activity increase during the presentation period, followed by a trough, followed by a second increase during the delay period. These temporal patterns of the trial-averaged firing rate, together with a strong irregularity within trials, are reminiscent of observations by multiple groups in primate PFC during delay periods (Shafi et al., 2007; Brody et al., 2003; Murray et al., 2017).

To check whether these states are truly chaotic, we computed the temporal evolution of the distance between two network states with slightly different initial conditions (see Methods). Fig. 6G shows that an initial distance between two initial conditions of $4.5 \cdot 10^{-6}$Hz exponentially grows and then plateaus to an average of ~ 13Hz. This sensitivity to initial conditions, and initial exponential growth of the distance between perturbed and unperturbed network states is the defining feature of a chaotic system (Guckenheimer and Holmes, 2013). The divergence of the network states starts to be noticeable in the single neuron dynamics in about ~1s (see Fig. 6H). However, the overlap with the stored pattern remains high in both networks states (see Fig. 6I). Therefore, despite the growth of the distance between the two network states, their dynamics keep aligned to the 1-dimensional subspace (of the full N-dimensional network space) spanned by the retrieved memory, providing a low dimensional representation of each memory.

Across neurons, for both the background and retrieval state, the chaotic fluctuations in the rates have a distinctive times scale of about 100ms (see Fig. 7A). However, there is a broad diversity of time scales for individual neurons, ranging from about ~ 50ms to ~ 500ms (see Fig. 7A, light traces). Low firing rate neurons have slightly slower time scales than high firing rate neurons. Neurons are weakly correlated, for both background and retrieval states (see Fig. 7B). Lastly, the distributions of the mean firing rates are qualitatively similar to the ones described for the fixed-point attractor scenario (compare Fig. 3C and F with Fig. 7C), but with a higher proportion of neurons at very low rates.

Discussion

We have shown that a learning rule inferred from data generate attractor dynamics, without any need for parameter adjustment or tuning, except for the condition that the dependence of the learning rule on the presynaptic rate should be ‘balanced’ (i.e. have a zero average over the distribution of visual responses, see below). Furthermore, this rule produces a storage capacity that is close to the maximal capacity, in the space of unsupervised Hebbian learning rules with sigmoidal dependence on both pre and post-synaptic firing rates. Remarkably, similar to the learning rules inferred from ITC recordings, learning rules derived from memory storage maximization depress the bulk of the distribution of the learned inputs (those that lead to low to intermediate firing rates) while potentiating outliers (those that lead to high rates), leading to a sparse representation of stored memories. The attractor states generated by our model are characterized by graded activity with a continuous range of firing rates (Treves, 1990*a,b*; Festa et al., 2014). Most of the distribution lies in the low rate region of the neuronal transfer function, leading to a strongly skewed distribution, with a small fraction of neurons firing at higher rates. These observations are consistent with the available data in ITC during delay match to sample experiments (Miyashita, 1988; Nakamura and Kubota, 1995).

For a range of parameters values consistent with learning rules inferred from data, our model presents irregular temporal dynamics for retrieval states, similar to the temporal and across trial variability observed during delay periods in multiple studies (Murray et al., 2017). In this regime, retrieval states are chaotic, yet they maintain non-zero overlap with the corresponding memories. Thus, the network performs robustly as an associative memory device, even though strong fluctuations are internally generated by its own chaotic dynamics.

Distribution of firing rates

Our model naturally gives rise to highly skewed distributions of firing rates, consistent with those that have been observed during presentation of visual stimuli in ITC (Lehky et al., 2011; Lim et al., 2015) and during delay periods of DMS tasks (Miyashita, 1988; Nakamura and Kubota, 1995). By construction of the model, it also reproduces the decrease in the mean response with familiarity, and the increase in selectivity with familiarity. Our model shows for most of the explored parameter space a weak bimodality in the distribution of firing rates due to neuronal saturation in response to familiar stimuli, with a tiny peak close to neuronal saturation, when the network is homogeneous. When heterogeneity in maximal firing rates is implemented in the network, the peak at high firing rates disappears and the distribution of firing rates becomes unimodal.

Learning rule

The learning rule we have used in our network model was inferred from ITC data (Lim et al., 2015). It is an unsupervised Hebbian rule, as it only depends on the pre and post-synaptic firing rates, and it leads to potentiation for large pre and post-synaptic rates. As other popular examples of Hebbian rules such as the covariance rule (Sejnowski, 1977) or the BCM rule (Bienenstock et al., 1982), it is separable in pre and post-synaptic rates. Unlike the covariance rule, but similar to other Hebbian rules (Bienenstock et al., 1982; Senn et al., 2001; Pfister and Gerstner, 2006), it is strongly non-linear as a function of the post-synaptic firing rate. It reproduces some of the phenomenology of the dependence of synaptic plasticity on pre and post-synaptic firing rates in cortical slices; in particular, large pre and post-synaptic firing rates lead to LTP (Sjöström et al., 2001). Large pre-synaptic firing rate in conjunction with low post-synaptic firing rate, lead to depression, consistent with ‘pairing’ experiments in which LTD is triggered by pre-synaptic activity, together with intermediate values of the membrane potential (Ngezahayo et al., 2000). Plasticity at low pre-synaptic firing rates could be due to plasticity mechanisms leading to ‘normalization’ or homeostasis. Indeed, our plasticity rule could be written as $\Delta J_{ij} = \Delta J_{ij}^{Hebb} + \Delta J_{ij}^{hom}$ where $\Delta J_{ij}^{Hebb} = Af(r_i)(g(r_j) - g(0))$, $\Delta J_{ij}^{hom} = Af(r_i)g(0)$. The ‘homeostatic’ component ΔJ_{ij}^{hom} leads to a decrease in the efficacy of all synapses onto a post-synaptic neuron when the neuron is firing at high rates, while it leads to an increase when the neuron fires at low rates (since $g(0) < 0$). Note that such a homeostatic mechanism would also automatically lead to a ‘balanced’ dependence of the rule of the pre-synaptic firing rate, which is necessary for the network to be able to store a large nuber of patterns. The analysis described in the Supplementary Material shows that if g has a non-zero average, then the mean of the noise term due to other patterns stored in the connectivity matrix would no longer be zero, but rather scale as $\alpha cN\langle g\rangle$, where $\langle g\rangle$ is the average of g over the distribution of visual responses. This has the consequence that the network would be able to store only a finite number of patterns. A precise balance could be restored by the homeostatic mechanism mentioned above - for a non-zero $\langle g\rangle$, this homeostatic term would become $\Delta J_{ij}^{hom} = Af(r_i)(g(0) - \langle g\rangle)$, which would ensure that the average synaptic strength (and consequently mean firing rate) onto a neuron remains constant with learning.

The synaptic connectivity matrix we used is assumed to be generated through multiple presentations of initially novel patterns. The simplest implementation of this plasticity rule consists in adding a term ΔJ_{ij} to the current matrix, as described above, but only when a novel pattern is presented to the network. This would require a novelty detector that would gate plasticity, perhaps through neuromodulators. An interesting hypothesis is that novelty detection could be generated by the network itself, through its mean activity (which is significantly higher for novel than for familiar stimuli). This novelty signal could in principle then be used to trigger learning.

To derive the learning rule, we used a subset of the data recorded by Woloszyn and Sheinberg (2012), i.e. excitatory neurons that show negative changes at low rates and positive changes at high rates. Those neurons are approximately half (14/30) of the neurons that showed significant differences between the distributions of visual responses for familiar and novel stimuli. Out the remaining 16 neurons, 10 showed negative changes for all rates, while 6 showed the opposite pattern of positive changes for all rates. This heterogeneity in inferred learning rules could be due to a heterogeneity in neuronal properties - for instance, it could be that the ‘putative’ excitatory neurons recorded in this study form a heterogeneous group of cells, some of which might actually be inhibitory. Consistent with this, some inhibitory neuron classes have electrophysiological properties (and in particular, spike width) that are closer to pyramidal cells that to fast-spiking interneurons. Another possibility is that part of the apparent heterogeneity stems form the same underlying learning rule, but with heterogeneous parameters. For instance, inferred learning rules with negative changes at all rates are consistent with a sigmoidal post-synaptic dependence f, but with a high threshold x_f that lies above the range of firing rates elicited in that particular experiment. Elucidating which of these scenarios hold in IT cortex will need recordings from more neurons, as well as recordings of single neurons with more stimuli.

Our approach is complementary to other studies that have inferred learning rules from *in vitro* studies, and then shown that these rules lead to attractor dynamics in large networks of spiking neurons (Litwin-Kumar and Doiron, 2014; Zenke et al., 2015). In contrast to these studies, we showed that a network with a learning rule inferred from *in vivo* data can achieve a high storage capacity, and generate graded distributions of firing rates during visual presentation and delay periods. An important difference between the studies of Litwin-Kumar and Doiron (2014) and Zenke et al. (2015) is that they used an online learning rule that is constantly active, while our connectivity matrix is assumed to be frozen following the learning process. It will be interesting to investigate whether, and in which conditions spike-timing and voltage based learning rules used in such studies can produce a firing rate dependence that is consistent with the rule used here.

Time-varying neural representations

In recent years, the standard attractor network scenario has been challenged by multiple observations of strong variability and non-stationarity during the delay period in prefrontal cortex (Compte et al., 2003; Shafi et al., 2007; Barak et al., 2010; Barak and Tsodyks, 2014; Kobak et al., 2016; Murray et al., 2017). Statistical analysis of recordings in this area during two different working memory tasks has shown that variability observed during delay periods is consistent with static coding of the stimulus kept in memory (Murray et al., 2017). Various models have been proposed to account for variability and/or non-stationarity (Barbieri and Brunel, 2007; Mongillo et al., 2008; Lundqvist et al., 2010; Mongillo et al., 2012; Druckmann and Chklovskii, 2012).

Here we propose an alternative mechanism where chaotic attractors with associative memory properties naturally generate the time-varying irregular activity observed during delay periods in associative memory tasks. In this state, chaotic attractors correspond to internal representations of stored memories. Each chaotic attractor state maintains a positive overlap with the corresponding stored memory. In this scenario, the network performs as an associative memory device where temporal variability is generated internally by chaos. This model naturally exhibits the combination of strong temporaly dynamics yet stable memory encoding which has been demonstrated in PFC by various groups (Druckmann and Chklovskii, 2012; Murray et al., 2017). It will be interesting to compare this model to existing data, using for instance methods used in Murray et al. (2017).

There has been a longstanding debate whether the type of chaotic states seen in firing rate models can be seen also in spiking network models under the form of ‘rate chaos’. Recent studies indicate that this type of chaos can be observed provided coupling is sufficiently strong, as in firing rate models Ostojic (2014); Harish and Hansel (2015); Kadmon and Sompolinsky (2015). Thus, it is reasonable to expect that the type of retrieval chaotic states we observed in our network can also be realized in networks of spiking neurons.

Optimality criteria for information storage

Here, we have argued that learning rules that are inferred from electrophysiological recordings in ITC of behaving primates are close to optimizing information storage, in the space of unsupervised Hebbian learning rules that have a sigmoidal dependence on both pre and post-synaptic firing rates. Such learning rules are appealing because synapses do not need to know anything beyond the firing rates of pre and post-synaptic neurons to form memories, two quantities that are easily available at a synapse. However, one cannot exclude that the dependence of plasticity on neuronal activity takes other forms than the one investigated here. In particular, a potentially more powerful approach proposed by Gardner

(1987) relies in maximizing the number of attractors in the space of all possible synaptic matrices. Unsurprisingly, this approach leads in general to a larger capacity than the ones that can be achieved by unsupervised Hebbian rules, but it turns out that in sparse coding limit, the covariance rule reaches asymptotically the Gardner bound (Tsodyks and Feigel'Man, 1988; Tsodyks, 1988). These results have been obtained in networks of binary neurons, and it remains to be investigated whether similar results could be obtained in networks of analog firing rate neurons. An additional challenge in comparing the two approaches in such networks is that the stored attractors are in our case not identical to the pattern that was initially shown to the network, while in the standard Gardner approach, the two were constrained to be identical.

Another motivation for considering the Gardner approach is provided by a recent study that showed that synaptic connectivity in a network of excitatory binary neurons that maximizes storage capacity in the space of all possible matrices reproduces a number of basic experimental facts on cortical excitatory connectivity (Brunel, 2016): Low connection probability (Markram et al., 1997; Sjöström et al., 2001; Lefort et al., 2009), in spite of full potential connectivity (Kalisman et al., 2005); And strong over-representation of bidirectionnally connected pairs of neurons compared to a random Erdos-Renyi network (Sjöström et al., 2001). In contrast with the network studied by Brunel (2016), the synaptic connectivity of the model proposed here has the unrealistic feature that it does not obey Dale's law. One could reconcile the present model with cortical connectivity by using a connectivity matrix that is a rectified version of Eq. (2) - such a connectivity matrix would then obey Dale's law, be sparse and be more symmetric than a random Erdos-Renyi network, making it therefore consistent with slice data. Such a generalization is beyond the scope of the present paper and will be the subject of a future study.

Altogether, our results strongly reinforce the link between attractor network theory and electrophysiological data during delayed response tasks in primates. Furthermore, they suggest that learning rules in association cortex are close to maximizing the number of possible internal representations of memories as attractor states.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Nicolas Brunel (nicolas.brunel@duke.edu).

METHOD DETAILS

Mean field theory—Here we present the main results of our mean field analysis that quantifies the retrieval of a particular familiar pattern during the delay period. Detailed calculations for this case are presented in section 1 of methods S1. The analysis is performed in the limit $p, N \to \infty$ and $c \ll 1$.

In our model, memories are defined as the patterns of external synaptic inputs that were present when the corresponding stimulus was shown for the first time to the network. These external synaptic inputs $\{\vec{\xi}^{k}\}_{k=1}^{p}$, where k labels individual memories, are independent and identical distributed (i.i.d.) Gaussian random variables with zero mean and unit variance. These memories are imprinted in the connectivity matrix using the learning rule described in Eq. (2). The firing rates $r_i(t)$ of neurons $i = 1, \ldots, N$ evolve according to the rate equations (Grossberg, 1969; Hopfield, 1984), i.e. Eq. (1).

During the delay period, the external stimulus I_i is set to be zero. The steady states or fixed point attractors for the dynamics are given by the following set of nonlinear equations

$$r_i = \phi\left(\sum_{i \neq j}^{N} J_{ij} r_j\right) \quad i = 1, \ldots, N. \quad (3)$$

To describe the statistics of the firing rates in a fixed point described by Eq. (3), we first need to compute the statistics of the incoming current to a given neuron, $h_i = \sum_{i \neq j}^{N} J_{ij} r_j$, assuming that the network state is correlated with one of the stored patterns (without loss of generality, we choose here the first pattern ξ_i^1), but uncorrelated with all other patterns. In the large N limit, the distribution of this current, conditioned on the value of ξ_i^1, becomes a Gaussian. The mean μ conditioned on ξ_i^1 is given by

$$\mu(\xi_i^1) = Af(\phi(\xi_i^1))q, \quad (4)$$

where q is the covariance between a non-linear transformation of the pattern $g(\phi(\xi_i^1))$ and the firing rates in the current network state r_i,

$$q = \frac{1}{N}\sum_{i=1}^{N} g(\phi(\xi_i^1))r_i \quad (5)$$

The ‘overlap’ m described in the main text is the corresponding correlation coefficient, i.e. normalized by the square root of the variances of $g(\phi(\xi_i^1))$ and r_i.

The variance of input currents (due to the other stored patterns that act as a quenched source of noise on the retrieval of the pattern of interest) is given by

$$\sigma^2 = \alpha\gamma M. \quad (6)$$

Where γ depends on the learning rule and statistics of the patterns as

$$\gamma \equiv A^2 \int_{-\infty}^{\infty} \mathscr{D}z f(\phi(z))^2 \int_{-\infty}^{\infty} \mathscr{D}z g(\phi(z))^2 \quad (7)$$

while *M* is the average squared firing rate

$$M = \frac{1}{N} \sum_{i=1}^{N} r_i^2. \quad (8)$$

The next step is to compute self-consistent equations for the 'order parameters' q and M, that fully describe the macroscopic behavior of the network. Inserting Eq. (3) in Eq. (5), using the fact that h_i has a Gaussian distribution with mean μ and variance σ^2, and replacing the sum over i by an integral over ξ_i, we obtain

$$q = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathscr{D}z \mathscr{D}y g(\phi(z)) \phi(qAf(\phi(z)) + \sqrt{\alpha\gamma M} y) \quad (9)$$

where the integral over z corresponds to an integral over the distribution of the patterns ξ_i^1, while the integral over y corresponds to an integral over the distribution of the 'quenched noise' due to other stored patterns.

Similarly, inserting Eq. (3) in Eq. (8) and using again the fact that h_i is Gaussian distributed, we find

$$M = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathscr{D}z \mathscr{D}y \phi^2(qAf(\phi(z)) + \sqrt{\alpha\gamma M} y). \quad (10)$$

For a given value of α, and functions ϕ, f and g, Eqs. (9,10) are solved numerically by using a gradient free approach where the equations are iterated as a discrete map from an arbitrary initial condition (i.e. $q_0 > 0$ and $M_0 > 0$) until convergence. Note that Eq. (9) always have a solution $q = 0$, which correspond to a background state which is uncorrelated with all stored patterns. Solutions of these equations with $q > 0$ indicate the presence of retrieval states.

The distribution of firing rates can be obtained as

$$p_r(r) = \int_{-\infty}^{\infty} \mathscr{D}z \frac{e^{-\frac{(\phi^{-1}(r) - Af(\phi(z))q)^2}{2\alpha\gamma M}}}{\sqrt{2\pi\alpha\gamma M}} \frac{d\phi^{-1}(r)}{dr}, \quad (11)$$

where the order parameters q and M are determined by the self-consistent equations (9) and (10). The overlap m is given by

$$m = \frac{q}{(M - R^2)\sqrt{\int_{-\infty}^{\infty} \mathscr{D}z g(\phi(z))^2}}, \quad (12)$$

where R is the mean firing rate in the attractor state given by

$$R = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathscr{D}z\mathscr{D}y\phi(qAf(\phi(z)) + \sqrt{\alpha\gamma M}y)\,. \quad (13)$$

Similar analysis can be performed in the presence of an external input (presentation period), for both familiar and novel stimuli. Details are presented in section 2 of methods S1. The calculations proceed along the lines of the calculations presented above, except that: (1) During the presentation of a familiar stimulus, the external input currents are set to $I_i = I_0\xi_i^1$; (2) During the presentation of a novel stimulus, the external inputs are set to $\vec{I} = I_0\vec{\eta}$, where $\vec{\eta}$ is an independent and identical distributed standard normal pattern of currents (i.e. $\eta_i \overset{iid}{\sim} \mathscr{N}(0, 1)$ with $i = 1, 2, \ldots, N$), uncorrelated with all learned patterns. For simulations in Figures 3 and 6 $I_0 = 1$ during the presentation period.

Simulations—For most simulations shown in this paper, the probability of connections was set to 0.5% (i.e. $c = 0.005$) and the number of neurons to $N = 50000$, which implies an average number of connections per neuron of $Nc = 250$. The choice of a low connection probability was motivated by the fact that the MFT is exact in the sparse connectivity limit (see methods S1 and Derrida et al. (1987); Kree and Zippelius (1987)). We have also simulated networks with with various values of N and c (see Fig. S5). These simulations show that our theory gives good quantitative predictions for denser connectivities. The single neuron time constant was chosen as $\tau = 20ms$, similar to time constants of single neurons (McCormick et al., 1985) and synapses (Destexhe et al., 1998), and with the decay time constant of cortical activity as measured *in vivo* (Reinhold et al., 2015). Open source built-in linear algebra methods in scipy and numpy Python packages suited for sparse matrices were used to generate the connectivity matrix. For simulating the networks dynamics, the Euler method was used with a time step size of 0.5ms. For a few parameter sets, we checked that results are unchanged when a smaller value of $dt = 0.1$ms is used. In the simulations, the background state was sometimes unstable, and the dynamics in this case converged to one of the 'memory states'. This tended to happen in particular for small values of α.

In Fig. 6 G–I, the Runge-Kutta fourth-order method with $dt = 0.1$ms was used. In Fig. 7 the autoand cross-correlation functions are computed over 100 realizations of a 8s network simulation. For retrieval states, in each realization the input current is given by the current corresponding to the stored pattern plus a random vector whose entries are i.i.d. random Gaussian variables with zero mean and S.D. 0.2. For the background state, the initial condition of the dynamics are the firing rates obtained from passing an i.i.d. standard normal vector through the transfer function Φ. The first second of simulation is not taken into account to compute auto and cross-correlation functions. Only neurons with mean firing rates between 1Hz and 65Hz are selected in order to avoid numerical artifacts arising from neurons whose mean firing rates stay close to zero or to the maximum firing rate during most of the simulation.

To measure the sensitivity of the network dynamics to small perturbations, we choose two slightly different initial conditions and follow the dynamics of the network following both initial conditions, to investigate whether these two initial conditions converge to the same state (indicating non-chaotic dynamics), or vice versa diverge exponentially (indicating chaotic dynamics). These two slightly different initial conditions are generated as follows

$$\vec{r}^{\,(1)}_{k}(0) = \phi(\vec{\xi}^{\,k}) \quad (14)$$

$$\vec{r}^{\,(2)}(0) = \phi(\vec{\xi}^{\,k}) + \vec{\eta}\,\frac{\delta}{\|\vec{\eta}\|_2}. \quad (15)$$

where the index k corresponds to one of the p stored patterns (i.e. $k \in \{1, 2, \ldots, p\}$), $\delta = 10^{-3}$ is the distance between the initial conditions and $\vec{\eta}$ is an independent and identically distributed Gaussian vector. Thus, $\vec{r}^{\,(1)}_{k}(0)$ is the firing rate produced by the kth stored pattern, while $\vec{r}^{\,(2)}(0)$ is a slightly perturbed version of this pattern. We define the distance between the two network states during the time evaluation of the dynamics by

$$d_k(t) = \frac{\left\|\vec{r}^{\,(1)}_{k}(t) - \vec{r}^{\,(2)}_{k}(t)\right\|_2}{\sqrt{N}}. \quad (16)$$

This distance gives the typical difference between the firing rates of a single neuron between two network states produced by slightly different initial conditions at time t, for the retrieval state corresponding to pattern k, and has units of Hz.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data analysis—We reanalyze the data recorded by Luke Woloszyn and David Sheinberg (Woloszyn and Sheinberg, 2012) using the method described in Lim et al. (2015). This data consists in trial-averaged firing rates of individual neurons in ITC (in a time window between 75 ms and 200 ms after stimulus onset) in response to 125 novel and 125 familiar stimuli measured, during a passive fixation task. We focused on the 30 putative excitatory neurons whose distributions of visual responses for novel and familiar stimuli were significantly different, using the Mann-Whitney U test at 5 significance level. In these neurons, the postsynaptic dependence of the learning rule, was inferred using the method described in Lim et al. (2015). In this subset of neurons, we focused on 14 excitatory neurons, the ones that show negative input changes for low firing rates and positive input changes for high firing rates. For these 14 neurons, the transfer function ϕ, and the postsynaptic dependence of the learning rule, f, are inferred using the method described in Lim et al. (2015).

The first step is to infer the transfer function ϕ. We assume that inputs to neurons during presentation of novel stimuli have a Gaussian distribution. The transfer function is then obtained as the function ϕ that maps a standard Gaussian to the empirical distribution of firing rates for novel stimuli (Lim et al., 2015). In practice, the function is obtained by building a quantile-quantile plot between the distribution of firing rates for novel stimuli and the assumed standard normal distribution of inputs (see Fig. 2 A and B and S2–3). The obtained transfer function (blue circles in Fig. 2) was fitted with the sigmoidal function

$$\phi_i(\xi) = \frac{r_m^{(i)}}{1 + e^{-\beta_T^{(i)}(\xi - h_0^{(i)})}} \quad (17)$$

where $r_m^{(i)}$ is the maximal firing rate, $\beta_T^{(i)}$ measures the slope at the inflection point, and $h_0^{(i)}$ is the location of this inflection point. h_0 is also the current leading to half maximal firing rate. These parameters were obtained by minimizing the squared error. We thus obtained for each of the 14 neurons the best estimators $r_m^{(i)}, \beta_T^{(i)}$ and $h_0^{(i)}$ with $i = 1, 2, \ldots, 14$ whose statistics are summarized in Fig. 2D.

The next step is to infer the postsynaptic dependence of the learning rule, f. For this, we use the difference between the distributions of visual responses to novel and familiar stimuli (Lim et al., 2015). In the model, learning of a novel stimulus defined by inputs ξ_i^k that leads to firing rates $r_i^k = \phi(\xi_i^k)$ leads to changes in recurrent inputs, due to changes in synaptic inputs

$$\Delta J_{ij} = \frac{Ac_{ij}}{cN} f(r_i^k) g(r_j^k) \quad (18)$$

This leads to a change in total inputs to neurons that is proportional to

$$\Delta h_i = Af(r_i^k) \frac{1}{cN} \sum_j c_{ij} g(r_j^k) r_j^k \quad (19)$$

In the large N limit, Eq. (19) becomes

$$\Delta h_i = Af(r_i^k) \int_{-\infty}^{\infty} \mathscr{D}z g(\phi(z))\phi(z) . \quad (20)$$

where $\mathscr{D}z$ is the standard Gaussian measure, $\mathscr{D}_z = dze^{-z^2/2}/\sqrt{2\pi}$. Eq. (20) give us the relationship between changes of total inputs to a neuron with learning of a particular stimulus, and the firing rate of the neuron upon presentation of that stimulus for the first time. This relationship can be inferred from the data by computing the difference between the quantile function of visual responses to familiar stimuli and the quantile function of visual responses to novel stimuli, and by plotting this difference as a function of visual response to novel stimuli (Lim et al., 2015). We then fitted the input change with a sigmoidal function given by

$$\Delta h_i^{fit}(r) = \frac{C^{(i)}}{2}\left[2q_f^{(i)} - 1 + \tanh(\beta_f^{(i)}(r - x_f^{(i)}))\right]. \quad (21)$$

where $C^{(i)}$ gives the amplitude of the total changes, q_f^i measures the vertical offset of the curve (for $q_f = 1$, Δh is non-negative at all rates, while for $q_f = 0$ it is non-positive at all rates), $\beta_f^{(i)}$ measures the slope at the inflection point, and $x_f^{(i)}$ is the rate at the inflection point. In the following, we refer to $x_f^{(i)}$ as the threshold since it is typically very close to the rate at which Δh changes sign. For each of the 14 neurons, the parameters $C^{(i)}$, $q_f^{(i)}, \beta_f^{(i)}$ and $x_f^{(i)}$ with $i = 1, 2, \ldots, 14$ were estimated by minimizing the squared error. The inferred function f for each neuron is given by

$$f_i(r) = \frac{\Delta h_i^{fit}(r)}{C^{(i)}} = \frac{1}{2}\left[2q_f^{(i)} - 1 + \tanh(\beta_f^{(i)}(r - x_f^{(i)}))\right]. \quad (22)$$

The parameter A is then obtained as

$$A^{(i)} = \frac{C^{(i)}}{\int_{-\infty}^{\infty} \mathscr{D}z g(\tilde{\phi}(z))\tilde{\phi}(z)}, \quad (23)$$

where $\tilde{\phi}$ is the sigmoidal transfer function in Eq. (23) whose parameters are the medians of the fitted parameters. The function g was also chosen to be a sigmoid, given by

$$g(r) = \frac{1}{2}[2q_g - 1 + \tanh(\beta_g(r - x_g))], \quad (24)$$

with q_g set such that the average change in connection strength due to learning of a single pattern is zero, i.e.

$$\int_{-\infty}^{\infty} \mathscr{D}z g(\tilde{\phi}(z)) = 0 . \quad (25)$$

Note that g is unconstrained by data. For most of the paper, we set the slope and the threshold for g to the median of the fitted parameters for f, i.e. $\beta_g = \tilde{\beta}_f$ and $x_g = \tilde{x}_f$. We also explored how the capacity depends on β_g and x_g, as shown in Fig. 3.

DATA AND SOFTWARE AVAILABILITY

Software was written in the Python (http://python.org) programming language. Network simulations and algorithms for solving the mean field equations and computing the capacity of the network are available at the GitHub repository: https://github.com/ulisespereira/AttractorDynamics

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Yonatan Aljadeff and Yali Amit for useful discussions and comments on previous versions of the manuscript. We also thank David Sheinberg for providing his data; Sukbin Lim for providing data analysis codes, and useful comments on the manuscript; Maxwell Gillett for useful discussions on optimizing network simulations; And anonymous reviewers whose comments have helped to improve the paper significantly. This work was supported by NIH R01 MH115555, NIH R01 EB022891 and ONR N00014-16-1-2327.

References

Amari S-I. Learning patterns and pattern sequences by self-organizing nets of threshold elements. IEEE Transactions on Computers. 1972; 100(11):1197–1206.

Amit D, Gutfreund H, Sompolinsky H. Statistical mechanics of neural networks near saturation. Annals of physics. 1987; 173(1):30–67.

Amit DJ. Modeling brain function: The world of attractor neural networks. Cambridge University Press; 1992.

Amit DJ. The hebbian paradigm reintegrated: local reverberations as internal representations. Behav. Brain Sci. 1995; 18:617.

Amit DJ, Brunel N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cerebral cortex. 1997; 7(3):237–252. [PubMed: 9143444]

Amit DJ, Fusi S. Learning in neural networks with material synapses. Neural Computation. 1994; 6(5): 957–982.

Anderson JS, Lampl I, Gillespie DC, Ferster D. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. Science. 2000; 290:1968–1972. [PubMed: 11110664]

Barak O, Sussillo D, Romo R, Tsodyks M, Abbott LF. From fixed points to chaos: three models of delayed discrimination. Prog. Neurobiol. 2013; 103:214–222. [PubMed: 23438479]

Barak O, Tsodyks M. Working models of working memory. Curr. Opin. Neurobiol. 2014; 25:20–24. [PubMed: 24709596]

Barak O, Tsodyks M, Romo R. Neuronal population coding of parametric working memory. J. Neurosci. 2010; 30:9424–9430. [PubMed: 20631171]

Barbieri F, Brunel N. Irregular persistent activity induced by synaptic excitatory feedback. Frontiers in Computational Neuroscience. 2007; 1:5. [PubMed: 18946527]

Bienenstock E, Cooper L, Munro P. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. J. Neurosci. 1982; 2:32–48. [PubMed: 7054394]

Brody CD, Hernández A, Zainos A, Romo R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cerebral cortex. 2003; 13(11):1196–1207. [PubMed: 14576211]

Brunel N. Network models of memory. In: Chow C, Gutkin B, Hansel D, Meunier C, Dalibard J, editorsMethods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School 2003. Elsevier; 2005.

Brunel N. Is cortical connectivity optimized for storing information? Nature neuroscience. 2016

Brunel N, Wang XJ. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J. Comput. Neurosci. 2001; 11:63–85. [PubMed: 11524578]

Buzsaki G, Mizuseki K. The log-dynamic brain: how skewed distributions affect network operations. Nat. Rev. Neurosci. 2014; 15:264–278. [PubMed: 24569488]

Compte A, Constantinidis C, Tegnér J, Raghavachari S, Chafee M, Goldman-Rakic PS, Wang X-J. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J. Neurophysiol. 2003; 90:3441–3454. [PubMed: 12773500]

Derrida B, Gardner E, Zippelius A. An exactly solvable asymmetric neural network model. Europhys. Lett. 1987; 4:167–173.

Destexhe A, Mainen ZF, Sejnowski TJ. Kinetic models of synaptic transmission. In: Koch C, Segev I, editorsMethods in Neuronal Modeling. 2. MIT press; Cambridge, MA: 1998. 1–25.

Druckmann S, Chklovskii DB. Neuronal circuits underlying persistent representations despite time varying activity. Current Biology. 2012; 22(22):2095–2103. [PubMed: 23084992]

Festa D, Hennequin G, Lengyel M. Analog memories in a balanced rate-based network of e-i neurons. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editorsAdvances in Neural Information Processing Systems 27. Curran Associates, Inc; 2014. 2231–2239.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. Cerebral Cortex. 2006; 16(11):1631–1644. [PubMed: 16400159]

Funahashi S, Bruce CJ, Goldman-Rakic PS. Mnemonic coding of visual space in the monkey’s dorsolateral prefrontal cortex. Journal of neurophysiology. 1989; 61(2):331–349. [PubMed: 2918358]

Fuster JM, Alexander GE, et al. Neuron activity related to short-term memory. Science. 1971; 173(3997):652–654. [PubMed: 4998337]

Fuster JM, Jervey JP. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science. 1981; 212(4497):952–955. [PubMed: 7233192]

Gardner E. Maximum storage capacity in neural networks. EPL (Europhysics Letters). 1987; 4(4):481.

Gerstner W, van Hemmen JL. Associative memory in a network of ?spiking?neurons. Network: Computation in Neural Systems. 1992; 3(2):139–164.

Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995; 14(3):477–485. [PubMed: 7695894]

Grossberg S. On learning, information, lateral inhibition, and transmitters. Mathematical Biosciences. 1969; 4(3–4):255–310.

Guckenheimer J, Holmes P. Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Vol. 42. Springer Science & Business Media; 2013.

Guo ZV, Li N, Huber D, Ophir E, Gutnisky D, Ting JT, Feng G, Svoboda K. Flow of cortical activity underlying a tactile decision in mice. Neuron. 2014; 81:179–194. [PubMed: 24361077]

Harish O, Hansel D. Asynchronous rate chaos in spiking neuronal circuits. PLoS Comput Biol. 2015; 11(7):e1004266. [PubMed: 26230679]

Hebb D. The organization of behavior: A neuropsychological theory. John Wiley; 1949.

Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences. 1982; 79(8):2554–2558.

Hopfield JJ. Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. U.S.A. 1984; 81:3088–3092. [PubMed: 6587342]

Hromadka T, Deweese MR, Zador AM. Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol. 2008; 6:e16. [PubMed: 18232737]

Inagaki HK, Fontolan L, Romani S, Svoboda K. Discrete attractor dynamics underlying selective persistent activity in frontal cortex. biorxiv. 2017

Kadmon J, Sompolinsky H. Transition to chaos in random neuronal networks. Phys. Rev. X. 2015; 5:041030.

Kalisman N, Silberberg G, Markram H. The neocortical microcircuit as a tabula rasa. Proc Natl Acad Sci U S A. 2005; 102:880–885. [PubMed: 15630093]

Kobak D, Brendel W, Constantinidis C, Feierstein CE, Kepecs A, Mainen ZF, Qi XL, Romo R, Uchida N, Machens CK. Demixed principal component analysis of neural population data. Elife. 2016; 5

Kobatake E, Wang G, Tanaka K. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. Journal of Neurophysiology. 1998; 80(1):324–330. [PubMed: 9658053]

Koch KW, Fuster JM. Unit activity in monkey parietal cortex related to haptic perception and temporary memory. Exp. Brain Res. 1989; 76:292–306. [PubMed: 2767186]

Kree R, Zippelius A. Continuous-time dynamics of asymmetrically diluted neural networks. Phys Rev A Gen Phys. 1987; 36:4421–4427. [PubMed: 9899399]

Lansner A. Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. Trends Neurosci. 2009; 32(3):178–186. [PubMed: 19187979]

Lefort S, Tomm C, Floyd Sarria JC, Petersen CC. The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. Neuron. 2009; 61:301–316. [PubMed: 19186171]

Lehky SR, Kiani R, Esteky H, Tanaka K. Statistics of visual responses in primate inferotemporal cortex to object stimuli. J. Neurophysiol. 2011; 106:1097–1117. [PubMed: 21562200]

Li L, Miller EK, Desimone R. The representation of stimulus familiarity in anterior inferior temporal cortex. Journal of neurophysiology. 1993; 69(6):1918–1929. [PubMed: 8350131]

Lim S, McKee JL, Woloszyn L, Amit Y, Freedman DJ, Sheinberg DL, Brunel N. Inferring learning rules from distributions of firing rates in cortical neurons. Nature neuroscience. 2015

Litwin-Kumar A, Doiron B. Formation and maintenance of neuronal assemblies through synaptic plasticity. Nat Commun. 2014; 5:5319. [PubMed: 25395015]

Liu D, Gu X, Zhu J, Zhang X, Han Z, Yan W, Cheng Q, Hao J, Fan H, Hou R, Chen Z, Chen Y, Li CT. Medial prefrontal activity during delay period contributes to learning of a working memory task. Science. 2014; 346:458–463. [PubMed: 25342800]

Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. Current Biology. 1995; 5(5):552–563. [PubMed: 7583105]

Lundqvist M, Compte A, Lansner A. Bistable, irregular firing and population oscillations in a modular attractor memory network. PLoS Comput. Biol. 2010; 6:e1000803. [PubMed: 20532199]

Markram H, Lubke J, Frotscher M, Roth A, Sakmann B. Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. J. Physiol. (London). 1997; 500:409–440. [PubMed: 9147328]

McCormick DA, Connors BW, Lighthall JW, Prince DA. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. Journal of neurophysiology. 1985; 54(4):782–806. [PubMed: 2999347]

Mézard M, Nadal J-P, Toulouse G. Solvable models of working memories. J. Physique. 1986; 47:1457–.

Miyashita Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature. 1988; 335(6193):817–820. [PubMed: 3185711]

Mongillo G, Barak O, Tsodyks M. Synaptic theory of working memory. Science. 2008; 319:1543. [PubMed: 18339943]

Mongillo G, Hansel D, van Vreeswijk C. Bistability and spatiotemporal irregularity in neuronal networks with nonlinear synaptic transmission. Phys. Rev. Lett. 2012; 108:158101. [PubMed: 22587287]

Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang XJ. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. U.S.A. 2017; 114:394–399. [PubMed: 28028221]

Nakamura K, Kubota K. Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. Journal of neurophysiology. 1995; 74(1):162–178. [PubMed: 7472321]

Ngezahayo A, Schachner M, Artola A. Synaptic activity modulates the induction of bidirectional synaptic changes in adult mouse hippocampus. J. Neurosci. 2000; 20:2451–2458. [PubMed: 10729325]

Ostojic S. Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons. Nature neuroscience. 2014; 17(4):594–600. [PubMed: 24561997]

Parisi G. A memory which forgets. Journal of Physics A: Mathematical and General. 1986; 19(10):L617.

Pfister J, Gerstner W. Triplets of spikes in a model of spike timing-dependent plasticity. J. Neurosci. 2006; 26:9673–9682. [PubMed: 16988038]

Rauch A, Camera GL, Lüscher H-R, Senn W, Fusi S. Neocortical pyramidal cells respond as integrate-and-fire neurons to *in vivo*-like input currents. J. Neurophysiol. 2003; 90:1598–1612. [PubMed: 12750422]

Reinhold K, Lien AD, Scanziani M. Distinct recurrent versus afferent dynamics in cortical visual processing. Nat. Neurosci. 2015; 18:1789–1797. [PubMed: 26502263]

Romo R, Brody CD, Hernández A, Lemus L. Neuronal correlates of parametric working memory in the prefrontal cortex. Nature. 1999; 399:470–474. [PubMed: 10365959]

Roxin A, Brunel N, Hansel D, Mongillo G, van Vreeswijk C. On the distribution of firing rates in networks of cortical neurons. Journal of Neuroscience. 2011; 31(45):16217–16226. [PubMed: 22072673]

Sejnowski TJ. Storing covariance with nonlinearly interacting neurons. Journal of mathematical biology. 1977; 4(4):303–321. [PubMed: 925522]

Senn W, Markram H, Tsodyks M. An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing. Neural Comput. 2001; 13:35–67. [PubMed: 11177427]

Shafi M, Zhou Y, Quintana J, Chow C, Fuster J, Bodner M. Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience. 2007; 146(3):1082–1108. [PubMed: 17418956]

Sjöström PJ, Turrigiano GG, Nelson SB. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron. 2001; 32(6):1149–1164. [PubMed: 11754844]

Sompolinsky H, Crisanti A, Sommers H-J. Chaos in random neural networks. Physical Review Letters. 1988; 61(3):259. [PubMed: 10039285]

Tirozzi B, Tsodyks M. Chaos in highly diluted neural networks. EPL (Europhysics Letters). 1991; 14(8):727.

Toyoizumi T, Kaneko M, Stryker MP, Miller KD. Modeling the dynamic interaction of hebbian and homeostatic plasticity. Neuron. 2014; 84(2):497–510. [PubMed: 25374364]

Treves A. Graded-response neurons and information encodings in autoassociative memories. Physical Review A. 1990a; 42(4):2418.

Treves A. Threshold-linear formal neurons in auto-associative nets. Journal of Physics A: Mathematical and General. 1990b; 23(12):2631.

Treves A. Mean-field analysis of neuronal spike dynamics. Network: Computation in Neural Systems. 1993; 4(3):259–284.

Tsodyks M. Associative memory in asymmetric diluted network with low level of activity. EPL (Europhysics Letters). 1988; 7(3):203.

Tsodyks M, Feigel'Man M. The enhanced storage capacity in neural networks with low activity level. EPL (Europhysics Letters). 1988; 6(2):101.

Vogels T, Sprekeler H, Zenke F, Clopath C, Gerstner W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. Science. 2011; 334(6062):1569–1573. [PubMed: 22075724]

Wang X-J. Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 2001; 24:455–463. [PubMed: 11476885]

Woloszyn L, Sheinberg DL. Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. Neuron. 2012; 74(1):193–205. [PubMed: 22500640]

Zenke F, Agnes EJ, Gerstner W. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. Nat Commun. 2015; 6:6922. [PubMed: 25897632]

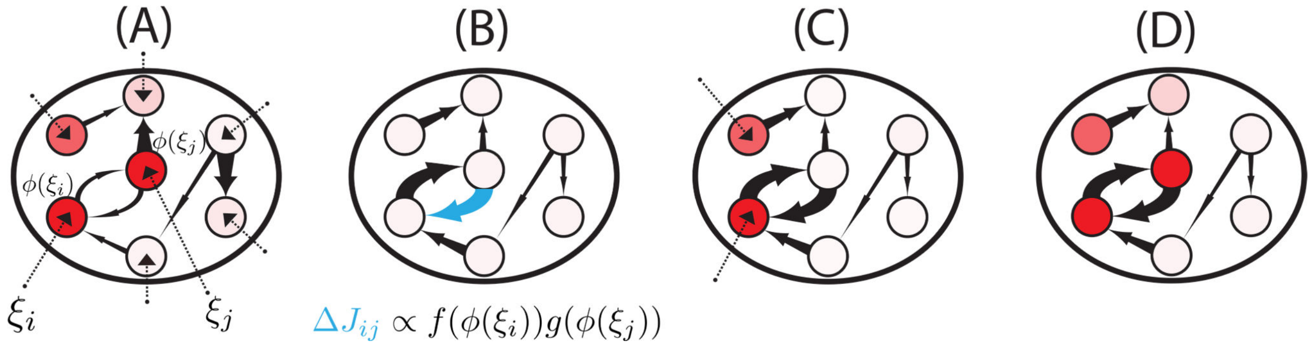

Figure 1.

Learning and retrieval in recurrent neural networks with unsupervised Hebbian learning rules. (**A**) When a novel pattern is presented to the network, synaptic inputs to each neuron in the network (ξ_I, for neurons $I = 1, \ldots, N$) are drawn randomly and independently from a Gaussian distribution. Synaptic inputs elicit firing rates through the static transfer function, i.e. $\phi(\xi_I)$. Some neurons respond strongly (red circles), others weakly (white circles). (**B**) The firing rate pattern produced by the synaptic input currents modifies the network connectivity according to an unsupervised Hebbian learning rule. The connection strength is represented by the thickness of the corresponding arrow (the thicker the arrow the stronger the connection). (**C**) After learning, a pattern of synaptic inputs that is correlated but not identical to the stored pattern is presented to the network. (**D**) Following the presentation, the network goes to an attractor state which strongly overlaps with the stored pattern (compare with panel A), which indicates the retrieval of the corresponding memory.

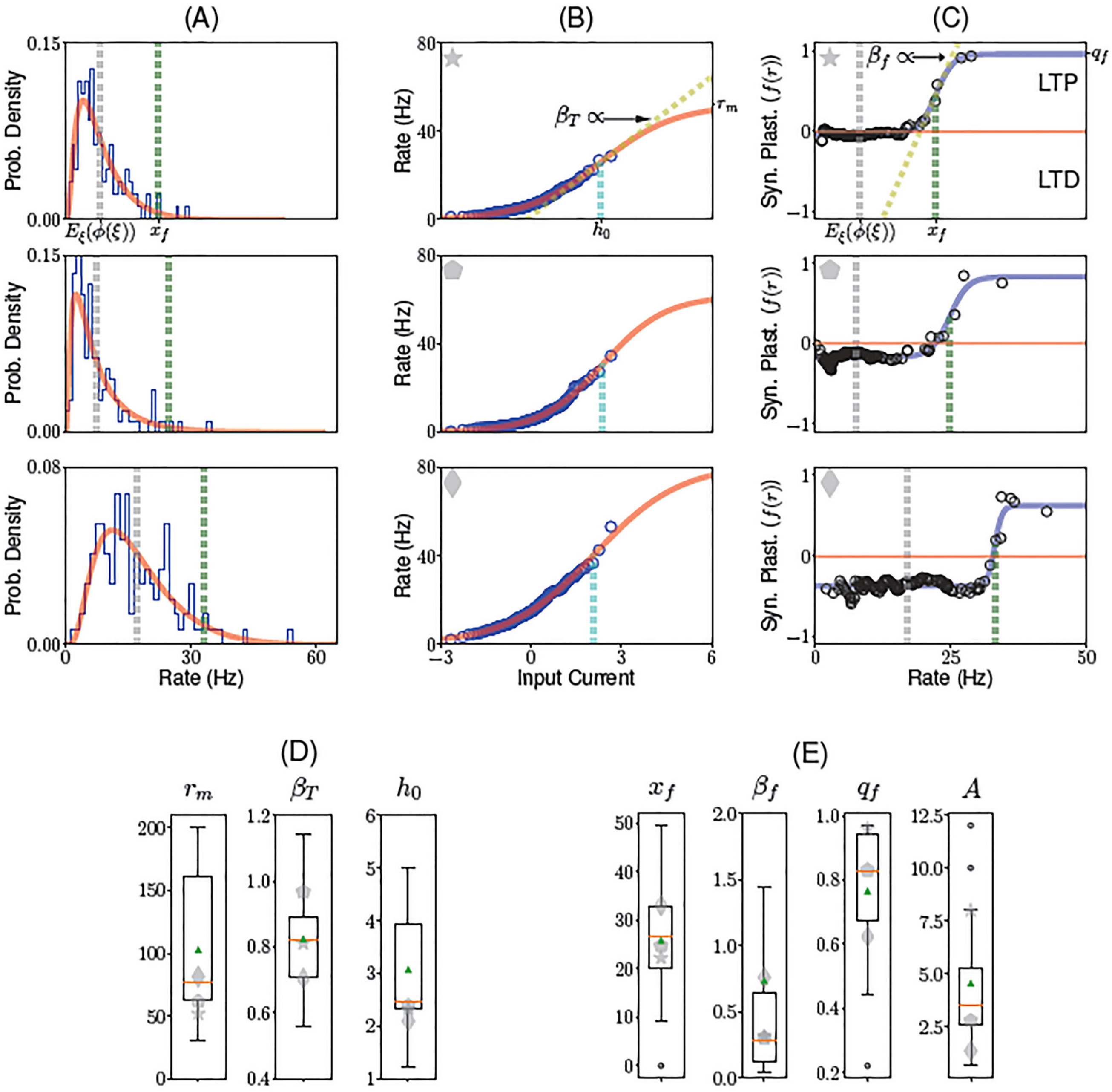

Figure 2.

Inferring transfer function and learning rule from ITC data. (**A**) Distributions of firing rates in response to novel stimuli, for three different ITC neurons. Blue histogram: histogram of experimentally recorded visual responses. Red: Distribution of firing rates obtained from passing a standard normal distribution through the sigmoidal transfer function shown in B. Gray vertical line: average firing rate. Green vertical line: learning rule threshold x_f (see C). (**B**) Static transfer function ϕ derived from the distribution of visual responses for novel stimuli (see A), assuming a Gaussian distribution of inputs (see (Lim et al., 2015) and Methods) for the same three neurons shown in A. The data (blue circles) was fitted using a sigmoidal function (red line; see Methods, Eq. (17)), defined by three parameters: the current h_0 that leads to half the maximal firing rate (cyan dashed lines), a slope parameter β_T (dashed yellow line in top plot), and maximal firing rate r_m. (**C**) Dependence of the synaptic plasticity rule on the postsynaptic firing rate as a function of firing rate (i.e. $f(r)$). The data (black circles) was fitted with a sigmoidal function (blue line; see Methods, Eq. (22)), defined by three parameters: maximum potentiation q_f; threshold x_f (see green dashed line); and slope parameter β_f (dashed yellow line in top plot). On the right axis is indicated the maximum potentiation of the fit q_f. (**D**) Boxplot for the fitted parameters r_m, β_T and h_0 of the transfer function. (**E**) Boxplot for the fitted parameters x_f, β_f, q_f of the dependence of the synaptic plasticity rule on the postsynaptic firing rate, and A, the learning rate. The red line and green triangle indicate the median and the mean of the fitted parameters, respectively. Gray symbols indicate the parameters of the three neurons shown in A,B,C.

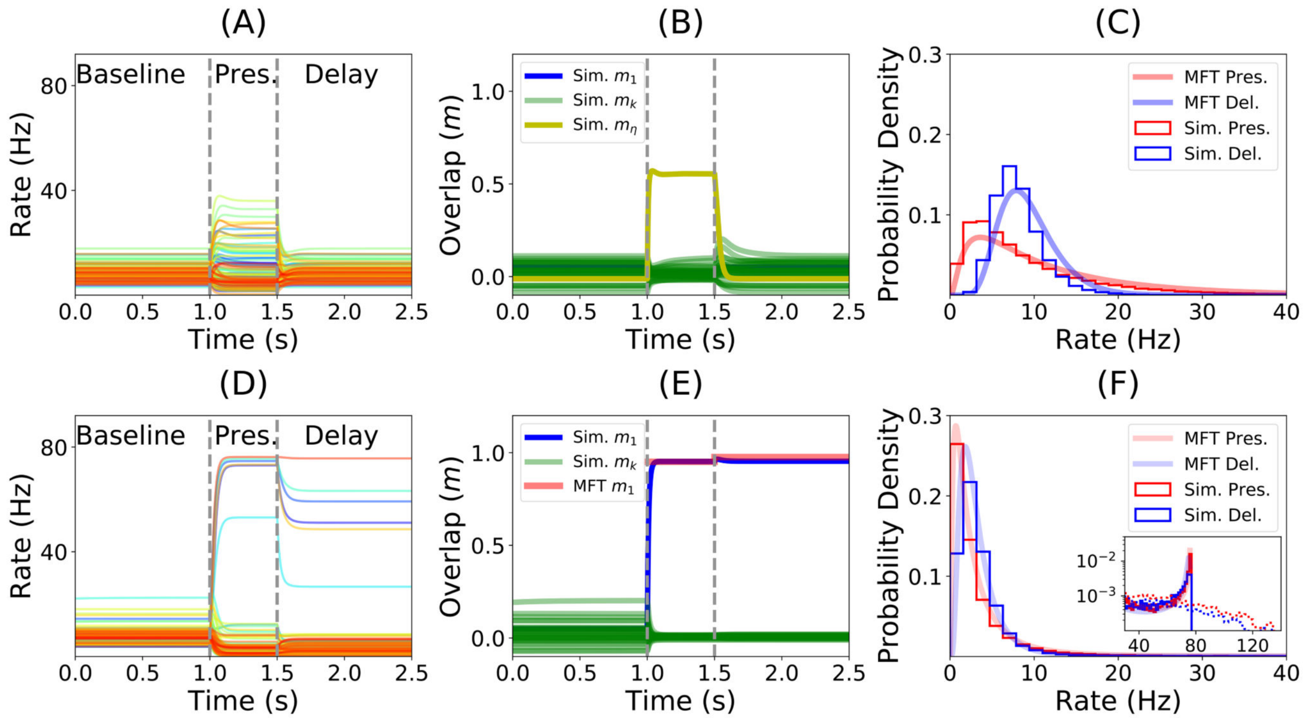

Figure 3.

Dynamics of the network before, during and after the presentation of novel (top row) and familiar (bottom row) stimuli, mimicking the initial part of a trial of a delay match to sample (DMS) experiment. (**A**) Firing rate of a randomly sampled subset of 100 neurons of a simulated network before, during and after the presentation of a novel stimulus. Vertical dashed lines indicate the beginning and the end of the presentation. Note that the firing rates of all neurons decay to baseline following removal of the stimulus. (**B**) Dynamics of the overlaps with the stored patterns. Green traces show overlaps computed numerically from the network simulation corresponding to each of the stored patterns. The yellow trace shows the overlap of the network state with the shown novel pattern. (**C**) Distribution of firing rates during the presentation (red) and delay (blue) periods. Smooth curves correspond to the predictions of the MFT, histograms are obtained from network simulations. (**D**) Similar to A, except that the shown stimulus is familiar. Note that this time firing rates do not decay to baseline during the delay period, but to a value that is strongly correlated (but not identical) to the visual response. (**E**) Dynamics of overlaps when a familiar stimulus is presented. The blue trace shows the numerically computed overlap with the pattern presented during the presentation period. The red trace shows the corresponding overlap computed from MFT. (**F**) Distribution of firing rates during the presentation (red) and delay (blue) periods in response to the presentation of a familiar stimulus. The vast majority of the neurons fire in the 0–10Hz range. A closer inspection of the tail of the distribution shows a tiny peak close to saturation in homogeneous networks (full lines), while this peak disappears when the heterogeneity in maximal firing rates is included (dashed lines).

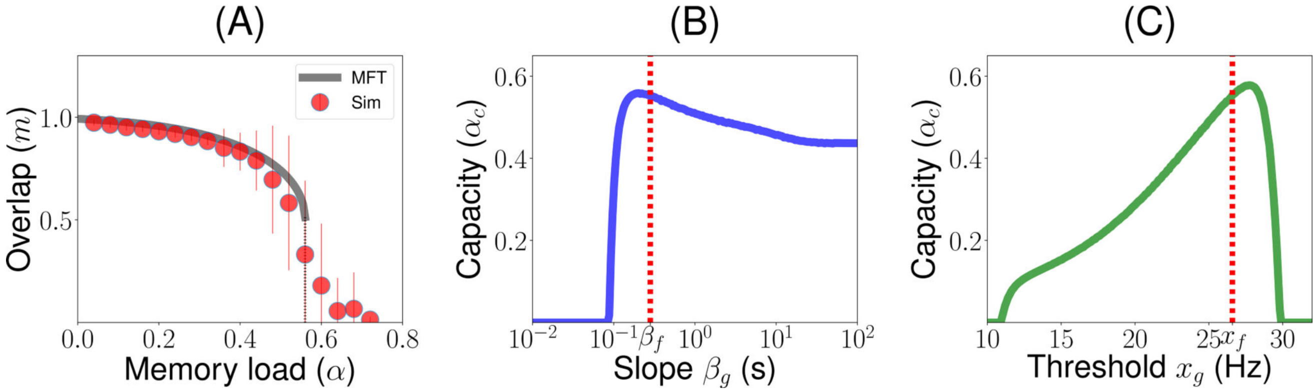

Figure 4.

Storage capacity of the network, and its dependence on g. (**A**) Overlap as a function of memory load α (number of patterns stored divided by average number of connections per neuron). Grey: MFT. Red circles: Numerical simulations (average and standard deviations computed from 100 realizations with $N = 5 \cdot 10^4$). The overlap stays positive until $\alpha \sim 0.56$. Parameters of g are chosen to be identical to those of f. (**B**) Capacity vs β_g. The capacity is maximized for $\beta_g \sim \beta_f$ (dashed red line $\beta_g = \beta_f$). (**C**) Capacity vs x_g. The capacity is close to being maximized for $x_f \sim x_g$ (dashed red line $x_g = x_f$). Other parameters as in Fig. 3.

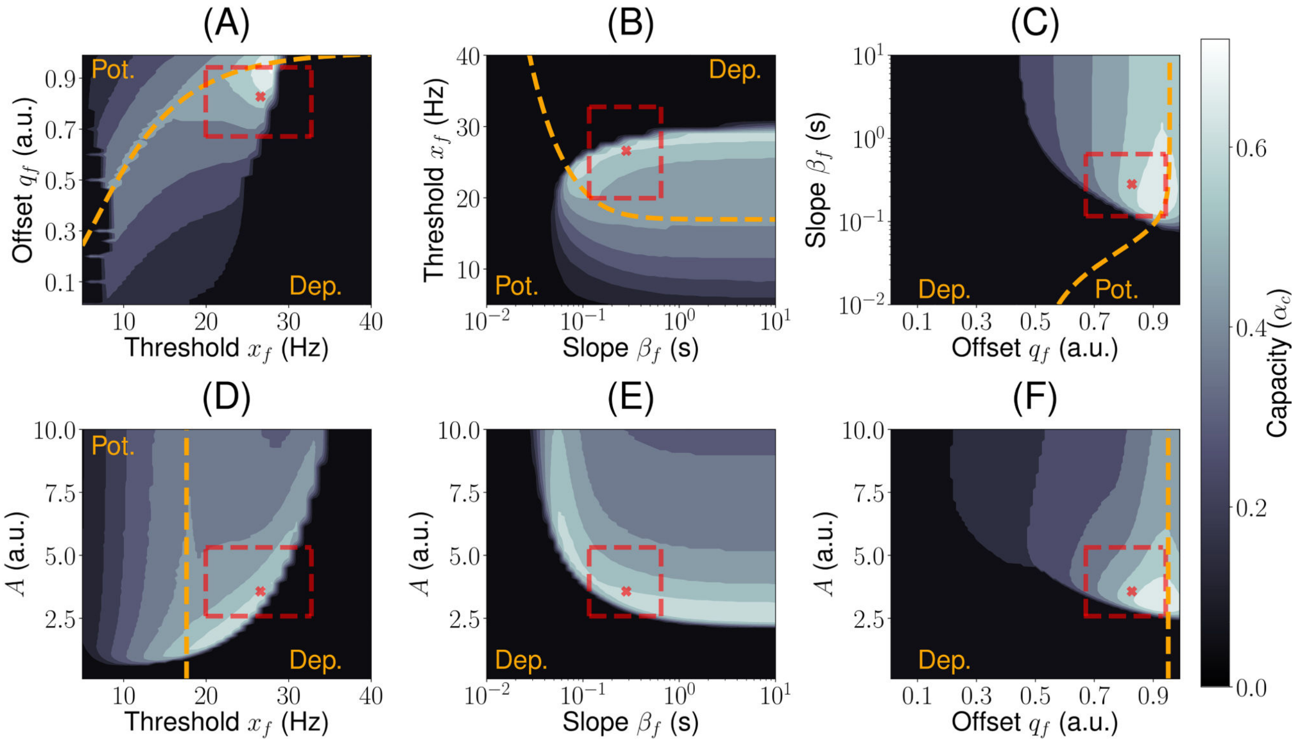

Figure 5.

Inferred learning rules from ITC are close to maximizing memory storage. Contour plots for the capacity of the network as a function of two parameters. In each plot, two parameters are set to the median best-fit parameters, and the other two are varied. The yellow dashed line indicates the curve where potentiation and depression are balanced in average (i.e. $\int \mathscr{D}\xi f(\phi(\xi)) = 0$). It separates the potentiation (i.e. $\int \mathscr{D}\xi f(\phi(\xi)) > 0$) and depression (i.e. $\int \mathscr{D}\xi f(\phi(\xi)) < 0$) regions. The parameter region corresponding to the interquartile range is indicated with a red dashed rectangle. The median best-fit parameters are shown as a red cross mark. The parameters of g: $x_g = x_f$ and $\beta_g = \beta_f$.

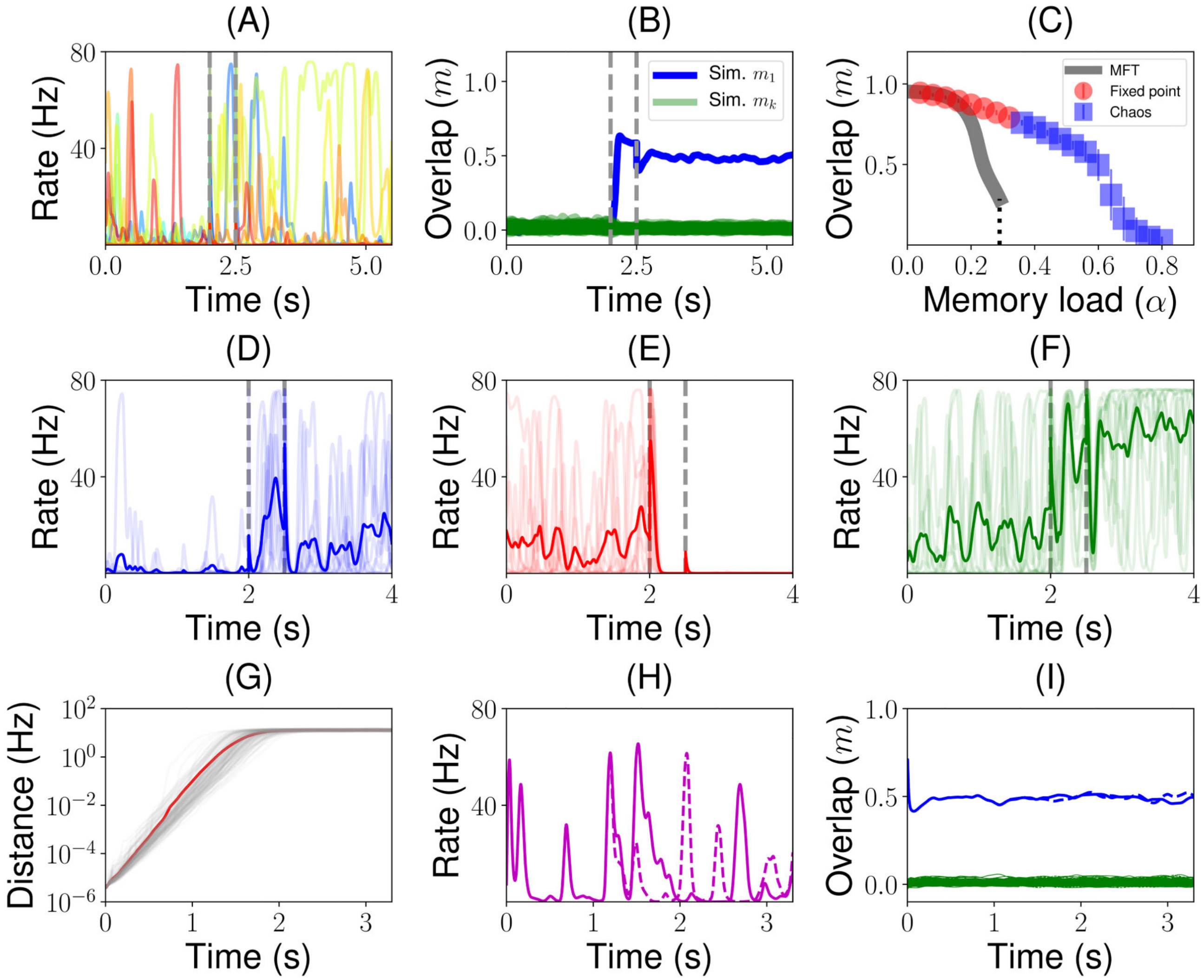

Figure 6.

Chaotic background and retrieval states, for a network with parameters as in Fig. 3, except for the learning rate (A = 10.65) and memory load (α = 0.48 in all panels except in C). (**A**) Firing rate dynamics for a randomly sampled subset of 10 neurons of a simulated network when a familiar stimulus (i.e. one of the stored patterns) is presented. (**B**) Dynamics of the overlaps before, during and after the presentation of a familiar stimulus. Green traces shown all the overlaps computed numerically from the network simulation corresponding to each of the stored patterns except the one with the presented pattern, shown in blue. (**C**) Overlap vs memory load. Gray curve: MFT. Red circles: simulations in which the dynamics converge to fixed point attractors. Blue square: simulations in which the dynamics converge to chaotic states. (**D–F**) Dynamics of the firing rate of three example neurons in 10 different trials (random initial conditions - transparent traces). Trial-averaged firing rate (over 20 trials) is shown with an opaque trace. (**G**) Light gray traces: exponential initial growth followed by saturation of the distance between pairs of retrieval states corresponding to the same stored pattern but slightly different initial conditions (see Methods). Red curve: average distance between pairs of retrieval states with slightly different initial conditions. (**H**) Firing rate of a single neuron starting from two slightly different initial conditions (continuous vs dashed). (**I**) Overlaps with the retrieved pattern (blue) and all other stored patterns (green) again for a pair of initial conditions (continuous vs dashed). As in Fig. 3, in A, B and D–F vertical dashed lines indicate the beginning and the end of the presentation period.

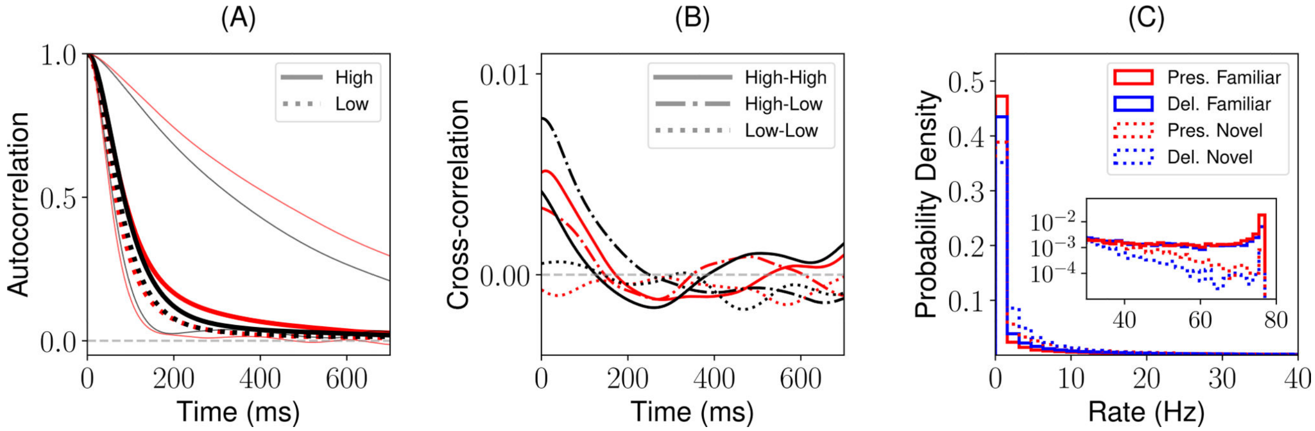

Figure 7.

Statistical properties of the chaotic background and retrieval states, for a network with parameters as in Fig 6. (A) Red: background state. Black: retrieval state. Thick traces: mean autocorrelation (AC) functions across 100 randomly sampled neurons with mean firing rate between 1Hz and half of the maximal firing rate (low mean firing rates; dashed) and between half of the maximal firing rate and 65Hz (high mean firing rates; solid). Light traces: AC function for neurons with the fastest and slowest decays, showing a broad range of individual AC timescales. (B) Mean cross-correlation (CC) functions across 200 randomly chosen pairs of neurons with high (i.e. high-high), low (i.e. low-low) and with one neuron high and the other low (i.e. high-low) mean firing rates. Same color code than panel A. (C) Distribution of mean firing rates during the presentation (red) and delay (blue) periods for novel (dashed) and familiar (solid) stimuli.

Key Resource Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Network simulations and algorithms for solving the mean field equations and computing the capacity of the network.	This paper	https://github.com/ulisespereira/AttractorDynamics
Other		
Data from electrophysiological recordings in ITC of behaving primates.	Woloszyn and Sheinberg, 2012; Lim et al., 2015	N/A