

# Short term memory properties of sensory neural architectures

Alexis M. Dubreuil

Group for Neural Theory, Laboratoire de Neurosciences Cognitives Computationnelles, Département d'Étude Cognitive, Ecole Normale Supérieure, INSERM U960, PSL University, Paris, France

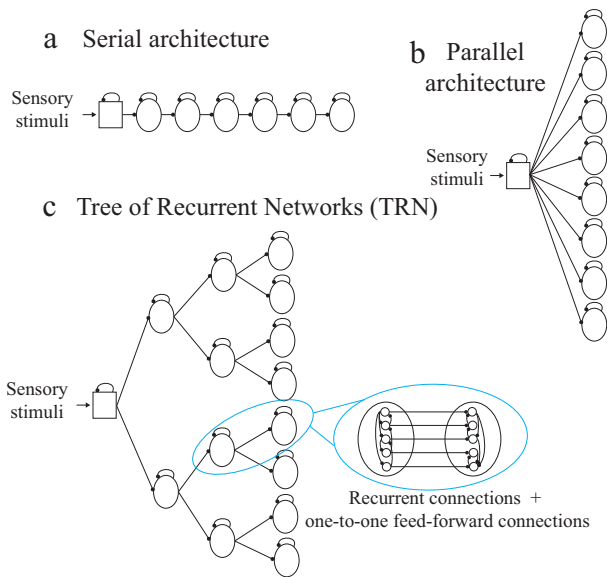
**Abstract** A functional role of the cerebral cortex is to form and hold representations of the sensory world for behavioral purposes. This is achieved by a sheet of neurons, organized in modules called cortical columns, that receives inputs in a peculiar manner, with only a few neurons driven by sensory inputs through thalamic projections, and a vast majority of neurons receiving mainly cortical inputs. How should cortical modules be organized, with respect to sensory inputs, in order for the cortex to efficiently hold sensory representations in memory? To address this question we investigate the memory performance of trees of recurrent networks (TRN) that are composed of recurrent networks, modeling cortical columns, connected with each others through a tree-shaped feed-forward backbone of connections, with sensory stimuli injected at the root of the tree. On these sensory architectures two types of short-term memory (STM) mechanisms can be implemented, STM via transient dynamics on the feed-forward tree, and STM via reverberating activity on the recurrent connectivity inside modules. We derive equations describing the dynamics of such networks, which allow us to thoroughly explore the space of possible architectures and quantify their memory performance. By varying the divergence ratio of the tree, we show that serial architectures, where sensory inputs are successively processed in different modules, are better suited to implement STM via transient dynamics, while parallel architectures, where sensory inputs are simultaneously processed by all modules, are better suited to implement STM via reverberating dynamics.

**Keywords** Modular network · Short-term memory · Cortical hierarchy

## 1 Introduction

Neural activity in the neocortex forms sensory representations of ongoing stimuli that can be held in memory for behavioral purposes (Fuster, 1995). Such operations are performed by cortical circuits which pos-

sess a peculiar organization, as described by anatomical studies. For instance cortical connectivity has been shown to be spatially clustered, which allows to conceptually segregate the cortical sheet into groups of neurons called cortical columns (Bosking et al., 1997; DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Pucak et al., 1996). Regarding the transmission of sensory information to cortical circuits, anatomical studies have suggested that a small subset of neurons are directly driven by thalamic projections, while a vast majority are in charge of processing these stimuli (Braitenberg and Schütz, 1991). This vast majority of neurons are sometimes called interneurons as they are not directly under the influence of sensory inputs, nor directly involved in generating a cortical output. In this paper we investigate how the organization of these interneurons, with respect to sensory inputs, impact the memory performance of a cognitive architecture. Two extreme types of organization can be considered. A serial organization, where a sensory input is successively passed to different groups of neurons, with one group operating on the output of the previous one (Fig. 1a). A parallel organization, where a sensory input is simultaneously sent to all groups of neurons, where it is processed independently from one group to another (Fig. 1b). The neural architectures we consider range from serial to parallel, and consist of modules, made out of recurrent neural networks modeling cortical columns, that for simplicity we consider connected in a feed-forward manner. These neural architectures implement two kinds of short-term memory mechanisms. First, the recurrent connectivity inside each module creates feed-back loops in which neural activity can reverberate leading to sustained patterns of activity coding for specific stimuli, i.e. modules are attractor neural networks (Amit, 1989; Amit and Brunel, 1997; Hopfield, 1982; Tsodyks and Feigel'man, 1988). Second, the feed-forward connectivity between modules allows to implement delay-lines on which sequences of stimuli propagate. Delay-lines have been shown to be efficient neural architectures to hold sensory representations in the transient dynamics of a neural network (Ganguli et al., 2008; Jaeger and Haas, 2004; Lim and Goldman, 2012; Maass et al., 2002; White et al., 2004). The advantage of the attractor neural network scenario is that memory can be held on a very long timescale. A drawback is that only a limited number of already learnt sensory representations can be maintained, since maintenance of a representation relies on a specific set of synaptic connections. We will refer to this number as the *storage capacity* of the architecture. On the opposite, short-term memory via transient dynamics allows to memorize arbitrary stimuli, but the timescale on which these



**Figure 1** Trees of recurrent networks (TRN) **a** Purely serial architecture, obtained as a special case of TRN with divergence ratio  $d = 1$ . The root module (squared box), models the thalamus conveying sensory inputs to modules modeling cortical columns. Sensory inputs propagate from left to right through the feed-forward connections between modules. **b** Purely parallel architecture, obtained as a special case of TRN with divergence ratio  $d = M$ , where  $M$  is the number of modules composing the network. **c** TRN with divergence ratio  $d = 2$ ,  $M = 14$  and  $L = 3$ . Each module is an attractor network of  $N$  neurons with recurrent connectivity storing  $\alpha N$  patterns of activity as fixed points of the module's dynamics. Connections between a module and each of its  $d$  children are one-to-one, i.e. a neuron of the parent module projects to a single neurons in each child module and each neuron of a child module receives a single projection.

stimuli can be maintained is limited. We will refer to this time scale as the *buffering capacity* of the network.

In the first section of this paper we formally define the class of neural architectures called Tree of Recurrent Networks (TRN) that we study. In order to reach a full comprehension of the memory properties of these architectures we take the following incremental steps. We first study the storage capacity of an isolated module, as well as the dynamics of memory retrieval (second section). We then focus on a single feed-forward path of a TRN, showing in particular how the recurrent connections inside modules degrade stimuli as they propagate throughout the path. An important phenomenon that eventually limits the storage and buffering capacity of the path (third section). We finally examine how these two quantities behave in a full TRN, in particular we compare the memory properties of the extreme cases that are the fully serial and fully parallel architectures (fourth section). To conclude we summarize our results and discuss how TRN can be mapped onto cortical sensory hierarchies.

## 2 Definition of the model

We consider feed-forward trees of  $M$  modules (Fig. 1) such that for each step up the tree, the number of modules increases by a factor  $d$  where  $d$  labels the divergence ratio of the TRN. The divergence ratio ranges from  $d = 1$ , defining a purely serial architecture composed of a single path of length  $L = M$  (Fig. 1a), to  $d = M$ , defining a purely parallel architecture composed of  $M$  paths of length  $L = 1$  (Fig. 1b). For intermediate divergence ratios, the length of the paths is related to  $M$  and  $d$  as detailed in section 5, and illustrated in Fig. 1c.

Each module operates as a fixed-points attractor neural network in which independent memories are stored in the recurrent connections  $J_{ij}^m$  ( $m = 1, \dots, M$  and  $i, j = 1, \dots, N$ , where  $N$  is the number of neurons in each module).

### 2.1 Dynamics of the network

Neuron  $i$  in module  $m$  is described by a binary variable  $\sigma_i^m(t)$  whose dynamics is:

$$\sigma_i^m(t+1) = \Theta(h_i^m(t) - \theta) \quad (1)$$

where  $\Theta(\cdot)$  is the Heaviside function,  $\theta$  an activation threshold and  $h_i^m$  is the total synaptic input on the neuron, which is the sum of a feedforward input, coming from a single neuron of the previous module, and a recurrent input:

$$h_i^m(t) = \sigma_i^{a(m)}(t) + \sum_{j=1}^N J_{ij}^m \sigma_j^m(t) \quad (2)$$

where module  $a(m)$  is the module preceding  $m$  in the feed-forward structure.

### 2.2 Recurrent connections

The recurrent connections of a module store  $P$  random independent binary patterns,  $\{\xi^{\mu,m}\}$ ,  $\mu = 1, \dots, P$  with a coding level  $f$ ,

$$\xi_i^{\mu,m} = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{with probability } 1 - f \end{cases} \quad (3)$$

Patterns are stored using a covariance learning rule (Sejnowski, 1977; Tsodyks and Feigel'man, 1988) which implements feed-back loops allowing activity to reverberate such that the  $P$  patterns can become fixed points of the network dynamics

$$J_{ij}^m = \frac{1}{Nf(1-f)} \sum_{\mu=1}^P (\xi_i^{\mu,m} - f)(\xi_j^{\mu,m} - f) \quad (4)$$

We introduce the storage load  $\alpha = P/N$  which quantifies the number of memories stored in each module. Note that in what follows, the coding level is set to  $f = 0.01$  unless stated otherwise, we have checked that other choices of coding levels ( $f = 0.001$  or  $f = 0.1$ ) lead to qualitatively similar results.

### 2.3 Inputs to the network

At each time step, a new input is represented by the state of the root module  $\sigma^0$ . In order to probe short-term memory properties via transient dynamics and reverberating activity, we consider the TRN under two kinds of stimulation. In the first scenario, inputs are drawn independently of the stored patterns (see Fig. 4a). In the second scenario, the network receives a similar stream of inputs, but at  $t = 0$  the root module represents a pattern stored in module  $m_0$  (see Fig. 5a).

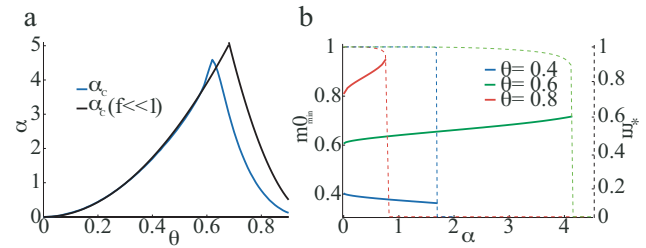
### 3 Storage capacity of a single module

The recurrent connectivity of each module is the substrate allowing TRN to hold stimuli in short-term memory via reverberating activity. With our choice of connectivity matrix (4), modules operate as attractor neural networks and  $\alpha N$  stimuli can be held in short-term memory under the form of persistent activity. We quantify the ability of a module to implement such short-term memory mechanism with the storage capacity (Amit, 1989). It is defined as the maximal storage load  $\alpha$  such that imprinted patterns are stable states of the network's dynamics. We also describe the basins of attraction associated to each stored memory.

#### 3.1 Dynamical equations for macroscopic variables

Without loss of generality we focus on the retrieval and maintenance of a specific memory, e.g. pattern  $\xi^1$ . The state of the module is described by the overlap  $m(t)$  between the instantaneous network state and the corresponding pattern

$$m(t) = \frac{1}{Nf(1-f)} \sum_i^N \langle (\xi_i^1 - f)\sigma_i(t) \rangle, \quad (5)$$



**Figure 2** Short-term memory via reverberating activity in a single module **a** The blue curve shows the storage capacity for  $f = 0.01$  obtained by iterating (7). The black curve shows the analytical approximation (9) obtained for  $f \ll 1$ . **b** Size of the basins of attraction associated to each stored patterns (full lines), together with the overlap  $m^*$  between the fixed point of the network and the memory pattern to be stored (dashed lines). We measure the size of the basins of attraction as the minimal initial overlap  $m_{0_{min}}$  that leads to pattern retrieval when solving (7) with  $\mu(t=0) = f$ .

and the average activity in the module,

$$\mu(t) = \frac{1}{N} \sum_i^N \langle \sigma_i(t) \rangle \quad (6)$$

where  $\langle \cdot \rangle$  corresponds to averaging over different microscopic initial states (that have the same macroscopic description). The dynamics of these order parameters can be derived using standard methods (see Supplementary materials, (Evans, 1989; Sompolinsky and White, 2003)):

$$\begin{aligned} m(t+1) &= H\left(\frac{\theta - (1-f)m(t)}{\sqrt{\alpha\mu(t)}}\right) - H\left(\frac{\theta + fm(t)}{\sqrt{\alpha\mu(t)}}\right) \\ \mu(t+1) &= fH\left(\frac{\theta - (1-f)m(t)}{\sqrt{\alpha\mu(t)}}\right) + \\ &\quad (1-f)H\left(\frac{\theta + fm(t)}{\sqrt{\alpha\mu(t)}}\right) \end{aligned} \quad (7)$$

where

$$H(x) = \int_x^{+\infty} dz \frac{e^{-z^2/2}}{\sqrt{2\pi}}. \quad (8)$$

#### 3.2 Stable states and storage capacity

We label  $(m^*, \mu^*)$  fixed point of (7) for the overlap and mean activity. A particular fixed point is the background state  $m^* = 0$  and  $\mu^* = 0$ . Retrieval states that have a high overlap with a stored pattern  $m^* \simeq 1$  exist as long as the storage load remains below a critical value  $\alpha < \alpha_c$ . This critical value is shown by the blue curve in Fig. 2a that we obtain by numerically finding the fixed points of equation (7) for  $f = 0.01$ . The retrieval states have been described previously for  $f \ll 1$  (Tsodyks

and Feigel'man, 1988; White et al., 2004). In this regime the storage capacity  $\alpha_c$  can be approximated by

$$\alpha_c \underset{f \ll 1}{=} \min \left[ \frac{\theta^2}{2f|\log f|}, \frac{1}{2f}(1-\theta)^2 \right] \quad (9)$$

The curve  $\alpha_c(f \ll 1)$  vs  $\theta$  is shown in black in Fig. 2a. It is composed of two quadratic branches in  $\theta$ . On the first one, retrieval states are destabilized by an increase in the activity of background neurons (neurons  $i$  such that  $\xi_i^1 = 0$ ), while on the second branch, retrieval states are destabilized by a silencing of foreground neurons (neurons  $i$  such that  $\xi_i^1 = 1$ ).

By solving equations (7) numerically, we have also found an 'active' background state ( $m^* = 0, \mu^* \gg f$ ) and 'weak' retrieval states characterized by ( $m^* = O(1), \mu^* \gg f$ ). We study such states in Supplementary materials. Note that these states exist at low activation thresholds, which are sub-optimal for TRN memory properties, as will be shown in the following.

### 3.3 Basins of attraction

The state in which the network settles depends on the two initial values  $m_0 = m(t=0)$  and  $\mu_0 = \mu(t=0)$ . We measure the size of the basins of attraction by  $m0_{min}$  the minimal value of the overlap that is required to reach a retrieval state for an initial mean activity  $\mu_0 = f$ . In Fig. 2b we show  $m0_{min}$  as a function of  $\alpha$  for  $f = 0.01$  and  $\theta = 0.4, 0.6$  and  $0.8$ . For a given value of  $\theta$ ,  $m0_{min}$  is only weakly modulated by the storage load, with  $m0_{min} \simeq \theta$ . This can be seen from equations (7), which give, for small  $f$ ,

$$m(t+1) \simeq H \left( \frac{\theta - m(t)}{\sqrt{\alpha\mu(t)}} \right) \quad (10)$$

When  $\alpha\mu(t) \ll 1$  we have

$$m(t+1) = \begin{cases} 1 & \text{if } m(t) > \theta \\ 0 & \text{if } m(t) < \theta \end{cases} \quad (11)$$

## 4 Memory properties of a single path

After isolating a single module from a TRN, we isolate a single feed-forward path composed of  $L$  modules, from the root module up to one of the leaves of the tree. Stimuli propagate along the feed-forward structure from the root module. We first describe the propagation of stimuli orthogonal to the stored patterns. This allows us to measure the buffering capacity, which quantifies the ability of the neural activity of the path

to represent a sequence of previously observed stimuli (Fig. 4a). We then consider a stimulus that is stored in one of the module, and describe under which conditions it can trigger a state of persistent activity in this module, as illustrated in Fig. 5a. This allows us to quantify the storage capacity of a path of length  $L$ .

### 4.1 Dynamical equations for propagation of activity patterns

Here modules form a single feed-forward path and are indexed with  $l = 1 \dots L$ . A stream of sensory stimuli imposes the state of the root module  $\sigma^0$  at all times with

$$\sigma_i^0(t) = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{with probability } 1 - f \end{cases} \quad (12)$$

An input represented in the root module at time  $t-l$  reaches module  $l$  at time  $t$ . The dynamics of the path is characterized by  $\{\mu^l(t)\}_{l=1 \dots L}$  the mean activities in the different modules, and  $\{n^l(t)\}_{l=1 \dots L}$  the overlaps between  $\sigma^0(t-l)$  and  $\sigma^l(t)$

$$n^l(t) = \frac{1}{Nf(1-f)} \sum_i^N \langle (\sigma_i^0(t-l) - f)\sigma_i^l(t) \rangle \quad (13)$$

The time evolution of these macroscopic variables is governed by (see Supplementary materials for derivation)

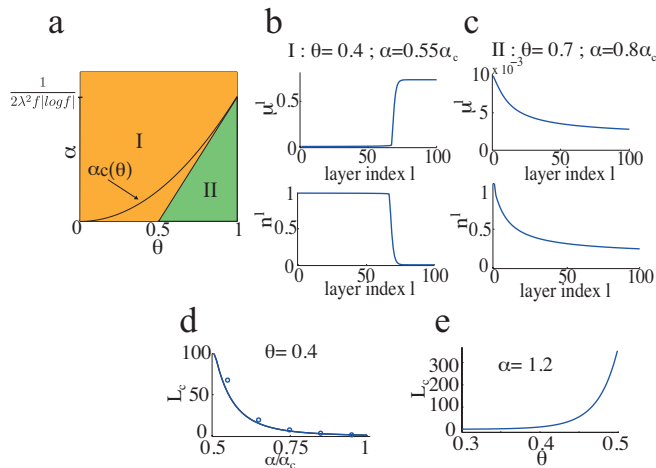
$$n^l(t+1) = n^{l-1}(t) \left[ H \left( \frac{\theta - 1}{\sqrt{\alpha\mu^l(t)}} \right) - H \left( \frac{\theta}{\sqrt{\alpha\mu^l(t)}} \right) \right] \quad (14)$$

together with

$$\mu^l(t+1) = \mu^{l-1}(t) \left[ H \left( \frac{\theta - 1}{\sqrt{\alpha\mu^l(t)}} \right) - H \left( \frac{\theta}{\sqrt{\alpha\mu^l(t)}} \right) \right] + H \left( \frac{\theta}{\sqrt{\alpha\mu^l(t)}} \right) \quad (15)$$

### 4.2 Mean activity in steady state

To understand how stimuli propagate throughout the path, it is informative to first examine the profile of mean activity the path settles into under the influence of the stream of stimuli. At the macroscopic level this profile is given by the steady states of (14)(15).



**Figure 3** Propagation of stimuli throughout the path. **a** Mean activity in a path. Regions in the  $\theta - \alpha$  plane where the different regimes are found for  $f \ll 1$ . In regime I, the orange region above the oblique line,  $\mu^l$  is increasing with  $l$ . In regime II, green region below the oblique line,  $\mu^l$  is decreasing with  $l$ . The parabolic line represents the storage capacity  $\alpha_c$ . **b** Top: average activities in steady state for a path of  $L = 100$  modules. Bottom: overlap  $n^l$  between a pattern sent at time  $t$  and the state of module  $l$  at time  $t + l$ . **c** Same as b for regime II. **d**  $L_c$  as a function of the normalized storage load  $\frac{\alpha}{\alpha_c}$  for  $\theta = 0.4$ . **e**  $L_c$  as a function of  $\theta$  for  $\alpha = 1.2$ .

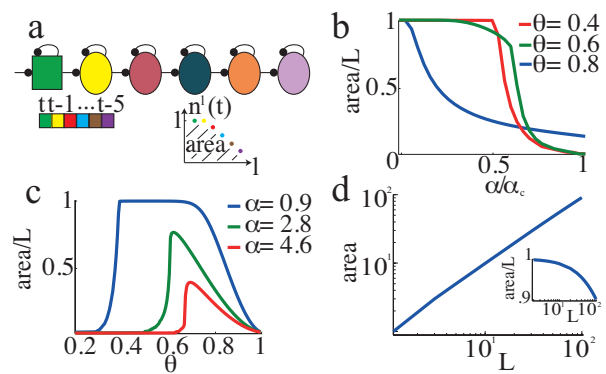
At the microscopic level a neuron  $i$  in module  $l$  is active if

$$1 + Q_i^l > \theta \quad \text{if neuron } i \text{ of module } l - 1 \text{ is active} \quad (16)$$

$$Q_i^l > \theta \quad \text{if neuron } i \text{ of module } l - 1 \text{ is silent} \quad (17)$$

where  $Q_i^l$  corresponds to the recurrent noise coming from the storage of patterns in module  $l$ .  $Q_i^l$  is drawn from a normal distribution  $\mathcal{N}(0, \sqrt{\alpha\mu^l})$ . Depending on the value of  $\theta$  and the standard deviation  $\sqrt{\alpha\mu^l}$ , two behaviors are possible, either the average activity in each module,  $\mu^l$ , increases along the path (regime I), either it decreases (regime II).

In the first regime the increase is characterized by a critical depth  $L_c$  from which activity explodes in the network:  $\mu^{l < L_c} = O(f)$  and  $\mu^{l > L_c} \gg f$ . The  $\alpha - \theta$  region of parameters in which this behavior is observed is shown in orange in Fig. 3a. The frontiers of this region can be derived analytically for small coding levels  $f \ll 1$  (see Appendix 7.1). It corresponds to low activation thresholds ( $\theta < \frac{1}{2}$ ) or high activation threshold paired with high recurrent noise ( $\theta > \frac{1}{2}$  and  $\alpha > (2\theta^{-1} - \theta^{-2})\alpha_c$ ). An example of profile of average activity  $\mu^l, l = 1 \dots L$  is shown in the top panel of Fig. 3b. Values of  $L_c$  can also be derived analytically (see Appendix 7.1) and its dependence on  $\alpha$  and  $\theta$  are shown in Fig. 3d. As the storage load  $\alpha$  is increased the recurrent noise becomes more effective at activating neu-



**Figure 4** Buffering of arbitrary input sequences. **a** Illustration of a path buffering the sequence of stimuli: 'magenta,brown,blue,red,yellow,green', with the 'green' stimulus presented at time  $t$  in the root module. The buffering capacity of a path of length  $L$  is measured by the area under the curve  $n^l$ , measuring the overlap between the activities in individual modules and patterns of activity previously presented in the root module. **b** Area as a function of the amount of recurrent connections  $\alpha/\alpha_c(\theta)$  for  $L = 100$ . **c** Normalized buffering capacity as a function of  $\theta$ ,  $L = 100$ . **d** Buffering capacity as a function of path length for  $\theta = 0.6$  and  $\alpha = 2$ . Inset shows how the normalized buffering capacity slightly decreases with  $L$ .

rons and the explosion of activity occurs at modules closer to the root module (top panel). As the activation threshold  $\theta$  is increased, it becomes more difficult for the recurrent noise to activate neurons and  $L_c$  increases (bottom panel).

In the complementary region of parameter space (green in Fig. 3a), the mean activity  $\mu^l$  decreases with  $l$ . In this regime the recurrent noise mainly affects the mean activity with a negative input. An example of  $\mu^l, l = 1 \dots L$  in this case is shown in the top panel of Fig. 3c.

### 4.3 Short-term memory via transient activity in a path

In this section we quantify the ability of a path to maintain a memory of a sequence of stimuli in its transient neural activity, i.e. its buffering capacity. When there are no recurrent connections ( $\alpha = 0$ ), inputs propagate through the path without being corrupted, with the input sent at time  $t$  by the root module being perfectly represented in module  $l$  at time  $t+l$  (i.e.  $n^l(t) = 1$  for all  $l, t$ ). The transient state of the path thus buffers memory sequences that last  $L$  time steps, where  $L$  is the total number of modules in the path. In the presence of recurrent connections feed-forward inputs carrying the stimuli are in competition with recurrent inputs acting as random noise (because stimuli are uncorrelated with  $J_{ij}^{m,l}$ ). Stimuli are thus progressively degraded as they travel along the path, according to (14). This is illustrated in the bottom panels of Fig. 3b-c, where we show  $n^l$  in the different regimes of mean

activity described above. In regimes I, patterns are slightly degraded by their travel through each module, until they reach module  $l \geq L_c$  where  $n^{l > L_c}$  drops to a value close to zero. Before reaching this critical module, the overlap decreases moderately since degradation mainly occurs by activating silent neurons, which does not dramatically affect the overlaps for the small coding levels considered here. In regime II, the overlap decreases with  $l$ , smoothly but relatively fast, since silencing neurons that should be active has a strong impact on the value of the overlaps. In order to quantify the ability of a path to buffer inputs, we define the buffering capacity as the area under the  $n^l$  versus  $l$  curve (White et al., 2004). In Fig. 4b we show how this area decreases with the amount of recurrent connections for different values of  $\theta$ , and  $L = 100$ . For  $\theta = 0.4$ , the profile of mean activity corresponds to regime I and the area under the  $n^l$  curve is close to  $L$  ( $L$  inputs are perfectly represented in the path of length  $L$ ) until  $L_c(\alpha)$  becomes smaller than  $L$  and the area decreases sharply with the amount of recurrent connections. For  $\theta = 0.6$  and  $0.8$  the observed decrease occurs in regime II. The specific effect of  $\theta$  on the buffering capacity is shown in Fig. 4c. The optimal value of  $\theta$  varies from  $\theta = \frac{1}{2}$  at  $\alpha \simeq 0$  to  $\theta \simeq 0.65$  at  $\alpha = 4.6$ , the storage capacity of a single module. Finally in Fig. 4d, we show how the buffering capacity scales with  $L$  for fixed parameters  $\theta = 0.6$  and  $\alpha = 2$ .

#### 4.4 Short-term memory via reverberating activity in a path

In this section the root module conveys a stream of random stimuli, except at  $t = 0$  for which the root module is set in the state  $\sigma^0(t = 0) = \xi^{\mu_0, l_0}$  corresponding to a pattern stored in the synaptic matrix of module  $l_0$ . This scenario is illustrated in Fig. 5a. To describe the propagation and retrieval of this pattern we introduce  $m^l(t)$ , the overlap between  $\sigma^l(t)$ , the state of module  $l$  at time  $t$ , and  $\xi^{\mu_0, l_0}$

$$m^l(t) = \frac{1}{Nf(1-f)} \sum_i^N \langle (\xi_i^{\mu_0, l} - f) \sigma_i^l(t) \rangle \quad (18)$$

In modules  $l < l_0$ , pattern  $\xi^{\mu_0, l_0}$  has the same status than a random stimulus, and the order parameter  $m^{l < l_0}(t)$  is governed by equation (14). In module  $l = l_0$ , recurrent connections tend to sustain the pattern  $\xi^{\mu_0, l_0}$ , and the overlap obeys equation (28) which can be found in the Appendix 7.2. Once the pattern is retrieved in module  $l_0$ , it propagates further in deeper modules  $l > l_0$ .

Retrieval in module  $l_0$  differs from the single module case in two aspects: first, the retrieval state has to persist despite the presence of feed-forward noise caused by random patterns of activity that enter the path at  $t \neq 0$ . And second, at  $t = l_0$ , module  $l_0$  is cued with a noisy version of  $\xi^{\mu_0, l_0}$ . Below we study these two effects separately and then show how they impact the storage capacity of a path.

##### 4.4.1 Effect of feed-forward inputs on the existence of a retrieval state

The ongoing stream of stimuli in the path tends to destabilize a retrieved pattern by activating neurons that should not be active. This increase of activity ( $\mu^l \simeq 2f$  in modules close to the root module) mediates an increase in recurrent noise since the recurrent connections also reflect the storage of patterns  $\xi^{\mu, m_0} \neq \xi^{\mu_0, m_0}$ . In order to accommodate this higher recurrent noise, modules embedded in a path need to have a storage load smaller than  $\alpha_c$ . The specific condition is given by (see Appendix 7.3 for derivation):

$$\alpha < \frac{\alpha_c}{1 + \frac{\mu^{l_0}}{f}} \quad (19)$$

where  $\mu^{l_0}$  is the mean activity in module  $l_0$  induced by the random stream of inputs, before retrieval, which is obtained from the steady-state solution of equation (15), and  $\alpha_c$  is the storage capacity of an isolated module. For modules close to the root module for which  $\mu^{l_0} \simeq f$ , equation (19) leads to a maximal storage load  $\frac{\alpha_c}{2}$  under the presence of feed-forward noise. In deeper modules, the value of the maximal storage load for which the retrieval state is stable depends on the profile of mean activity  $\mu^l, l = 1 \dots L$ . In regime I (Fig. 3a), for modules close to the critical module  $l \lesssim L_c$ , where the mean activity is larger than  $f$ , the maximal storage load will be even more reduced. In regime II, the mean activity decreases with  $l$  and the maximal storage load compatible with the feed-forward drive can in principle be higher than for the initial module. Note however that we do not take advantage of this effect in the present work, since for simplicity we only consider equal storage load in each module.

##### 4.4.2 Retrieval from corrupted signal

The signal eliciting retrieval in module  $l_0$  is a corrupted version of pattern  $\xi^{\mu_0, l_0}$ , i.e.  $m^{l_0}(t = l_0) = m_0 < 1$ . The success of retrieval thus depends on whether this corrupted version is within the basin of attraction of

$\xi^{\mu_0, l_0}$ . Similarly to the single module case, assuming  $\mu^{l_0} = O(f)$  and  $f \ll 1$ , the condition for retrieval is:

$$m^{l_0}(t = l_0 + 1) \simeq \begin{cases} 1 & \text{if } m_0 > \theta \\ 0 & \text{if } m_0 < \theta \end{cases} \quad (20)$$

#### 4.4.3 Storage capacity of a path

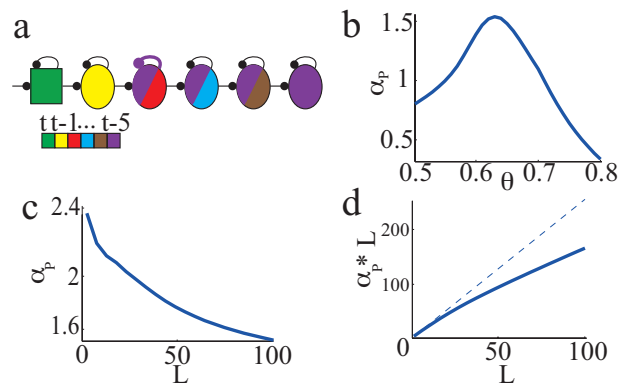
Because of the two phenomenons described above, modules of the TRN should have a storage load lower than the storage capacity of an isolated module ( $\alpha < \alpha_c$ ) in order for a path to serve as an auto-associative memory. We have iterated the dynamical equations (28) (see Appendix 7.2) in order to measure the storage capacity of paths of length  $L$

$$\alpha_L = L * \alpha_P(L) \quad (21)$$

Where  $\alpha_L$  quantifies the total number of patterns that can be retrieved by cueing from the root module. In the definition of  $\alpha_L$  we also impose that retrieved memories can be read-out in the last module of the path, which we define as  $m^L(t \gg L) > \frac{1}{2}$  (see Discussion).  $\alpha_L$  crucially depends on the parameter  $\theta$  because i) it controls the size of the basins of attractions, cf (20); ii) the single module storage load  $\alpha_c$  depends on  $\theta$ ; and iii) the corruption of patterns during feed-forward propagation strongly depends on  $\theta$  (e.g. Fig. 4c). In Fig. 5b we show the storage capacity  $\alpha_{L=100}$  as a function of  $\theta$ , a maximum of  $\alpha_L \simeq 0.3 * \alpha_c * L$  is reached for  $\theta \simeq 0.6$ . In other words, to accommodate all the constraints related to having an attractor network embedded in a feed-forward path of length  $L = 100$ , the storage load of each module  $\alpha_P = \frac{\alpha_L}{L}$  has to be reduced by  $\simeq 1/3$  compared to the case of an isolated module. This ratio decreases further as the length of the path increases (Fig. 5c), going from a ratio of  $\frac{1}{2}$  for  $L = 1$  (maximal ratio that allows to sustain a pattern with a mean activity  $2f$  in the presence of feed-forward inputs according to (19)) to  $\simeq 0.23$  for  $L = 50,000$  (not shown). Figure 5d shows how the capacity of a path  $\alpha_L$  increases sub-linearly with  $L$ .

## 5 Memory properties of the full TRN

As seen in the previous section, the memory properties of a path crucially depends on its length  $L$ . The buffering capacity scales with the length of the path, if working with appropriate parameters, e.g. avoiding explosion of mean activity with storage load small enough to have  $L_c > L$ . And the maximal storage load per



**Figure 5** Storage capacity of a path. **a** Illustration of retrieval in a path, a sequence of stimuli 'magenta,brown,blue,red,yellow,green' is presented by the root module. The recurrent connectivity of the second module is designed to maintain the magenta stimulus in short-term memory. Once retrieval has been elicited in this module, the activity in this module and subsequent ones represent the magenta stimulus together with the ongoing sequence. **b** Optimal single module storage load as a function of  $\theta$  for a path of length  $L = 100$ . **c** Optimal single module storage load as a function of path's length. **d** Sub-linear increase of storage capacity as a function of path's length. Storage capacity is optimized over  $\theta$  for each value of  $L$  here.

module  $\alpha_P(L)$  decreases with  $L$  due to increasing noise accumulation with paths lengths. We now use these results to describe the memory properties of full TRNs of  $M$  modules. TRN are composed of feed-forward paths of length  $L$ , where  $L$  is related to  $M$  through the divergence ratio  $d$ , measuring the ratio between the numbers of modules in two successive levels of the TRN (see Fig. 1):

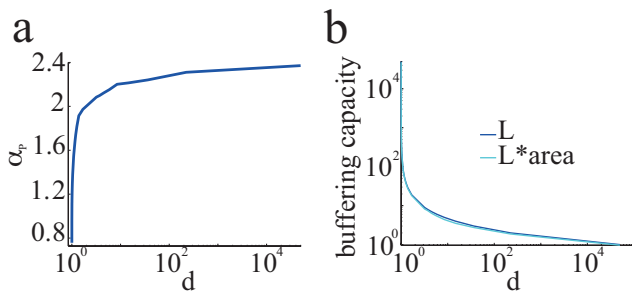
$$L(M, d) = \frac{\log\left(1 + \frac{M(d-1)}{d}\right)}{\log(d)} \quad (22)$$

### 5.1 Storage capacity of TRN

The TRN being composed of feed-forward paths of length  $L$ , its storage capacity is given by

$$\alpha_{TRN} = M \alpha_P(L(M, d)) \quad (23)$$

as defined above  $\alpha_P(L)$  is the maximal storage load of a single module that ensures proper working of a path of length  $L$ . As shown in Fig. 5b,  $\alpha_P(L)$  is a decreasing function of  $L$ , and as such the capacity of the TRN is maximized for a fully parallel architecture  $L = 1$  ( $d = M$ ) and minimized for a fully serial architecture  $L = M$  ( $d = 1$ ). This is shown in Fig. 6 where we show  $\alpha_{TRN}$  as a function of  $d$  for a TRN of  $M = 50,000$  modules, which would for instance correspond to a sensory-hierarchy of  $500 * 10^6$  neurons with columns



**Figure 6** Memory properties of full TRNs of  $M = 50,000$  modules. On the  $x$ -axis  $d = 10^0$  corresponds to a fully serial architecture as shown in Fig. 1a, while  $d = M$  corresponds to a fully parallel architecture as shown in Fig. 1b. **a** Storage capacity as a function of the expansion ratio. **b** Lower and upper bounds (cyan and blue curves) on the buffering capacity as a function of the expansion ratio, for  $\theta = 0.4$ ,  $\alpha = 0.85$ .

composed of  $10^4$  neurons. From the fully serial to the fully parallel organization, the total capacity  $\alpha_{TRN}$  is increased by a factor  $\simeq 3$ .

## 5.2 Buffering inputs in the TRN

The ability of the TRN to buffer inputs is strongly dependent on the length of its paths, as shown in Fig. 4d. In principle the buffering capacity of a TRN of depth  $L$  is larger than the buffering capacity of a path of length  $L$ , since the same patterns are represented in multiple paths and this redundancy could be exploited by a readout. On the other hand the buffering capacity is upper bounded by the length of the paths  $L$ . This lower and upper bounds correspond to the cyan and blue curves in Fig. 6b, which decreases sharply with  $d$ .

## 6 Conclusion

In this paper, we have studied the short-term memory properties of trees of recurrent networks (TRN). For this neural architecture the feed-forward tree-like backbone of modules supports short-term memory via transient dynamics (Ganguli et al., 2008; Jaeger and Haas, 2004; Lim and Goldman, 2012; Maass et al., 2002; White et al., 2004), and the recurrent connectivity inside each module supports short-term memory via reverberating activity (Amit, 1989; Amit and Brunel, 1997; Hopfield, 1982; Tsodyks and Feigl'man, 1988). The ability of TRN to memorize sequences of stimuli via transient dynamics has been quantified using the buffering capacity (White et al., 2004), a measure of how many stimuli, successively presented at the first module, are well represented in the pattern of activity of the TRN at a given time. Networks with feed-forward structures

have been shown to be optimal to implement such memory mechanisms, since they avoid interference between patterns presented at different times (Ganguli et al., 2008; Lim and Goldman, 2012). In TRN, stimuli propagate in a feed-forward manner and reach modules of increasing depth as time is incremented. The buffering capacity of a TRN is thus circumscribed by its depth, and crucially depends on a faithful propagation of stimuli along the network. We have shown that such a faithful propagation is mainly controlled by two factors. The amount of recurrent connections in each module, controlling the strength of the recurrent inputs elicited by a given level of activity in each module; those inputs act as noise by competing with the feed-forward inputs. The activation threshold of neurons, which sets the levels of activity in the modules, as well as tunes signal corruption, by controlling whether recurrent noise tends to activate neurons that should be silent or active according to the incoming feed-forward input.

The ability of TRN to hold a stimulus in memory via reverberating activity has been quantified with the storage capacity, i.e. the number of patterns of activity that modules can maintain in persistent activity, see e.g. (Amit, 1989). Compared to standard studies of an isolated attractor network, the capacity analysis of TRN requires to take into account two additional constraints. First, a state of persistent activity in a module needs to be robust to feed-forward inputs that act as a source of noise. The consequence of this noise is an increased mean activity in the module, leading to more recurrent noise. This requires to decrease the amount of stored patterns per modules compared to the case of an isolated network. Second, the network is cued with a corrupted signal, due to noisy propagation of signals along a TRN. This requires to be able to compute the extent to which patterns of activity are corrupted when traveling through a TRN, as well as to compute the size of the basins of attractions associated to each stored memory pattern. Again the activation threshold plays here a crucial role, since it controls signal propagation, the size of the basins of attractions, as well as the single module storage capacity. Because a pattern of activity is all the more corrupted as it has travelled through many modules, TRN with low depth are better suited to reach high storage capacities.

TRNs are able to support both kinds of short-term memory mechanisms, with the non trivial observation that similar values of a key parameter, the activation threshold, are optimal in both settings. Although both kinds of short-term memory are optimized by similar values of the activation threshold, the buffering and storage capacity of a TRN depend strongly on its depth, with a purely serial TRN architecture being optimal for short-



term memory via transient dynamics and a purely parallel architecture being optimal for short-term memory via reverberating activity.

We have introduced TRN in order to investigate memory properties of sensory cortices, which can be described as made of a few neurons driven by sensory stimuli via thalamic inputs, and a vast majority of interneurons mainly receiving cortical inputs (Braitenberg and Schütz, 1991). Cortico-cortical connectivity has been shown to be spatially clustered (Bosking et al., 1997; DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Pucak et al., 1996), leading to the notion of cortical column. These columns are modeled by the modules of our network, which we have taken to be attractor neural networks, a class of networks whose connectivity profile is consistent with local connectivity in sensory cortices (Brunel, 2016), and that has been shown to efficiently encode sensory stimuli (Tkačik et al., 2010). The extent to which cortical columns are organized in series or in parallel with respect to the thalamic inputs is not known, although, in the visual hierarchy, available data makes it clear that there is some form of serial feed-forward organization. Anatomy has shown that as the depth in the hierarchy is increased, networks are less and less under the influence of the thalamus, with  $\simeq 0.2\%$  of inputs to V1 neurons having a thalamic origin, versus  $\simeq 0.1\%$  for V2 and  $\simeq 0.05\%$  for V4 neurons (Markov et al., 2010). It has also been shown that activity in V2 is strongly dependent on activity in V1, as suggested by lesions of V1 suppressing activity in V2 (Girard and Bullier, 1989). Moreover, along the visual hierarchy, stimuli are processed with increasing delays, with 40 ms and 80 ms delays in V1 and infero-temporal cortex respectively, suggesting a form of serial processing (Thorpe and Fabre-Thorpe, 2001). However, whether such serial organization supports short-term memory via transient dynamics is not obvious since this difference in processing delays between V1 and infero-temporal cortex is much smaller than the behavioral time scales associated with short-term memory. Nevertheless it can be argued that more advanced analysis could reveal short-term memory via transient dynamics at the scale of a brain area (Klampfl et al., 2012; Nikolić et al., 2009).

Neural activity in a TRN forms sensory representations of ongoing stimuli that can be held in short-term memory. A question we have not discussed so far is the one of the reading of such representations by a read-out system. For short-term memory with reverberating activity, once a stimulus has elicited a pattern of persistent activity in one of the module, this

pattern propagates down to sub-sequent modules (see Fig. 5a). A natural read-out for such a representation would be the last module of each path, which would correspond to associative areas in a cortical hierarchy. In Supplementary materials we show how reading-out through the last module of a path is done at the expense of reducing the storage load to avoid dramatic signal degradation from the module maintaining persistent activity to the end of the paths. Regarding short-term memory with transient dynamics, which buffers sequences of stimuli encoded on the whole depth of a TRN, a read-out system has to be connected to modules of all depths in order for reading such sensory representations (White et al., 2004). In a cortical setting such a read-out system could also be implemented by areas on the top of sensory hierarchies, with feed-back connections, not modeled in a TRN, sampling from areas lower in the hierarchy.

Anatomical studies reveal a cortical architecture that is much richer than the simple architecture of TRN, which include neither feed-back nor shortcuts between modules. Other simplifications are the one-to-one projections from one layer to the next, and the fact that patterns stored in different layers have no semantic relationships. While these simplifications allowed us to quantify the memory properties of such networks in depth, it remains to be investigated how these properties will be affected in more realistic networks. Nevertheless it is remarkable that it has remained possible to track analytically the dynamics of such cellular-resolution large-scale networks, and evaluate the computational performance of various connectivity schemes, such as serial or parallel. Such an approach, linking structural and functional properties of this type of network models, together with the advent of large-scale single-cell resolution connectomes, will allow to shed new lights on the computational roles of cortical networks.

## 7 Appendix

### 7.1 Steady-state mean activity profiles for $f \ll 1$

We consider a module with mean activity  $\mu = O(f)$  receiving feed-forward inputs from a module whose mean activity is  $g = \frac{f}{s}$ . From equations (15), the fixed point equations relating  $\mu$  and  $g$  is

$$\mu = gH\left(\frac{\theta - 1}{\lambda\sqrt{\alpha\mu}}\right) + (1 - g)H\left(\frac{\theta}{\lambda\sqrt{\alpha\mu}}\right) \quad (24)$$

Using the estimate  $H(x) \underset{x \gg 1}{\simeq} \frac{1}{x\sqrt{2\pi}} e^{-\frac{x^2}{2}} \simeq e^{-\frac{x^2}{2}}$ , and  $\frac{f|\log f|}{g|\log g|} \simeq \frac{f}{g}$  the fixed point equation can be rewritten

$$\mu \approx g \left( 1 + g^{sx-1} - g^{sx(1-\theta^{-1})^2} \right) \quad (25)$$

with  $x = \frac{\theta^2}{2\lambda^2\alpha f|\log f|}$ . Comparing the two last terms of equation (25) allows to understand whether  $\mu$  increases or decreases compared to  $g$  and thus to describe the two regimes of mean activity profiles of section 4.2. Moreover, for the regime of increasing activity along the path, the transition from  $\mu = O(f)$  to  $\mu = O(1)$  arises for  $s = 1/x$ , i.e.  $g = xf$ . By differentiating (24) as shown in Supplementary materials, this allows to give expressions for the depth  $L_c$  at which the transition occurs. For  $\theta < \frac{1}{2}$

$$L_c = 2\sqrt{\frac{\pi}{xf^{2(x-1)}|\log f|}} \quad (26)$$

and for  $\theta > \frac{1}{2}$  and  $\alpha > (2\theta^{-1} - \theta^{-2})\alpha_c$  the critical depth scale as

$$1/L_c \propto \frac{1}{(\theta^{-1} - 1)\sqrt{x|\log f|}} f^{x(1-\theta^{-1})^2} - \frac{1}{\sqrt{x|\log f|}} f^{x-1} \quad (27)$$

## 7.2 Dynamical equations for memory retrieval in a path

In order to describe the retrieval of pattern  $\xi^{l_0,1}$  in a module  $l_0$  receiving feed-forward inputs from module  $l_0 - 1$  (see section 4.4), we have used the following dynamical equations, whose derivation is detailed in Supplementary Materials.  $m^{l_0}$  (resp.  $m^{l_0-1}$ ) is the overlap between the activity in module  $l_0$  (resp.  $l_0 - 1$ ) and  $\xi^{l_0,1}$ .

$$\begin{aligned} m^{l_0}(t+1) &= m^{l_0-1}(t)\{(1-f)[I_t(1,1) - I_t(1,0)] \\ &\quad + f[I_t(0,1) - I_t(0,0)]\} \\ &\quad + \mu^{l_0-1}(t)\{I_t(1,1) - I_t(1,0) \\ &\quad - [I_t(0,1) - I_t(0,0)]\} \\ &\quad + I_t(1,0) - I_t(0,0) \end{aligned} \quad (28)$$

and

$$\begin{aligned} \mu^{l_0}(t+1) &= m^{l_0-1}(t)f(1-f)\{I_t(1,1) - I_t(1,0) \\ &\quad - [I_t(0,1) - I_t(0,0)]\} \\ &\quad + \mu^{l_0-1}(t)\{f[I_t(1,1) - I_t(1,0)] \\ &\quad + (1-f)[I_t(0,1) - I_t(0,0)]\} \\ &\quad + fI_t(1,0) + (1-f)I_t(0,0) \end{aligned} \quad (29)$$

with

$$I_t(a,b) = H\left(\frac{\theta - (a-f)m^{l_0}(t) - b}{\sqrt{\alpha\mu^{l_0}(t)}}\right) \quad (30)$$

## 7.3 Impact of feed-forward noise on persistent activity

The random sequence of inputs is a form of noise that reduces the capacity for retrieval states. To evaluate the existence of a retrieval state  $\xi^{\mu_0,m_0}$  under these conditions, we examine the stability of the module  $l$  assuming that it receives random feed-forward inputs from the module  $l-1$  with a steady state mean activity  $\mu_{eq}^{l-1}$ . We re-write the equations for retrieval in module  $l$  in terms of the order parameters  $f_0^l$  and  $f_1^l$ , which measure the fraction of background neurons (neurons  $i$  such that  $\xi_i^{\mu_0,m_0} = 0$ ) that are active and the fraction of foreground neurons ( $i$  such that  $\xi_i^{\mu_0,m_0} = 1$ ) neurons that are active. These order parameters are related to  $m^l$  and  $\mu^l$  by  $m^l = f_1^l - f_0^l$  and  $\mu^l = f * f_1^l + (1-f) * f_0^l$ .

$$\begin{aligned} f_1^l(t+1) &= \mu_{eq}^{l-1} H\left(\frac{\theta - m^l(t) - 1}{\sqrt{\alpha\mu^l(t)}}\right) + \\ &\quad (1 - \mu_{eq}^{l-1}) H\left(\frac{\theta - m^l(t)}{\sqrt{\alpha\mu^l(t)}}\right), \quad t \geq 1 \\ f_0^l(t+1) &= \mu_{eq}^{l-1} H\left(\frac{\theta - 1}{\sqrt{\alpha\mu^l(t)}}\right) + \\ &\quad (1 - \mu_{eq}^{l-1}) H\left(\frac{\theta}{\sqrt{\alpha\mu^l(t)}}\right), \quad t \geq 1 \end{aligned} \quad (31)$$

If we assume that a pattern has been retrieved, i.e.  $m^l(t=1) \simeq 1$  and  $\mu^l(t=1) = f + \mu_{eq}^l$ . This pattern remains stable if  $f_1^l$  remains of order 1 and  $f_0^l$  remains of order  $f$ , which comes down to have, for  $\mu_{eq}^{l-1} = O(f) \ll 1$ ,

$$\begin{aligned} H\left(\frac{\theta - 1}{\sqrt{\alpha(f + \mu_{eq}^l)}}\right) &= 1 - f^{\frac{x(1-\theta^{-1})^2}{1 + \frac{\mu_{eq}^l}{f}}} = O(1) \\ H\left(\frac{\theta}{\sqrt{\alpha(f + \mu_{eq}^l)}}\right) &= f^{\frac{-x}{1 + \frac{\mu_{eq}^l}{f}} - 1} \ll 1 \end{aligned} \quad (32)$$

For  $f \ll 1$  this is satisfied if  $x > 1 + \frac{\mu_{eq}^l}{f}$ , hence the condition (19) on  $\alpha$ .

**Acknowledgements** I would like to thank Nicolas Brunel and Haim Sompolinsky for useful discussions throughout the course this work.

## References

- Amit DJ (1989) Modeling brain function. Cambridge University Press
- Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex* 7:237–252
- Bosking W, Zhang Y, Schofield B, Fitzpatrick D (1997) Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J Neurosci* 17:2112–2127
- Braitenberg V, Schütz A (1991) Anatomy of the cortex. Springer-Verlag
- Brunel N (2016) Is cortical connectivity optimized for storing information [quest]. *Nature neuroscience* 19(5):749–755
- DeFelipe J, Conley M, Jones EG (1986) Long-range focal collateralization of axons arising from corticocortical cells in monkey sensory-motor cortex. *J Neurosci* 6:3749–3766
- Evans M (1989) Random dilution in a neural network for biased patterns. *Journal of Physics A: Mathematical and General* 22(12):2103
- Fuster JM (1995) Memory in the cerebral cortex. MIT Press
- Ganguli S, Huh D, Sompolinsky H (2008) Memory traces in dynamical systems. *Proc Natl Acad Sci USA* 105:18970–18975
- Gilbert CD, Wiesel T (1989) Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J Neurosci* 9:2432–2442
- Girard P, Bullier J (1989) Visual activity in area v2 during reversible inactivation of area 17 in the macaque monkey. *Journal of neurophysiology* 62(6):1287–1302
- Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558
- Jaeger H, Haas H (2004) Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304(5667):78–80
- Klampafl S, David SV, Yin P, Shamma SA, Maass W (2012) A quantitative analysis of information about past and present stimuli encoded by spikes of a1 neurons. *Journal of Neurophysiology* 108(5):1366–1380
- Lim S, Goldman MS (2012) Noise tolerance of attractor and feedforward memory models. *Neural computation* 24(2):332–390
- Maass W, Natschläger T, Markram H (2002) Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput* 14(11):2531–2560
- Markov N, Misery P, Falchier A, Lamy C, Vezoli J, Quilodran R, Gariel M, Giroud P, Ercsey-Ravasz M, Pilaz L, et al. (2010) Weight consistency specifies regularities of macaque cortical networks. *Cerebral Cortex* 21(6):1254–1272
- Nikolić D, Häusler S, Singer W, Maass W (2009) Distributed fading memory for stimulus properties in the primary visual cortex. *PLoS biology* 7(12):e1000260
- Pucak ML, Levitt JB, Lund JS, Lewis DA (1996) Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *J Comp Neurol* 338:360–376
- Sejnowski TJ (1977) Storing covariance with nonlinearly interacting neurons. *J Math Biol* 4:303–
- Sompolinsky H, White O (2003) Theory of large recurrent networks: from spikes to behavior. In: *Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School*, p Chapter 8
- Thorpe SJ, Fabre-Thorpe M (2001) Seeking categories in the brain. *Science* 291(5502):260–263
- Tkačik G, Prentice JS, Balasubramanian V, Schneidman E (2010) Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences* 107(32):14419–14424
- Tsodyks M, Feigel'man MV (1988) The enhanced storage capacity in neural networks with low activity level. *Europhys Lett* 6:101–105
- White OL, Lee DD, Sompolinsky H (2004) Short-term memory in orthogonal neural networks. *Physical review letters* 92(14):148102