

Context-dependent representation in recurrent neural networks

Gilles Wainrib

Ecole Normale Supérieure, Département d'Informatique, Paris, France.

Abstract

In order to assess the short-term memory performance of non-linear random neural networks, we introduce a measure to quantify the dependence of a neural representation upon the past context. We study this measure both numerically and theoretically using the mean-field theory for random neural networks, showing the existence of an optimal level of synaptic weights heterogeneity. We further investigate the influence of the network topology, in particular the symmetry of reciprocal synaptic connections, on this measure of context dependence, revealing the importance of considering the interplay between non-linearities and connectivity structure.

Keywords: Short-term memory, Echo-state networks, Recurrent neural networks, Random matrix theory, Mean-field theory, Information

1. Introduction

The ability of the neural representation of a signal to depend on its past context and the capacity of short-term memory are essential for most perceptive and cognitive processes, from vision to language processing and decision making [29, 41, 6, 3]. Although adaptation and plasticity may play a crucial role to shape short-term memory, we study here the hypothesis that this phenomenon may be also explained by dynamical network effects in particular due to the recurrent nature of neural networks connectivity [12, 44, 31, 32]. To study this question from a theoretical standpoint, we consider the following classical model of recurrent neural network (RNN)[46, 1, 38, 28], often called *rate model*:

$$\mathbf{x}(t+1) = S(\mathbf{W}\mathbf{x}(t) + \mathbf{V}\mathbf{u}(t) + \eta(t))$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ describes the states of the n neurons, $\mathbf{u}(t) \in \mathbb{R}^m$ represents the signal, $S(\cdot)$ is typically a sigmoid function (*e.g.* $\tanh(\cdot)$), \mathbf{M} is a $n \times m$ matrix projecting the input into the recurrent network, and \mathbf{W} is a $n \times n$ matrix representing the internal connectivity of the recurrent network (\mathbf{W}_{ij} is the connection strength from neuron j to neuron i , and is often called *synaptic weight*). We will make various assumptions on the matrices \mathbf{V} and \mathbf{W} , and in particular we will investigate the disordered system where \mathbf{V}_{ij} are independent $\mathcal{N}(0, \kappa^2/m)$, and \mathbf{W}_{ij} are either independent $\mathcal{N}(0, \sigma^2/n)$ (random asymmetric connectivity) or constrained to be symmetric with $\mathbf{W}_{ij} = \mathbf{W}_{ji} \sim \mathcal{N}(0, \sigma^2/4n)$ (random symmetric connectivity). The parameter σ characterizes the

synaptic weights heterogeneity and is one of the most important parameter in this study because it controls the amount of recurrence in the generation of the representation. The variables $\eta(t)$ are independent random centered Gaussian vectors in \mathbb{R}^n with diagonal covariance $\epsilon^2\mathbf{I}$. In this model, we have assumed a form of "network noise" that is associated with the network dynamics. However, it is also possible to consider at least two other natural sources of noise, namely an "input noise" in which $\mathbf{u}(t) + \eta(t)$ replaces $\mathbf{u}(t)$, or an "output noise" in which the observation of the network state $\mathbf{x}(t)$ is perturbed by a measurement noise $\mathbf{x}(t) + \eta(t)$.

Most attempts to study theoretically short-term memory in similar random neural networks models have been focused on the linear case $S(x) = x$ because of the difficulty to handle this question in the non-linear regime. In the context of machine learning applications, [27] has shown the ability of this system to reconstruct faithfully up to n time-steps in the past. In [22], Fisher Information was used to assess how short-term memory depends upon the connectivity structure and the network size (see below). More recently, [11, 24] have studied the relevance of applying compressed sensing concepts to this problem, showing the possibility for very long short-term memory, scaling exponentially with n , for the recovery of sparse signals.

Before investigating the interplay between non-linearity and connectivity properties, which is the principal objective of the present article, we study statistical and information-theoretic measures (section 2.1) of the dependence of $\mathbf{x}(t)$ upon past stimulus $\mathbf{u}(t - k)$ for the linear model, and further introduce a new measure, the *context capacity* (section 2.2), that will be easier to study in the non-linear case (section 3).

2. Linear model $S(\mathbf{x}) = \mathbf{x}$

Preliminary results. For simplicity, we consider in this section the case of 1D input time-series, i.e. $m = 1$ and $\mathbf{V} = \mathbf{v} \in \mathbb{R}^n$. We introduce the transfer matrix:

$$\mathbf{M} = [\mathbf{v}, \mathbf{W}\mathbf{v}, \mathbf{W}^2\mathbf{v}, \dots, \mathbf{W}^{T-1}\mathbf{v}]$$

With this notation, in the absence of noise $\mathbf{x}(T)$ can be written as $\mathbf{x}(T) = \mathbf{M}\mathbf{u}$ where \mathbf{u} is the T -dimensional vector $(u(0), \dots, u(T - 1))$. Then, in the case of network noise $\mathbf{x}(T) = \mathbf{M}\mathbf{u} + \sum_{k=0}^{T-1} \mathbf{W}^k \eta(t - k)$, in the case of input noise $\mathbf{x}(T) = \mathbf{M}(\mathbf{u} + \eta)$ and finally in the case of output noise $\mathbf{x}(T) = \mathbf{M}\mathbf{u} + \eta$. From these expressions, it is possible to estimate the covariance structure of $\mathbf{x}(T)$, assuming that the time-series \mathbf{u} is a Gaussian process with zero mean and covariance matrix \mathbf{C} :

$$\mathbf{E} [\mathbf{x}(T)\mathbf{x}(T)'] = \mathbf{M}\mathbf{C}\mathbf{M}' + \Omega$$

where Ω is defined according to the type of noise considered:

- Network noise : $\Omega = \epsilon^2 \sum_{k=0}^{T-1} \mathbf{W}^k \mathbf{W}'^k$
- Input noise : $\Omega = \epsilon^2 \mathbf{M}\mathbf{M}'$
- Output noise : $\Omega = \epsilon^2 \mathbf{I}$.

2.1. Some statistical and information-theoretic measures

From the expression $\mathbf{x}(T) = \mathbf{M}\mathbf{u}$ in the absence of noise, one observes readily that when $T = n$, if the square matrix \mathbf{M} is invertible (which is almost surely the case with random connectivity matrices), then the time-series \mathbf{u} of length T can be recovered exactly with the observation of $n = T$ neurons. However, this basic result does not take into account the impact of noise on the representation. In other words, if the matrix \mathbf{M} is not well conditioned, then a small perturbation can lead to huge errors.

The Cramer-Rao bound states that the variance of the reconstruction error for any estimator of \mathbf{u} cannot be smaller than the inverse of the Fisher Information, therefore providing a universal bound for the input recovery problem. This measure of short-term memory has been studied in a remarkable work by [22] for the linear model with network noise, where it was shown that the Fisher matrix is given by

$$\mathcal{I}_{kl} = \mathbf{v}'\mathbf{W}'^k\Omega^{-1}\mathbf{W}^l\mathbf{v}$$

Here, the diagonal element \mathcal{I}_{kk} is the Fisher information that $\mathbf{x}(t)$ contains about the input at time $t - k$, and characterizes the memory decay of the network representation. Interestingly, this measure of memory is independent of the input and characterizes only the recurrent neural network. After rescaling by the noise level $\mathcal{I} \equiv \epsilon^2\mathcal{I}$, one defines the total memory $\bar{\mathcal{I}} = Tr(\mathcal{I})$ which satisfies the following fundamental distinction: if \mathbf{W} is a normal matrix, then $\bar{\mathcal{I}} = 1$, and otherwise, $\bar{\mathcal{I}} \leq n$ and may behave extensively with network size. This result shows that the underlying structure of the connectivity \mathbf{W} may have a profound impact on the dependence of the representation upon past context. In particular, random symmetric (normal) connectivity matrices appear to be less efficient than asymmetric ones (non-normal) in terms of short-term memory.

To further quantify this statement, one can evaluate the mutual information between $\mathbf{x}(T)$ and \mathbf{u} ,

$$\begin{aligned} I(\mathbf{x}(T); \mathbf{u}) &= \frac{1}{2} \log \frac{|\det(\Omega + \mathbf{M}\mathbf{C}\mathbf{M}')|}{|\det \Omega|} \\ &= \frac{1}{2} \log \left| \det \left(\mathbf{I}_n + (\Omega^{-\frac{1}{2}}\mathbf{M})\mathbf{C}(\Omega^{-\frac{1}{2}}\mathbf{M}') \right) \right| \\ &= \frac{1}{2} \log \left| \det \left(\mathbf{I}_T + (\mathbf{M}\mathbf{C}^{\frac{1}{2}})'\Omega^{-1}(\mathbf{M}\mathbf{C}^{\frac{1}{2}}) \right) \right| \end{aligned}$$

which is related to the Fisher information by¹:

$$I(\mathbf{x}(T); \mathbf{u}) = \frac{1}{2} \log |\det(\mathbf{I}_T + \mathbf{C}^{\frac{1}{2}}\mathcal{I}\mathbf{C}^{\frac{1}{2}})|$$

In FIG. 2 (left), the mutual information is displayed as a function of σ for both the symmetric and asymmetric model, showing a clear superiority of the asymmetric model. To understand theoretically this observation, it is possible to use random matrix theory to evaluate the mutual

¹notice that a similar formula can be found in [22].

information for $n \rightarrow \infty$, assuming $\mathbf{C} = \mu^2 \mathbf{I}_T$ and an output noise setting, for which $\Omega = \epsilon^2 \mathbf{I}$. When \mathbf{W} is random asymmetric then

$$I_{asym}(\mathbf{x}(T); \mathbf{u}) \sim \tilde{I}_{asym} = \frac{1}{2} \sum_{k=0}^T \log \left(1 + \frac{\mu^2}{\epsilon^2} n \sigma^{2k} \right)$$

while the case of symmetric connectivity yields

$$I_{sym}(\mathbf{x}(T); \mathbf{u}) \sim \tilde{I}_{sym} = \frac{1}{2} \log \left| \det \left(\mathbf{I}_T + \frac{\mu^2}{\epsilon^2} n \mathbf{L} \right) \right|$$

where \mathbf{L} is a checkerboard matrix filled with rescaled Catalan numbers:

$L_{ij} = C_{p-1} \left(\frac{\sigma}{2}\right)^{2(p-1)}$ if $i + j = 2p$ with $C_p = \frac{1}{p+1} \binom{2p}{p}$, and $L_{ij} = 0$ if $i + j$ is odd.

Indeed, to prove this result one only needs to evaluate the entries of the matrix $\mathbf{M}'\mathbf{M}$. For random asymmetric matrices, in the large n limit:

$$\frac{1}{n} (\mathbf{M}'\mathbf{M})_{ij} = \frac{1}{n} \mathbf{v}' \mathbf{W}^i \mathbf{W}^j \mathbf{v} \rightarrow \delta_{ij} \sigma^{i+j}$$

whereas for random symmetric matrices:

$$\frac{1}{n} (\mathbf{M}'\mathbf{M})_{ij} = \frac{1}{n} \mathbf{v}' \mathbf{W}^{i+j} \mathbf{v} \rightarrow L_{ij}$$

where the Catalan numbers arise as the even moments of the semi-circular law.

The above determinant appears to be rather difficult to compute analytically, however it is still possible to compare the mutual information for these two cases. Applying Hadamard inequality to the determinant in \tilde{I}_{sym} :

$$\tilde{I}_{sym} \leq \frac{1}{2} \sum_{i=1}^T \log \left(1 + \frac{\mu^2}{\epsilon^2} n C_{i-1} \left(\frac{\sigma}{2}\right)^{2(i-1)} \right)$$

and using the inequality $C_p \leq 4^p$, one concludes that the mutual information for symmetric linear RNN is smaller than the one for asymmetric case:

$$\tilde{I}_{sym} \leq \tilde{I}_{asym}.$$

In fact, for any symmetric connectivity matrix with an eigenvalue distribution compactly supported in $[-\sigma, \sigma]$ (not only the semi-circular law), the same conclusion remains valid. Therefore, we have shown in this section that, in the linear model, RNN with symmetric connectivities capture less information about the past inputs than asymmetric ones.

2.2. Context capacity

All the previous estimations were heavily relying on the linear relationship between \mathbf{x} and \mathbf{u} , enabling the use of linear algebra and (random) matrix tools. However, as noticed for instance in [22], extending the above analysis to non-linear models is very challenging. To circumvent this difficulty, we introduce a new measure that quantifies how much the representation of a signal depends upon its past context, and that is amenable to analysis in the non-linear setting. To define our measure of context-dependence, we decompose the input time-series into two parts: the *context* is the input from time $t = 1$ to $t = t_0$, and the *signal* is the input from time $t = t_0 + 1$ up to some time $t = t_0 + \tau$. From this decomposition, we first define the *context sensitivity* χ as follows (FIG. 1, left panel):

- (a) we consider the context to be randomly generated according to a given probability law. At each trial, an independent source of noise η is also generated.
- (b) the signal is kept fixed to a specific time-series $\bar{\mathbf{u}}$
- (c) we estimate the across-trial variance $\chi(\tau)$ of the representations $\mathbf{x}(t_0 + \tau)$ obtained for each realization.

This variance can be explained by two sources of variability, namely the various contexts generated before the signal and the presence of noise in the construction of the representation. In order to normalize this variance with respect to the pure impact of noise, we introduce a measure of the variability of the representation due to noise only, called the *unreliability* coefficient ρ (FIG. 1, right panel):

- (a) we consider the context to be a fixed time-series (say a fixed random sample generated according to the same given probability law), whereas at each trial, an independent source of noise η is generated.
- (b) the signal is kept fixed to a specific time-series $\bar{\mathbf{u}}$
- (c) we estimate the across-trial variance $\rho(\tau)$ of the representations $\mathbf{x}(t_0 + \tau)$ obtained for each realization.

Finally, we are in position to define the *context capacity* $C(\tau)$ as the ratio of the context sensitivity over the unreliability coefficient:

$$C(\tau) = \frac{\chi(\tau)}{\rho(\tau)}.$$

If the representation is context-independent, then $C(\tau) = 1$, and the higher is $C(\tau)$, the more context-dependent is the representation, and moreover, one expects $C(\tau)$ to be a decreasing function of τ . Before studying the non-linear system, we first evaluate the context capacity in the linear case to study its relationship with the above statistical and information-theoretic measures.

We make the assumption that t_0 is large. To compute the context sensitivity, we consider that, before time $t = t_0$, both the noise and the input are treated as random processes. Therefore $\mathbf{x}(t_0)$ is centered and has covariance $Cov(\mathbf{x}(t_0)) = \mathbf{MCM}' + \Omega$. Then, at time $T = t_0 + \tau$, $\mathbf{x}(T)$ is no longer centered, and has covariance $Cov(\mathbf{x}(T)) = \mathbf{W}^\tau (\mathbf{MCM}' + \Omega) \mathbf{W}'^\tau + \Omega_\tau$ where $\Omega_\tau = \sum_{k=0}^{\tau-1} \mathbf{W}^k \mathbf{W}'^k$. Since $t_0 \rightarrow \infty$, one has the identity $\mathbf{W}^\tau \Omega \mathbf{W}'^\tau + \Omega_\tau = \Omega$, and the context sensitivity is given by

$$\chi = Tr(\mathbf{W}^\tau \mathbf{MCM}' \mathbf{W}'^\tau + \Omega)$$

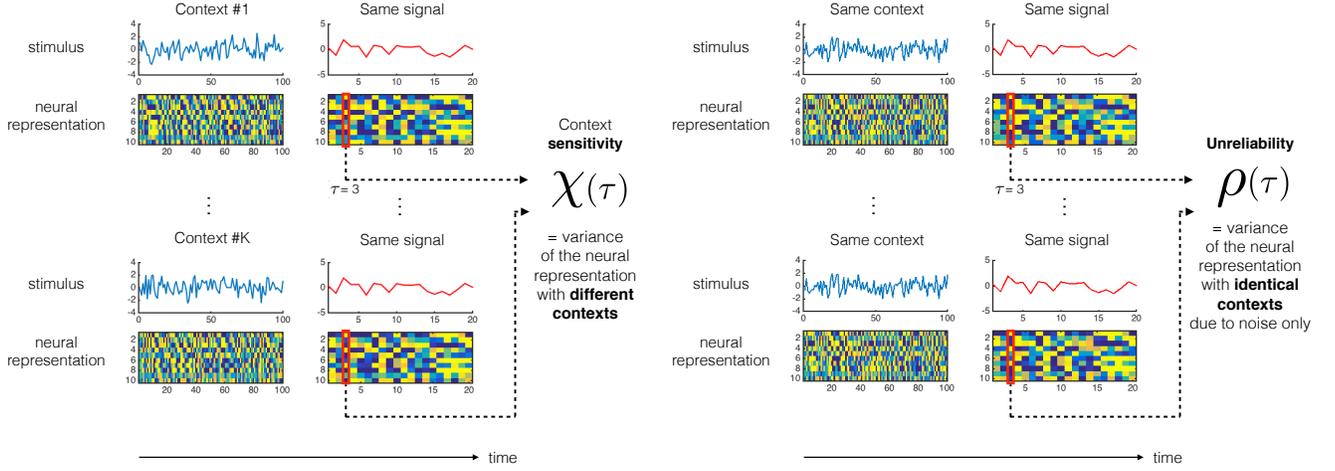


Figure 1: Schematic illustration of the definition of the unreliability coefficient ρ and context sensitivity χ .

To compute the unreliability coefficient, the input is always considered as deterministic, so the covariance of $\mathbf{x}(T)$ is $Cov(\mathbf{x}(T)) = \Omega$. Therefore, the unreliability coefficient is

$$\rho = Tr(\Omega)$$

Finally, the context capacity is given by:

$$C(\tau) = 1 + \frac{Tr(\mathbf{W}^\tau \mathbf{M} \mathbf{C} \mathbf{M}' \mathbf{W}'^\tau)}{Tr(\Omega)}$$

To investigate the impact of connectivity symmetry, we wish to analyze this formula from the RMT point of view, assuming $\mathbf{C} = \mu^2 \mathbf{I}$. First we need the following *trace lemma*: in the limit $n \rightarrow \infty$,

- for the asymmetric random model, $\frac{1}{n} Tr(\mathbf{W}^k \mathbf{W}'^k) \rightarrow \sigma^{2k}$,
- while for the symmetric random model, $\frac{1}{n} Tr(\mathbf{W}^k \mathbf{W}'^k) \rightarrow C_k \left(\frac{\sigma}{2}\right)^{2k}$.

Since $t_0 \rightarrow \infty$, one first remark that: $\mathbf{W}^\tau \mathbf{M}_{t_0} \mathbf{M}'_{t_0} \mathbf{W}'^\tau = \mathbf{M}_T \mathbf{M}'_T - \mathbf{M}_\tau \mathbf{M}'_\tau$ where \mathbf{M}_k is the extraction of the first k columns of \mathbf{M} . Considering first the case of asymmetric connectivity, since $Tr(\mathbf{M} \mathbf{M}') = Tr(\mathbf{M}' \mathbf{M})$ and since the diagonal terms of $\frac{1}{n} \mathbf{M}' \mathbf{M}$ converge to $\sigma^{2(i-1)} \mu^2$, then:

$$Tr(\mathbf{W}^\tau \mathbf{M} \mathbf{C} \mathbf{M}' \mathbf{W}'^\tau) \sim n \mu^2 \left(\frac{1}{1 - \sigma^2} - \frac{1 - \sigma^{2\tau}}{1 - \sigma^2} \right) = n \mu^2 \frac{\sigma^{2\tau}}{1 - \sigma^2}$$

It remains to evaluate $Tr(\Omega)$, which can be done using the trace lemma, yielding:

$$Tr(\Omega) \sim n \frac{1}{1 - \sigma^2}$$

Finally we obtain that when $n \rightarrow \infty$ and T is fixed,

$$\lim_{n \rightarrow \infty} C_{asym}(\tau) = 1 + \frac{\mu^2}{\epsilon^2} \sigma^{2\tau} = \tilde{C}_{asym}(\tau)$$

This formula is very simple and shows that the context capacity for the linear random asymmetric RNN is given by the product of the signal/noise ratio $\frac{\mu^2}{\epsilon^2}$ times a geometrically decaying term $\sigma^{2\tau}$ which is maximal when σ approaches ones. Notice that the same product was already appearing in the expression of the mutual information \tilde{I}_{asym} .

The situation is, again, different for symmetric connectivity. Indeed, the diagonal terms of $\frac{1}{n} \mathbf{M}' \mathbf{M}$ now converge to $C_{i-1} \left(\frac{\sigma}{2}\right)^{2(i-1)}$, and we apply the trace lemma to obtain that when $n \rightarrow \infty$ and T is fixed,

$$\lim_{n \rightarrow \infty} C_{sym}(\tau) = 1 + \frac{\mu^2}{\epsilon^2} \frac{\Theta_\tau}{\Theta_0} = \tilde{C}_{sym}(\tau)$$

where, for $\sigma < 1$:

$$\Theta_\tau = \int_{-\sigma}^{\sigma} \frac{x^{2\tau}}{1-x^2} p(x) dx$$

with $p(x)$ the density of the real eigenvalue distribution of \mathbf{W} . In the case of the semi-circular law, using the generating function of the Catalan numbers, one obtains:

$$\Theta_0 = \frac{2}{1 + \sqrt{1 - \sigma^2}} \text{ and } \Theta_\tau = \Theta_0 - \sum_{k=0}^{\tau} \frac{C_k}{4^k} \sigma^{2k}$$

As for the mutual information \tilde{I}_{sum} , this formula involves Catalan numbers and is more explicit because the evaluation of the trace is more straightforward than the evaluation of the determinant.

From this result, using $C_k \leq 4^k$, we deduce that the context capacity for symmetric random model is lower than its asymmetric counterpart $\tilde{C}_{sym}(\tau) \leq \tilde{C}_{asym}(\tau)$. Therefore the context capacity behaves similarly as more standard measures such as the mutual information, and, as we will see in the next section, offers an interesting alternative to investigate the interplay between non-linearity and connectivity properties.

3. Non-linear model

According to the mean-field theory [38, 9, 20, 8], in particular in the case of input-driven systems [36, 33, 21], σ is the most important parameter in this system since it controls an order-disorder phase transition between a "stimulus-driven regime" for $\sigma < \sigma_{critical}$ and a "chaotic regime" for $\sigma > \sigma_{critical}$. It has been argued that the regime close to criticality in such systems may be relevant in terms of information processing capabilities, both in the fields of neuroscience [2] and artificial intelligence [5, 4]. Therefore, our first aim is to understand how this heterogeneity parameter affects the context capacity in the disordered neural network model. To measure the capacity C we need to specify how we generate various contexts and select a given signal. For simplicity, we

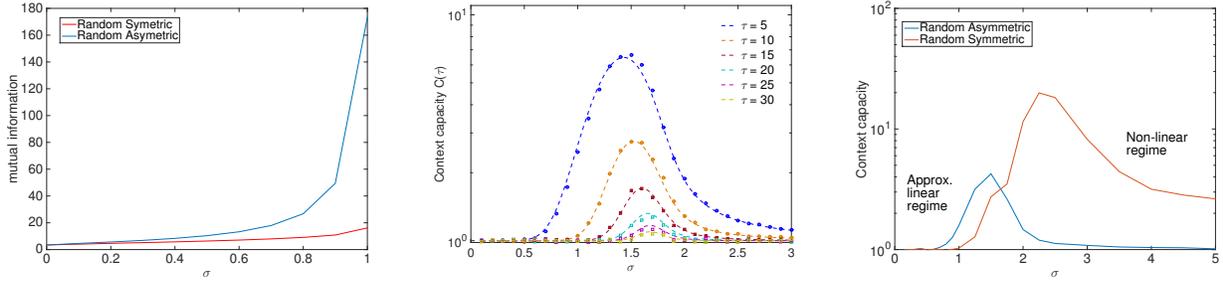


Figure 2: **(Left)** Mutual information as a function of σ for the linear model **(Middle)** Context dependence capacity $C(\tau)$ as defined in (2.2), for the non-linear system with random asymmetric connectivity, as a function of the standard deviation σ of the synaptic weights, for different values of the delay τ . Points correspond to stochastic simulations with $n = 1000$ neurons, and dotted lines correspond to theoretical predictions. **(Right)** Comparison of the context dependence capacity $C(\tau = 5)$ with respect to the symmetry of the connectivity. All numerical simulations of the non-linear model were done with $S(x) = \text{erf}(\sqrt{\pi/2}x)$.

assume that the input $\mathbf{u}(t)$ is a one-dimensional white noise process. In FIG. 2.(middle), we display the context capacity $C(\tau)$ for different values of τ as a function of the synaptic heterogeneity σ for the random asymmetric model. As expected, the context capacity $C(\tau)$ is a decreasing function of τ . More interestingly, it displays a maximal value for an intermediate value of $\sigma > 1$, revealing a trade-off between recurrence-induced memory and non-linear instabilities. To understand this observation from a theoretical standpoint, our strategy is to develop a mean-field approximation of $C(\tau)$. To estimate the context sensitivity, we consider for each trial $k \in \{1, \dots, K\}$ the solution $\mathbf{x}^{(k)}$ of

$$\mathbf{x}^{(k)}(t+1) = S(\mathbf{W}\mathbf{x}^{(k)}(t) + \mathbf{V}\mathbf{u}^{(k)}(t) + \eta^{(k)}(t))$$

with the same initial condition $\mathbf{x}(0)$ and where $\eta^{(k)}$ are independent realization of the random process $\eta(t)$ defined above, and:

- for $1 \leq t \leq t_0$, $\mathbf{u}^{(k)}(t)$ are independent standard Gaussian variables, representing various contexts,
- for $t_0 + 1 \leq t \leq t_0 + \tau$, for all trials, all the $\mathbf{u}^{(k)}(t)$ are equal to $\bar{\mathbf{u}}(t)$, which is a frozen realization of a white noise process.

In this problem, there are two different sources of randomness: the weights matrices \mathbf{W} and \mathbf{V} are randomly drawn once and for all, while the stochastic process η and the various contexts are drawn at each trial k . The first source of randomness, although frozen, will be responsible, in the large n limit, for a phenomenon of self-averaging, which is a well-established property of the mean-field theory [24, 34, 10]. More precisely, population averages of the form $\frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i(t))$ converge to the expectation $\mathbf{E}_{\mathbf{W}, \mathbf{V}}[f(\mathbf{x}_i(t))]$ over the law of the pair (\mathbf{W}, \mathbf{V}) , which is actually a quantity independent of i .

To estimate the across-trials variance of $\mathbf{x}_i^{(k)}(t)$, we need to compute:

$$v_i(t) := \langle \mathbf{x}_i(t)^2 \rangle_K - \langle \mathbf{x}_i(t) \rangle_K^2$$

where $\langle z \rangle_K = \frac{1}{K} \sum_{k=1}^K z^{(k)}$ denotes the across-trials average. Then we take the average over the neural population to obtain a scalar value $v(t) = [v]_N$ where $[z]_N = \frac{1}{n} \sum_{i=1}^n z_i$ denotes the population average. Introducing the sample covariance between two trials: $\rho_{k,l}(t) := [\mathbf{x}_i^{(k)}(t)\mathbf{x}_i^{(l)}(t)]_N$ we can rewrite:

$$v(t) = \langle \rho_{k,k}(t) \rangle_K - \langle \langle \rho_{k,l}(t) \rangle \rangle_K$$

First, the trace term $\gamma(t) := \langle \rho_{k,k}(t) \rangle_K$ can be obtained using classical mean-field theory. Indeed,

$$\gamma(t) = \langle [S(\mathbf{a}(t))]_N \rangle_K$$

where the variables

$$\mathbf{a}_i^{(k)}(t) = (\mathbf{W}\mathbf{x}^{(k)}(t) + \mathbf{V}\mathbf{u}(t) + \eta^{(k)}(t))_i$$

are asymptotically (for large n) independent Gaussian random variable with common zero mean and co-variance:

$$\begin{aligned} (*) \quad \mathbf{E} \left(\mathbf{a}_i^{(k)}(t^+) \mathbf{a}_i^{(l)}(t^+) \right) &= \mathbf{E} \left(\sum_{a,b} \mathbf{W}_{ia} \mathbf{W}_{ib} \mathbf{x}_a^{(k)}(t) \mathbf{x}_b^{(l)}(t) \right) \\ &+ \mathbf{E} \left(\sum_{a,b} \mathbf{V}_{ia} \mathbf{V}_{ib} \mathbf{u}_a^{(k)}(t) \mathbf{u}_b^{(l)}(t) \right) + \mathbf{E} \left(\sum_{a,b} \eta_a^{(k)}(t) \eta_b^{(l)}(t) \right) \end{aligned}$$

where $t^+ = t + 1$. Here, we denote by $\mathbf{E}(\cdot)$ the expectation with respect to the joint law of $(\mathbf{W}, \mathbf{M}, \eta, [\mathbf{u}]_1^{t_0})$. In fact, we only need to consider here the variance $\mathbf{E} \left(\mathbf{a}_i^{(k)}(t) \mathbf{a}_i^{(k)}(t) \right)$, which neither depend on i , since we took the average over (\mathbf{W}, \mathbf{M}) , nor on k since we took the average over η and $[\mathbf{u}]_1^{t_0}$. When considering $k = l$, the third term of the above sum is equal to ϵ^2 by assumption on η , and the second one is given by $\kappa^2 u^2$. The first term is more problematic, since \mathbf{W} and \mathbf{x} could be, in principle, correlated. However, and this is the key point in the mean-field theory, this dependence decays when n becomes very large, and in this asymptotic regime one can pretend \mathbf{x} and \mathbf{W} are independent (see [34, 10] for a rigorous justification, and [33] for a recent exposition of the application of the theory in the case of input-driven systems). Therefore, the first term in the sum can be approximated by $\sigma^2 \gamma(t)$, and we obtain formally: $\mathbf{E}(\mathbf{a}_i(t)^2) = \sigma^2 \gamma(t) + \kappa^2 u^2 + \epsilon^2$. Knowing the mean and variance of $\mathbf{a}_i^{(k)}$, since $\mathbf{x}(t+1) = S(\mathbf{a}(t))$, one can write a recurrent equation, also called the mean-field equation:

$$\gamma(t+1) = F(\sigma^2 \gamma(t) + u(t)^2 + \epsilon^2)$$

where

$$F(x^2) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} S(zx)^2 e^{-z^2/2} dz.$$

Then, the second term to compute is the average of the sample covariance $\rho_{k,l}$ that we denote by $\lambda(t)$. The situation is slightly different: since $\lambda(t+1) = \langle \langle [S(\mathbf{a}_i^{(k)})S(\mathbf{a}_i^{(l)})]_N \rangle \rangle_K$, it appears here

that knowing the variance of $\mathbf{a}_i^{(k)}$ is not sufficient, and we further need to estimate the covariance between $\mathbf{a}_i^{(k)}$ and $\mathbf{a}_i^{(l)}$ for $k \neq l$. The first term in the sum (*) is again given by $\sigma^2\lambda(t)$, the second and third terms are equal to zero since all the trials are independent. Therefore, we obtain:

$$\lambda(t+1) = G(\sigma^2\lambda(t), \sigma^2\gamma(t) + \kappa^2u^2 + \epsilon^2)$$

where:

$$G(x, y) := \int_{\mathbb{R}^2} S(a_1)S(a_2)e^{-a'\Sigma^{-1}(x,y)a} da$$

$$\Sigma(x, y) := \begin{pmatrix} y & x \\ x & y \end{pmatrix}$$

We have obtained a deterministic dynamical system describing the variance across trials $v(t) = \gamma(t) - \lambda(t)$, holding from time $t = 1$ up to time $t = t_0 - 1$:

$$(E) \begin{cases} \gamma(t+1) = F(\sigma^2\gamma(t) + \kappa^2u^2 + \epsilon^2) \\ \lambda(t+1) = G(\sigma^2\lambda(t), \sigma^2\gamma(t) + \kappa^2u^2 + \epsilon^2) \end{cases}$$

To compute the context sensitivity coefficient χ , one needs to solve the above dynamical system from time 1 to $t_0 - 1$. Then, at time $t = t_0$, one switches to a system where $\mathbf{u}^{(k)} = \bar{\mathbf{u}}$ is now the signal and is the same for all trials, so that from time $t = t_0$ to $t = t_0 + \tau$:

$$(E') \begin{cases} \gamma(t+1) = F(\sigma^2\gamma(t) + \kappa^2\bar{u}(t)^2 + \epsilon^2) \\ \lambda(t+1) = G(\sigma^2\lambda(t) + \kappa^2\bar{u}(t)^2, \sigma^2\gamma(t) + \kappa^2\bar{u}(t)^2 + \epsilon^2) \end{cases}$$

To summarize, after solving (E), followed by (E'), one obtains $\chi(\tau) = \gamma(t_0 + \tau) - \lambda(t_0 + \tau)$.

To compute the unreliability coefficient ρ , the only difference is that the context \mathbf{u} is now frozen and does not change across trials. For each trial $k \in \{1, \dots, K\}$, we consider $\mathbf{x}^{(k)}$ the solution of

$$\mathbf{x}^{(k)}(t+1) = S(\mathbf{W}\mathbf{x}^{(k)}(t) + \mathbf{M}\mathbf{u}(t) + \eta^{(k)}(t))$$

Therefore, the above derivation remains valid, with only minor modifications, yielding a slightly modified version of the mean-field dynamical system, for $t = 1$ to $t = t_0 - 1$:

$$(E'') \begin{cases} \gamma(t+1) = F(\sigma^2\gamma(t) + \kappa^2u(t)^2 + \epsilon^2) \\ \lambda(t+1) = G(\sigma^2\lambda(t) + \kappa^2u(t)^2, \sigma^2\gamma(t) + \kappa^2u(t)^2 + \epsilon^2) \end{cases}$$

and for $t = t_0$ to $t = t_0 + \tau$ we obtain exactly the same system (E') defined above. Therefore, after solving (E''), followed by (E'), one obtains $\rho(\tau) = \gamma(t_0 + \tau) - \lambda(t_0 + \tau)$ and finally $C(\tau) = \chi(\tau)/\rho(\tau)$. In FIG. 2 (middle), theoretical predictions are compared with numerical simulations showing a good agreement.

In light of the results obtained for the linear model, the next natural question is to compare the context capacity according to the connectivity properties, and in particular the symmetry of \mathbf{W} . So far, our theoretical results were based on mean-field theory, which heavily relies on the assumption of independent coefficients \mathbf{W}_{ij} (asymmetric random model). Indeed, symmetry introduces a large amount of dependence in the matrix \mathbf{W} and mean-field theory fails for the symmetric random model [17, 18, 19]. However, it is possible to evaluate numerically the context capacity in the symmetric case: as shown in FIG. 2 (right), C is much much lower in the symmetric case for small values of σ , corresponding to an "almost-linear" regime, in accordance with results for the linear model, whereas it becomes much higher for larger values of σ , a regime where the non-linear effects become prominent. Indeed, in this regime, the asymmetric model displays chaotic dynamics hence a poor context sensitivity due to a high unreliability, while the symmetric model has an energy function [16], which prevents chaos and ensures a higher context capacity. With this new observation, it appears that the subtle interplay between connectivity properties and non-linearities is crucial for shaping the way neural networks remember their inputs, and that studying the problem only from a linear algebra perspective may be misleading.

4. Discussion

The problem of short-term memory in recurrent neural network has been investigated using a variety of models, from discrete-time networks [27, 22], to continuous-time networks [26, 7] and spiking networks [30, 43]. Using various approaches from statistics to information theory and dynamical systems, existing literature has mainly focused on three major questions:

- How does memory relate to the number of nodes in the network ?

Since [27], it has been shown that the relationship between the memory capacity and the number n of nodes in the network is essentially linear. Beyond this linear relationship between the memory span and n , a recent study [11] has shown the ability of linear recurrent network to perform a compressed sensing operation and to achieve exponentially long memory for sparse inputs, echoing ideas introduced in [23].

- What is the role of non-linearities ?

While short-term memory in linear random recurrent networks has been studied extensively in [22, 25], the case of non-linear models is less well understood as the increase of the amount of recurrence (e.g. through the spectral radius of the connectivity matrix) controls simultaneously the memory and the amount of nonlinearity in the representation. This trade-off between memory and nonlinearity has been investigated in particular from a theoretical perspective in [14], which shows how these two components can be defined and measured, and how it imposes constraints on the overall performance of reservoir computing systems.

- What is the impact of the connectivity structure ?

As discussed in Section 2, the impact of connectivity structure on memory has been investigated in [22], showing the importance of non-normality of \mathbf{W} . In [45, 25], the specific case of

orthogonal connectivity matrices is also studied, showing the robustness of such structures to noisy perturbations, a type of matrix also under consideration in [11] to demonstrate the RIP property. Furthermore, in a series of articles [35, 39, 37, 40], several authors have explored the impact of connectivity structure in terms of prediction performance, showing that simple deterministic connectivity, such as linear chains, may perform very well in various tasks. The relationship between memory and performance is not straightforward as it may be very task-dependent. However, a recent study [13] of the performance of linear ESN has identified a connection between the Fisher memory curve of [22] and the mean-square-error prediction performance. Finally, various works have been interested in understanding the interplay between connectivity structure and autonomous non-linear reservoir dynamics (e.g. [15, 42]) but not in the perspective of understanding short-term memory properties.

The interplay between connectivity structure and non-linearity may have important consequences for short-term memory and theoretical studies of this problem remain scarce. The present theoretical analysis of short-term memory and context-dependent representation in recurrent neural networks, although limited by its modeling assumptions (choice of the dynamical system, connectivity models), has contributed to advance the understanding of this phenomenon:

1. Since [22], it is known that in linear models, the distinction between normal and non-normal connectivity matrices is very important.
2. Similar results hold for other memory measures (mutual information and context capacity) in the linear case: random symmetric connectivities capture less information about the input.
3. However, we have shown that this is no more the case in non-linear models : in the non-linear regime, symmetric random RNN outperform the asymmetric model.
4. Mean-field theory is reaching its limitation: we have shown how it can be used for the asymmetric model, but it does not provide the key to unlock the symmetric one (more generally, structured models).

References

References

- [1] Amari, S.-I. (1972). Characteristics of random nets of analog neuron-like elements. *Systems, Man and Cybernetics, IEEE Transactions on*, (5):643–657.
- [2] Beggs, J. M. (2008). The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1864):329–343.
- [3] Bernacchia, A., Seo, H., Lee, D., and Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience*, 14(3):366–372.
- [4] Bertschinger, N. and Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural computation*, 16(7):1413–1436.

- [5] Boedecker, J., Obst, O., Lizier, J. T., Mayer, N. M., and Asada, M. (2012). Information processing in echo state networks at the edge of chaos. *Theory in Biosciences*, 131(3):205–213.
- [6] Boutla, M., Supalla, T., Newport, E. L., and Bavelier, D. (2004). Short-term memory span: insights from sign language. *Nature neuroscience*, 7(9):997–1002.
- [7] Büsing, L., Schrauwen, B., and Legenstein, R. (2010). Connectivity, dynamics, and memory in reservoir computing with binary and analog neurons. *Neural Computation*, 22(5):1272–1311.
- [8] Cabana, T. and Touboul, J. (2013). Large deviations, dynamics and phase transitions in large stochastic and disordered neural networks. *Journal of Statistical Physics*, 153(2):211–269.
- [9] Cessac, B., Doyon, B., Quoy, M., and Samuelides, M. (1994). Mean-field equations, bifurcation map and route to chaos in discrete time neural networks. *Physica D: Nonlinear Phenomena*, 74(1):24–44.
- [10] Cessac, B. and Samuelides, M. (2007). From neuron to neural networks dynamics. *The European Physical Journal Special Topics*, 142(1):7–88.
- [11] Charles, A. S., Yap, H. L., and Rozell, C. J. (2014). Short-term memory capacity in networks via the restricted isometry property. *Neural computation*, 26(6):1198–1235.
- [12] Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923.
- [13] COUILLET, R., Wainrib, G., Sevi, H., and Ali, H. T. (2016). The asymptotic performance of linear echo state neural networks. *submitted*.
- [14] Dambre, J., Verstraeten, D., Schrauwen, B., and Massar, S. (2012). Information processing capacity of dynamical systems. *Nature Scientific Reports*, 2:id 514.
- [15] Del Molino, L. C. G., Pakdaman, K., Touboul, J., and Wainrib, G. (2013). Synchronization in random balanced networks. *Physical Review E*, 88(4):042824.
- [16] Dong, D. W. and Hopfield, J. J. (1992). Dynamic properties of neural networks with adapting synapses. *Network: Computation in Neural Systems*, 3(3):267–283.
- [17] Faugeras, O. and MacLaurin, J. (2013). A large deviation principle for networks of rate neurons with correlated synaptic weights. *BMC Neuroscience*, 14(Suppl 1):P252.
- [18] Faugeras, O. and Maclaurin, J. (2014a). Asymptotic description of stochastic neural networks. i. existence of a large deviation principle. *Comptes Rendus Mathématique*, 352(10):841–846.
- [19] Faugeras, O. and Maclaurin, J. (2014b). Asymptotic description of stochastic neural networks. ii. characterization of the limit law. *Comptes Rendus Mathématique*, 352(10):847–852.

- [20] Faugeras, O., Touboul, J., and Cessac, B. (2009). A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs. *Frontiers in computational neuroscience*, 3.
- [21] Galtier, M. and Wainrib, G. (2014). A local echo state property through the largest lyapunov exponent. *arXiv preprint arXiv:1402.1619*.
- [22] Ganguli, S., Huh, D., and Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975.
- [23] Ganguli, S. and Sompolinsky, H. (2010). Short-term memory in neuronal networks through dynamical compressed sensing. In *Advances in neural information processing systems*, pages 667–675.
- [24] Ganguli, S. and Sompolinsky, H. (2012). Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual review of neuroscience*, 35:485–508.
- [25] Hermans, M. and Schrauwen, B. (2010a). Memory in linear recurrent neural networks in continuous time. *Neural Networks*, 23(3):341–355.
- [26] Hermans, M. and Schrauwen, B. (2010b). Memory in reservoirs for high dimensional input. In *Proc. WCCI 2010 (IEEE World Congress on Computational Intelligence)*, pages 2662–2668.
- [27] Jaeger, H. (2001). *Short term memory in echo state networks*. GMD-Forschungszentrum Informationstechnik.
- [28] Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.
- [29] Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., and Moore, K. S. (2008). The mind and brain of short-term memory. *Annual review of psychology*, 59:193.
- [30] Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560.
- [31] Major, G. and Tank, D. (2004). Persistent neural activity: prevalence and mechanisms. *Current opinion in neurobiology*, 14(6):675–684.
- [32] Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84.
- [33] Massar, M. and Massar, S. (2013). Mean-field theory of echo state networks. *Physical Review E*, 87(4):042809.

- [34] Moynot, O. and Samuelides, M. (2002). Large deviations and mean-field theory for asymmetric random recurrent neural networks. *Probability Theory and Related Fields*, 123(1):41–75.
- [35] Ozturk, M. C., Xu, D., and Príncipe, J. C. (2007). Analysis and design of echo state networks. *Neural Computation*, 19(1):111–138.
- [36] Rajan, K., Abbott, L., Sompolinsky, H., et al. (2010). Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 82(1 Pt 1):011903–011903.
- [37] Rodan, A. and Tiño, P. (2011). Minimum complexity echo state network. *Neural Networks, IEEE Transactions on*, 22(1):131–144.
- [38] Sompolinsky, H., Crisanti, A., and Sommers, H. (1988). Chaos in random neural networks. *Physical Review Letters*, 61(3):259.
- [39] Song, Q. and Feng, Z. (2010). Effects of connectivity structure of complex echo state network on its prediction performance for nonlinear time series. *Neurocomputing*, 73(10):2177–2185.
- [40] Strauss, T., Wustlich, W., and Labahn, R. (2012). Design strategies for weight matrices of echo state networks. *Neural Computation*, 24(12):3246–3276.
- [41] Todd, J. J. and Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, 428(6984):751–754.
- [42] Wainrib, G. and Galtier, M. (2015). Regular graphs maximize the variability of random neural networks. *Physical Review E*, 92(3):032802.
- [43] Wallace, E., Hamid, R., and Latham, P. (2013). Randomly connected networks have short temporal memory. *Neural Computation*, 25:1408–1439.
- [44] Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463.
- [45] White, O. L., Lee, D. D., and Sompolinsky, H. (2004). Short-term memory in orthogonal neural networks. *Physical review letters*, 92(14):148102.
- [46] Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1.