Generation of scale-invariant sequential activity in linear recurrent networks

Yue Liu[1,2] and Marc W. Howard[1,2,3]

[1] *Department of Physics, Boston University, Boston, MA 02215*
[2] *Center for Systems Neuroscience, Boston University, Boston, MA 02215*
[3] *Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215*

Abstract

Sequential neural activity has been observed in many parts of the brain and has been proposed as a neural mechanism for memory. The natural world expresses temporal relationships at a wide range of scales. Because we cannot know the relevant scales *a priori* it is desirable that memory, and thus the generated sequences, are scale-invariant. Although recurrent neural network models have been proposed as a mechanism for generating sequences, the requirements for *scale-invariant* sequences are not known. This paper reports the constraints that enable a linear recurrent neural network model to generate scale-invariant sequential activity. A straightforward eigendecomposition analysis results in two independent conditions that are required for scale-invariance. First the eigenvalues of the network must be geometrically spaced. Second, the eigenvectors must be related to one another *via* translation. These constraints, along with considerations on initial conditions, provide a general recipe to build linear recurrent neural networks that support scale-invariant sequential activity.

1 Introduction

The natural world has temporal relationships on a wide range of timescales. Since we cannot predict the relevant timescales *a priori*, it would be desirable to have a dynamic representation of the world that is scale-invariant across time. Decades of research in cognitive psychology demonstrate that human timing and memory behavior exhibit the same signature on a wide range of timescales (Murdock (1962); Glenberg et al. (1980); Rakitin et al. (1998); Howard et al. (2008)). This is termed "scalar timing" (Gibbon et al. (1977)). In models of temporal pattern recognition, a spectrum of scale-invariant functions are used to encode the recent past. Indeed, a suitable representation of time is an essential ingredient in neural circuit models for temporal pattern recognition (Tank and Hopfield (1987); Hopfield and Brody (2000); Buonomano and Maass (2009); Gütig and Sompolinsky (2009)). Sequential neural activity has been observed in many areas of the brain and is thought to have important cognitive functions in memory and decision making, including representing internal thought processes leading up to decisions (Harvey et al. (2012)) and bridging temporal gaps between episodes (MacDonald et al. (2011)).

In light of these considerations, recently it has been proposed that "scale-invariance" is a desirable property for neural sequences (Shankar and Howard (2012, 2013)). In a scale-invariant neural sequence, the cells that are activated later have wider temporal receptive fields. More specifically, the responses of different cells have identical time courses when they are rescaled in time by their peaks. The hypothesis of scale-invariant neural sequences for time is consistent with recent electrophysiological recordings of "time cells" during a delay period when animals are performing various tasks (Pastalkova et al. (2008); MacDonald et al. (2011); Salz et al.

(2016); Tiganj et al. (2018)). The firing fields of time cells that fire later in the delay period are wider than the firing fields of time cells that fire earlier in the delay period.

Many researchers have studied recurrent neural networks that generate sequential activity (Goldman (2009); Rajan et al. (2016); Wang et al. (2018)), but not many of these works considered scale-invariant sequences (but see Voelker and Eliasmith (2018)). In this paper we seek to identify general constraints on the network connectivity for the generation of scale-invariant neural sequences in recurrent neural networks. We study a linear network of interacting neurons. The f-I curve of many neurons are observed to be largely linear (e.g. Chance et al. (2002)). The learning dynamics of linear feedforward neural networks exhibit many similarities compared to their non-linear counterparts (Saxe et al. (2013)). Therefore linear neural networks provide a good model for studying systems-level properties of real neuronal circuits.

The paper is organized as follows. In Section 2 the network constraints for the generation of scale-invariant neural sequences are derived analytically. In Section 3 two example networks with different single cell dynamics are constructed to illustrate the mathematical result. In Section 4, we compare the eigenvectors and eigenvalues of a chaining model and a random recurrent network with those of the examples in Section 3. It is shown that neither of these networks satisfy the structural constraints derived from Section 2, therefore neither of them support sequential neural activity that is scale-invariant.

2 Derivation of the constraints for scale-invariance

In this section we derive the constraints on the connectivity matrix of a linear recurrent network for it to support scale-invariant activity. In Section 2.1, we start with a formal definition of scale-invariance of network activity. In Section 2.2 and Section 2.3 we will derive the two necessary constraints on the connectivity matrix to achieve scale-invariance. Lastly in Section 2.4, we will point out that in addition to the two necessary constraints, a particular initial condition is required for the subsequent network dynamics to be scale-invariant.

2.1 Formulation of the problem

We consider the autonomous dynamics of a linear recurrent network with N neurons:

$$\dot{\mathbf{x}}(t) = \mathbf{M}\mathbf{x}(t), \tag{1}$$

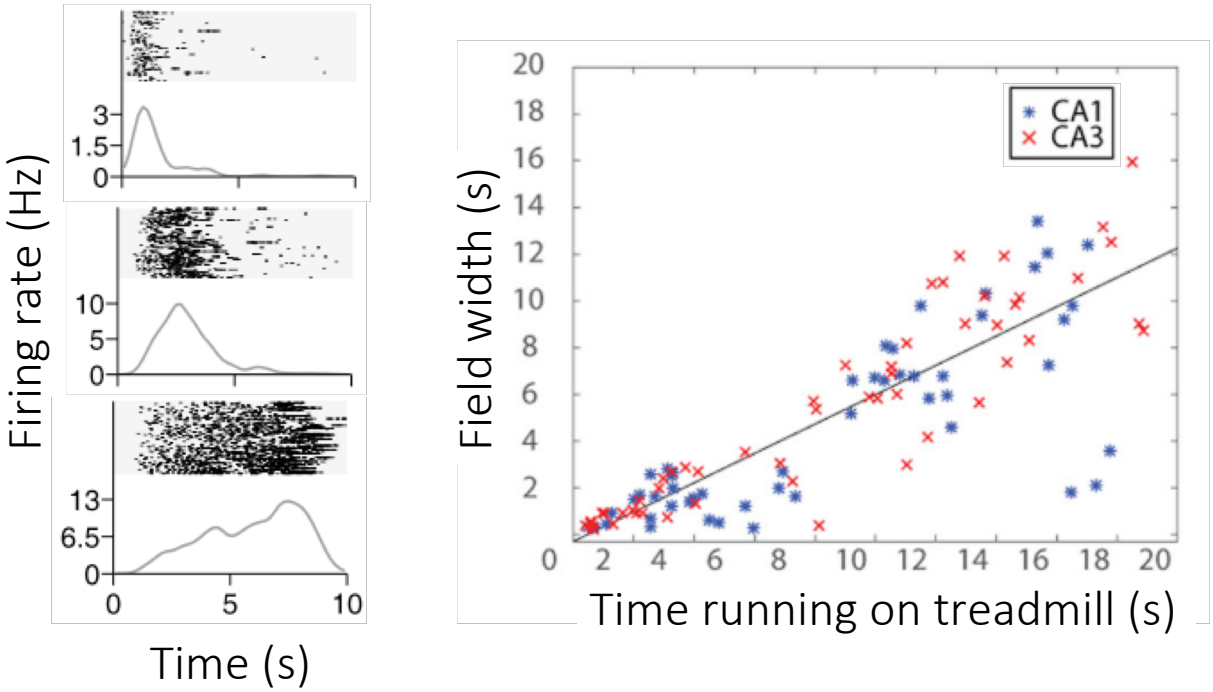
where $\mathbf{x}$ is an N-dimensional vector summarizing the activity of all the neurons in the network and $\mathbf{M}$ is the $N \times N$ connectivity matrix of the network. We consider the case where there is no input into the network since sequential neural activity is thought to be maintained by internal neuronal dynamics (Pastalkova et al. (2008)).

Scale-invariance of the network activity means that the activities of any two neurons in the sequence are rescaled version of each other in time (Figure 1b). Mathematically, this requirement can be written in the following form:

$$x_i(t) = x_j(\alpha_{ij}t), \qquad \forall i, j \in 1, 2, ..., N, \tag{2}$$

That is, for every pair of neurons i and j, their responses $x_i(t)$ and $x_j(t)$ are rescaled in time by a factor α_{ij}. In the following sections we are going to derive two conditions on the connectivity matrix $\mathbf{M}$ necessary for Equation 2 to hold and for the network to generate scale-invariant sequential activity.

a.



b.

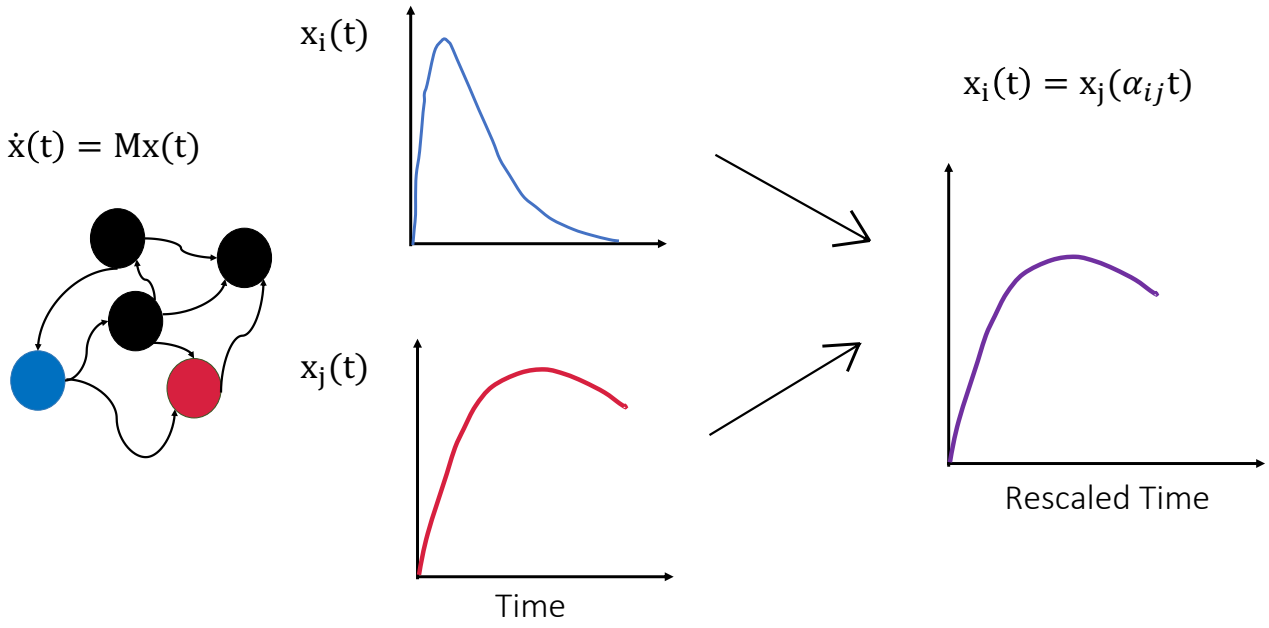


Figure 1: **Scale-invariance for neural sequences and the setup of the problem. a** Left: the raster plots (top) and trial-averaged firing rates (bottom) for three neurons from MacDonald et al. (2011). The neurons were recorded in the hippocampus of rats during a delay period when they were waiting to sample an odor. The neurons that fire later in the delay period show a wider response than neurons that fire earlier during the delay. Right: a scatter plot showing the relationship between the width of each neuron's response and the peak time at which that neuron fires for all neurons recorded from Salz et al. (2016)). Neurons were recorded in the CA3 region of the rat hippocampus during the delay period of a T-maze alteration task when the animal was running on a treadmill. **b.** In this work, we study the dynamics of a linear recurrent network. We seek constraints on the network connectivity matrix $\mathbf{M}$ (**b**, left) such that the activity of every pair of neurons (here for example neurons i and j, middle) are rescaled version of each other in time (right).

2.2 Constraint 1: Weber-Fechner spaced network timescales

We start by solving Equation 1 using the standard eigendecomposition technique. We diagonalize the connectivity matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$ where $\mathbf{\Lambda}$ is a diagonal matrix consisting of the eigenvalues and $\mathbf{U}$ is a matrix whose columns are the eigenvectors of $\mathbf{M}$. The solution of Equation 1 is then a linear combination of exponential functions:

$$x_i(t) = \sum_j U_{ij} A_j e^{\lambda_j t} \equiv \sum_j \tilde{U}_{ij} e^{\lambda_j t}, \tag{3}$$

where the A_i's are constants determined by the initial condition and are absorbed into the definition of the matrix $\tilde{\mathbf{U}}$.

Imposing the scale-invariance condition (Equation 2) on Equation 3, we have

$$\sum_k \tilde{U}_{ik} e^{\lambda_k t} = \sum_k \tilde{U}_{jk} e^{\alpha_{ij}\lambda_k t} \qquad \forall i, j \in 1, 2, ..., N. \tag{4}$$

For this equation to hold, the time-dependent parts on both sides of the equation must be identical. The only way to achieve this is to have a geometric series of eigenvalues (network timescales) (Figure 2a), so that for each λ_k there exists an integer δ such that $\lambda_{k+\delta} = \alpha_{ij}\lambda_k$. For this to hold, all the eigenvalues have to form a geometric series (for example, $\lambda_1 = -1$, $\lambda_2 = -2$, $\lambda_3 = -4$ ...). Therefore we arrive at the first constraint:

Constraint 1: The eigenvalues of the connectivity matrix must form a geometric series.[1]

2.3 Constraint 2: Translation-invariant eigenvectors

A second constraint for Equation 4 to hold is that the rows of $\tilde{\mathbf{U}}$ must satisfy a translation-invariant relationship. $\tilde{U}_{i,k+\delta} = \tilde{U}_{jk}$. This way the basis functions that different neurons pick out will be rescaled versions of each other. Recall from Equation 3 that $\tilde{U}_{ij} = U_{ij}A_j$. Therefore the condition above is equivalent to the columns of the matrix $\mathbf{U}$ being translation-invariant up to a constant. For example, the different eigenvectors could be $\mathbf{v_1} = [1, -1, 0, 0, 0]$, $\mathbf{v_2} = [0, 1, -1, 0, 0]$, $\mathbf{v_3} = [0, 0, 1, -1, 0]$, etc.. Notice that the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{M}$. Therefore, we reach the second constraint (see Figure 2b for a graphical illustration):

Constraint 2: The eigenvectors of the connectivity matrix must consist of the same motif (up to a scaling factor) at translated entries. In other words, the eigenvectors must be translation-invariant.

2.4 A note on initial conditions

Besides the constraints on the connectivity matrix, the initial condition of the network also affects the scale-invariance property of the network activity. Equation 3 requires each neuron to have a specific initial condition, i.e. $x_i(0) = \sum_{j=1}^{N} \tilde{U}_{ij}$, to ensure that the dynamics that ensues are scale-invariant. If in addition we specifically look at balanced motifs whose elements sum to 0 (which will be the case for the examples in Section 3), the constraint on the initial condition becomes:

$$x_i(t=0) = \begin{cases} 0, & \text{if } i \leq N - L + 1 \\ \sum_{j=1}^{N} \tilde{U}_{ij}, & \text{otherwise,} \end{cases} \tag{5}$$

[1] All of the eigenvalues have to have negative real parts to prevent the unbounded growth of network activity.

where L is the length of the repeating motif. In the case where the number of neurons in the network is much larger than the length of motif ($N \gg L$), this constraint on initial condition states that most of the neurons in the network need to be inactive at $t = 0$, whereas a few active neurons act as "input nodes" and propagate their activity to the rest of the network.

If the above initial condition is relaxed, so that instead of specifying the exact initial condition for the active neurons, we assume they have random initial conditions,

$$x_i(t=0) = \begin{cases} 0, & \text{if } i \leq N - L + 1 \\ \mathcal{N}(0,1), & \text{otherwise,} \end{cases} \tag{6}$$

global scale-invariance will generally be broken, but subsets of neurons in the network will still be scale-invariant. We did not study the exact mechanism for how this happens in this paper.

3 Examples

In this section we construct two example networks based on the analytical results derived above and show that they allow scale-invariant sequential dynamics. The connectivity matrices of these networks will have geometric series of eigenvalues and tranlation-invariant eigenvectors, as shown in Section 2. In the first example (Section 3.1), all eigenvalues are real and the neurons in the network have simple bell-shaped temporal receptive fields. In the second example (Section 3.2), all eigenvalues are complex, which gives rise to more complex damping oscillatory single neuron dynamics. We will also show that the network activity is no longer scale-invariant when either of the two constraints is violated. In Section 3.3, we will discuss the relationship of our results to a previously proposed network model that generates scale-invariant sequential activity (Shankar and Howard (2013)). In what follows, all simulations were performed in Python 3.6 using Euler's method with a time step of 0.01 ms.

3.1 Bell-shaped temporal receptive fields

In this example we construct a network that generates sequentially-activated cells with a scale-invariant property. The network consists of $N = 10$ cells. Each cell will have a bell-shaped temporal receptive field, similar to what was observed in electrophysiological recordings of "time cells" (MacDonald et al. (2011); Salz et al. (2016)).

We will construct the connectivity matrix from its eigendecomposition $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$. According to Constraint 1 (Section 2.2), the network must have geometrically spaced eigenvalues. We hence let $\mathbf{\Lambda}$ be a diagonal matrix whose diagonal elements are geometrically spaced between -0.1 and -5.12.

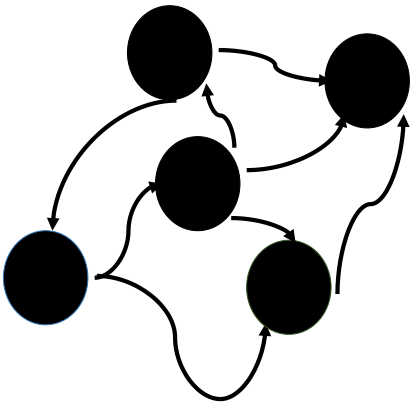
$$\mathbf{\Lambda} = \begin{bmatrix} -0.1 & 0 & 0 & \dots \\ 0 & -0.22 & 0 & \dots \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & -5.12 \end{bmatrix} \tag{7}$$

According to Constraint 2, the eigenvectors of the connectivity matrix $\mathbf{M}$ must consist of the same balanced motif. Equivalently, we will construct the matrix $\mathbf{U}$ such that its *rows* consist of the same motif. In this example we construct the motif to be [1, -1]. Therefore the matrix $\mathbf{U}$ was given by

$$\mathbf{U} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ U_{N,1} & U_{N,2} & U_{N,3} & \dots & U_{N,N} \end{bmatrix}, \tag{8}$$

a.

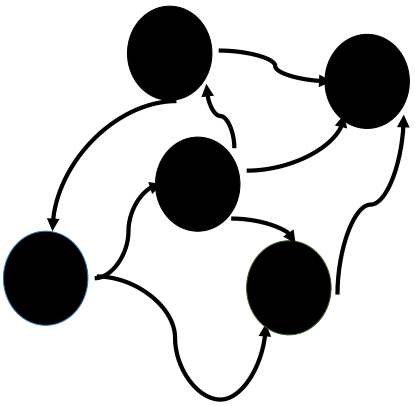
$$\mathbf{M} = \mathbf{U}\,\boldsymbol{\Lambda}\,\mathbf{U}^{-1}$$



$$\begin{bmatrix} \lambda & & \mathbf{0} \\ & k\lambda & \\ & & k^2\lambda \\ \mathbf{0} & & \ddots \end{bmatrix}$$

b.

$$\mathbf{M} = \mathbf{U}\,\boldsymbol{\Lambda}\,\mathbf{U}^{-1}$$



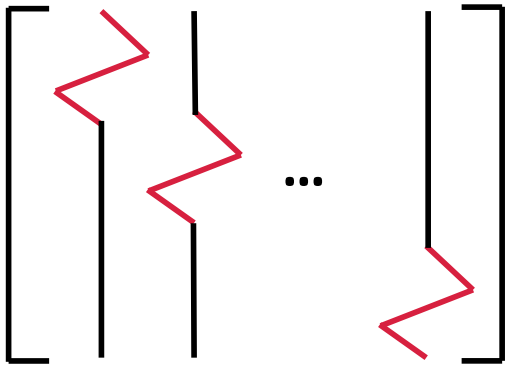


Figure 2: **Graphical illustrations of the constraints for scale-invariance.** To generate scale-invariant activity in linear recurrent networks, its network connectivity must have geometrically spaced eigenvalues (**a**). Furthermore, it must have translation invariant eigenvectors (columns of the matrix **U**) that consist of the same motif (in red) at translated entries (**b**).

where the last row can be arbitrary numbers that are not all zero. In this example they were sampled from $\mathcal{N}(0,1)$. Finally the connectivity matrix of the network $\mathbf{M}$ was computed from $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$.

The simulated activity of the network is shown in Figure 3a (top left). The initial condition is specified to satisfy Equation 5. The bottom left of Figure 3a shows the network activity rescaled along the time axis according to the peak times of each neuron. It is evident that the activations of the neurons are rescaled version of each other, confirming that the constraints derived above indeed lead to scale-invariant sequential activity.

We also broke each of the two constraints above and showed that the resulting activity became no longer scale-invariant (Figure 3a, right two panels). To break Constraint 1 (geometrically spaced eigenvalues), linearly spaced eigenvalues in the same range were used in constructing the matrix $\mathbf{\Lambda}$ instead of geometrically spaced eigenvalues. To break Constraint 2 (translation-invariant eigenvectors), a random vector was added to each row of the matrix $\mathbf{U}$ where each entry was sampled from $\mathcal{N}(0,1)$, making each motif different. As a result, both manipulations generated activity that was no longer scale-invariant (Figure 3a, bottom right two panels).

3.2 Complex single-cell tuning curves

Temporal coding needs not result in a bell-shaped sequential code. Instead, it could also be embedded in the collective activity of neuronal population where single neurons may exhibit highly complex dynamics (Machens et al. (2010)). In this subsection we show that the framework we presented above is sufficiently rich to allow for more complex single cell dynamics. In the previous example, the connectivity matrix has real eigenvalues. Therefore it can only give rise to single cell dynamics that are linear combinations of exponential functions. On the other hand, in this subsection we are going to show that more complex temporal dynamics can be generated if the connectivity matrix has complex eigenvalues.

Following a similar procedure as detailed in Section 3.1, the connectivity matrix with $N = 10$ neurons was set up so that all of its eigenvalues were complex, and their real and imaginary parts both formed geometric series. The real part of the eigenvalues were geometrically spaced between -0.1 and -1.6. The imaginary part of the eigenvalues were geometrically spaced between 0.5 and 8. Since the eigenvalues for real matrices come in complex conjugate pairs, we also included those eigenvalues. The motif was chosen to be [1,-1], same as the example in Section 3.1. The simulated and rescaled activity are shown in Figure 3b (left). The neural activity exhibits more complex dynamics at the same time maintaining scale-invariance.

We also broke each of the two constraints using similar protocols as in the previous example. To break Constraint 1, eigenvalues with linearly spaced real and imaginary parts in the same range were used instead of geometrically spaced ones. To break Constraint 2, a random vector was added to each row of the matrix $\mathbf{U}$ where each entry was sampled from $\mathcal{N}(0,1)$, making each motif different. As shown in Figure 3b (right two panels), the resulting neural activity is no longer scale-invariant.

3.3 A special case: Laplace and inverse Laplace transforms

It should be noted that the geometric series of time constants required by Constraint 1 do not necessarily have to be an emergent property of the network, but can instead be driven by physiological properties of single cells (Loewenstein and Sompolinsky (2003); Fransén et al. (2002); Tiganj et al. (2015); Liu et al. (2018)). Consequently, scale-invariant sequential activity could also be generated by feedforward networks where the neurons in the first layer receive inputs and decay exponentially with a geometric series of intrinsic time constants, and the

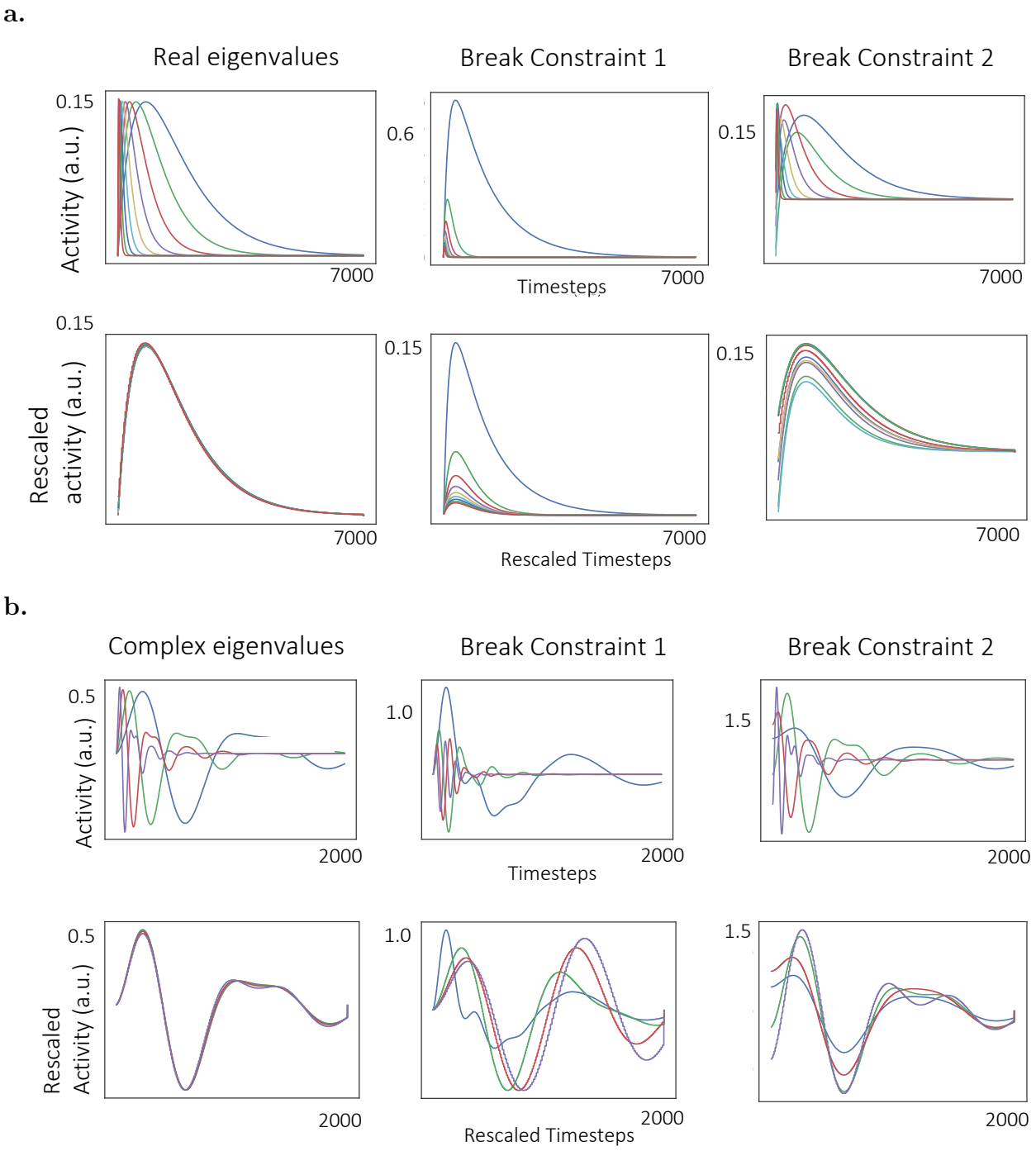


Figure 3: **Generating scale-invariant neural sequences with simple (a) and complex (b) single neuron dynamics**. Using the analytical result, we constructed networks with specific connectivity matrices so that they generate scale-invariant sequential activity (left column). We also broke each of the two constraints derived in Section 2 and showed that the resulting network activity breaks scale-invariance (middle and right columns) **a.** The network with real eigenvalues gives rise to bell-shaped single cell temporal receptive fields (**a**, left top, each line represents the activity of one neuron in the network). The activity of different neurons overlap with each other when rescaled according to their peak times (left bottom). The network activity becomes not scale-invariant when Constraint 1 was broken by choosing linearly spaced eigenvalues or Constraint 2 was broken by choosing random motifs (**a**, middle and right columns, see text for details). **b.** The network with complex eigenvalues gives rise to more complex single cell dynamics while the neural sequence is still scale-invariant (**b** left). When each of the two constraints are broken in the same way as in **a**, the network activity becomes not scale-invariant (right two columns).

neurons in the second layer are driven by the first layer via translation-invariant synaptic weights, implementing the eigendecomposition in Equation 3 explicitly.

One such model has been proposed by Shankar and Howard (Shankar and Howard (2013)). It is a two-layer feedforward neural network. In that model, the first layer neurons $\mathbf{F}$ encode the Laplace transform of the input and have exponentially decaying firing rates with a spectrum of decay constants.

$$\frac{dF_i(t)}{dt} = -\lambda_i F_i(t), \qquad i = 1, 2, ..., N. \tag{9}$$

The activity of the first layer constitutes a scale-invariant sequential activity (see Equation 2). It is also the basis functions that make up any general scale-invariant sequential activity (see Equation 3).

To generate bell-shaped single cell tuning curves, the neurons in the second layer $\tilde{\mathbf{f}}$ compute the inverse Laplace transform of the first layer under the Post approximation (Post (1930)).

$$\tilde{\mathbf{f}} = \mathbf{L}_k \mathbf{F}, \tag{10}$$

where $\mathbf{L}_k$ is a discretized approximation of the inverse Laplace transform of the kth order (Shankar and Howard (2013)).

From Equation 9 and Equation 10, the feedfoward dynamics above is equivalent to a linear recurrent dynamics involving only the second layer neurons $\tilde{\mathbf{f}}$:

$$\frac{d}{dt}\tilde{\mathbf{f}} = \mathbf{L}_k \mathbf{S} \mathbf{L}_k^{-1} \tilde{\mathbf{f}} \equiv \mathbf{M}_{\text{TILT}} \tilde{\mathbf{f}}, \tag{11}$$

where $\mathbf{S}$ is a diagonal matrix consisting of the single cell time constants λ_i's in Equation 9.

Therefore the dynamics of the neurons in the second layer are equivalent to the one generated by a linear recurrent network with connectivity matrix $\mathbf{M}_{\text{TILT}}$. Because the matrix representation of the inverse Laplace transform $\mathbf{L}_k$ is approximated by taking derivatives of nearby nodes, it has the same motif across columns (for details see Shankar and Howard (2013)). Therefore, the model in Shankar and Howard (2012), although a feedforward network, is a special case in the family of linear recurrent networks that can generate scale-invariant sequential activity. It might be the case that that the exponentially decaying basis functions are indeed maintained by a separate population of neurons, and the downstream neurons consititute a "dual" population. Such exponentially decaying cells have recently been identified in lateral entorhinal cortex (Tsao et al. (2018)), whose downstream regions have been identified as locations for time cells (MacDonald et al. (2011); Salz et al. (2016)).

4 Comparison with common network models

Scale-invariance puts stringent constraints on the architecture of recurrent neural networks. To illustrate this, in this section we consider two widely-used neural network models that do not generate scale-invariant sequential activity. Section 4.1 considers a simple chaining model; the following section considers a random network. We will see that the connectivity matrices for these two widely-used models violate the constraints derived above in Section 2 and that these models do not support scale-invariant sequential activity. By means of comparison we will compare these two networks with the two scale-invariant example networks described previously. All the networks in this section are simulated with $N = 20$ neurons.

4.1 Simple Feedforward Chaining Model

The simplest possible model to generate sequential activity is a simple feedforward chaining model. In this subsection we will analyze a simple feedforward chaining model and demonstrate

that a chaining model composed of elements with the same time constant cannot meet the requirements for scale-invariant sequential activity.

In this simple model the activity of the ith neuron in a chain of N neurons obeys

$$\frac{dx_i}{dt} = \begin{cases} -x_i + x_{i-1}, & 1 < i \leq N \\ -x_i, & i = 1 \end{cases} \tag{12}$$

Note that all of these neurons have the same time constant which is here set to one. Because an activation in the first unit spreads gradually across the network, this model generates sequential activity. Each neuron does not respond instantaneously to its input but has some finite integration time resulting in a spread of activity across time. As the activation spreads across the chain, the spread in time accumulates. However, as can be shown via the Central Limit theorem, the sequential activity is not scale-invariant because peak time of the ith unit goes up linearly in i but the width of the peak goes up with $\sqrt{i}$ (Liu et al. (2018)). Here we show that the eigenvalues and eigenvectors of the connectivity matrix of this simple model (Eq. 12) do not satisfy the constraints derived from Section 2.

Figure 4a (left) shows the connectivity matrix of the chaining model described in Eq. 12. This connectivity matrix has a simple motif that repeats across rows; the self-interaction term gives a -1 on the diagonal and the chaining term gives a +1 at the off-diagonal. This connectivity matrix has a single degenerate eigenvalue of -1 (Figure 4a, middle left), violating Constraint 1 that the network must have geometrically spaced eigenvalues. In addition, the connectivity matrix has only a single distinct eigenvector (Figure 4a, middle right). Because each eigenvector is identical, they are not translated versions of one another, violating Constraint 2. Therefore, although the network generates sequential activity (Figure 4a, left top), the activity of different neurons are not rescaled versions of each other (Figure 4a, left bottom).

For contrast, Figure 4b illustrates the connectivity matrix, eigenvalues and eigenvectors for the scale-invariant network that described previously in Section 3.1. In illustrating the connectivity matrix, we have ordered the neurons according to the sequence in which they are activated and we have only included the neurons in the sequence (the same for the complex example below). Recall that this matrix was constructed to obey the two constraints and has already been shown to generate scale-invariant sequential activity. Consequently, by construction, the eigenvalues are geometrically spaced and the eigenvectors are translated versions of one another. Although the connectivity matrix clearly has a rich structure, the rows of the connectivity matrix are certainly not translated versions of one another. The entries above the diagonal tend to be more negative, indicating that the connections from neurons later in the sequence to the neurons earlier in the sequence tend to be more inhibitory than the connections in the opposite direction. It is not at all obvious why this specific connectivity matrix yields scale-invariant sequential activity. This is much more clear from examining the eigenvalues and eigenvectors.

4.2 Random recurrent networks

Nonlinear neural networks with random connectivity matrices are able to generate chaotic activity (Sompolinsky et al. (1988)). With appropriately trained weights, they are also able to generate sequential neural activity similar to that obtained in actual recordings (Rajan et al. (2016)). However, generic random neural networks without training cannot produce scale-invariant, sequential activity due to the conflict with the two constraints derived above in Section 2.

Random neural networks have eigenvalue spectrums that are uniformly distributed inside a unit disc in the complex plane (Rajan and Abbott (2006); Girko (1985)), therefore not geometrically spaced as required by scale-invariance (Section 2.2). We simulated the activity of

an instance of a linear network with a random connectivity matrix and computed its eigenvalues and eigenvectors (Figure 4c). The network dynamics is described by

$$\frac{d\mathbf{x}(t)}{dt} = -\mathbf{x}(t) + \mathbf{M}\mathbf{x}(t) = (\mathbf{M} - \mathbf{I})\mathbf{x}(t), \tag{13}$$

where each element in $\mathbf{M}$ was sampled from a Gaussian distribution with mean 0 and variance $\frac{1}{N}$.

The connectivity matrix $\mathbf{M} - \mathbf{I}$ is shown in Figure 4c (left). The eigenvalues of the connectivity matrix are approximately uniformly distributed in a unit disc centered at (-1,0) (Figure 4c, middle left), confirming the results from Rajan and Abbott (2006) and Girko (1985). The eigenvectors also apparently do not have any translation-invariant structure (Figure 4c, middle right). Consequently, the network activity does not have the scale-invariant property (Figure 4c, right).

In contrast, the network constructed in the same way as in Section 3.2 has complex eigenvalues whose real and imaginary parts are both geometrically distributed (Figure 4d, middle left) and eigenvectors that consist of the same motif (Figure 4d, middle right). Consequently, it allows scale-invariant sequential activity (Figure 4d, right). The connectivity matrix has a structure where the connections from neurons later in the sequence to the ones earlier in the sequence tend to be more inhibitory (Figure 4a, notice neurons 10-18 correspond to the complex conjugate eigenvalues and hence have the same activity as neurons 1-9).

5 Discussion

In this paper we studied the constraints on the connectivity of a linear network for it to generate scale-invariant sequential neural activity. It is analytically shown that two necessary conditions need to hold for the structure of the connectivity matrix. First of all, it must have geometrically-spaced time constants. Second, its eigenvectors must contain the same motif with a translation-invariant structure. Intuitively, the same motif in different eigenvectors pick out different timescales by the same proportion, and the geometric spacing of timescales further ensures the activity of different cells are rescaled with each other. The generated activity is highly dynamic, providing a possibility for the dynamic coding of working memory (Stokes (2015)).

The first constraint we derived requires that the connectivity of the neural network has a spectrum of geometrically spaced eigenvalues. Geometrically spaced network eigenvalues can generically emerge from multiplicative cellular processes (Amir et al. (2012)). Mechanistically, a spectrum of timescales could be generated by positive feedback in recurrent circuits. It could also be an inhearent property of single neurons. We noted in our procedure to generate connectivity matrices that if the eigenvectors are sparse, the spectrum of timescales in the connectivity matrix would come mainly from the single cell time constants (notice the gradient in the diagonal elements of the connectivity matrices in Figure 4b, d). If the eigenvectors are dense, the spectrum of timescales would be a consequence of the network interaction.

In vitro slice experiments and modeling works have shown that a spectrum of slow timescales can be obtained in single cells by utilizing the slow dynamics of calcium-dependent currents (Loewenstein and Sompolinsky (2003); Egorov et al. (2002); Fransén et al. (2002); Mongillo et al. (2008); Tiganj et al. (2015); Liu et al. (2018)). *In vivo* experiments also showed a spectrum of timescales in cortical dynamics, both on the single cell and the population level. Bernacchia et al. (2011) showed that in monkey prefrontal, cingulate and parietal cortex, reward modulates neural activity multiplicatively with a spectrum of time constants (Bernacchia et al. (2011)). Murray et al. (2014) showed that the autocorrelation function of the spontaneous activity of

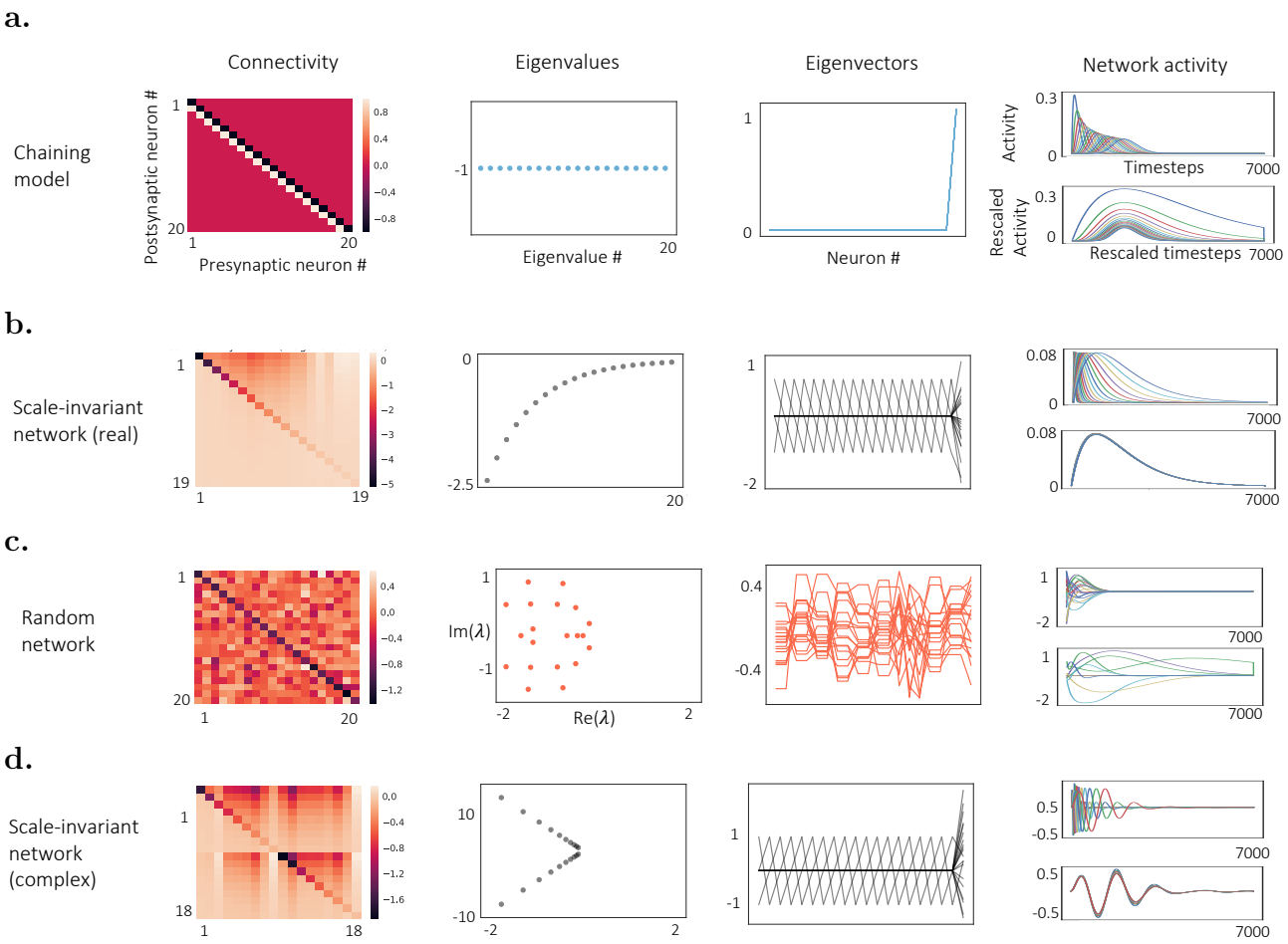


Figure 4: **Comparison with common network models. a, b** Comparison between the simple feedfoward model (**a**) and a network constructed from Section 3.1 that generates scale-invariant sequential activity (scale-invariant network (real), **b**). For the simple feedforward chaining model, the eigenvalues of its connectivity matrix are not geometrically spaced (**a**, middle left) and the eigenvectors do not consist of the same motif (**a**, middle right). Therefore, the network activity does not have the scale-invariant property (**a**, right). In contrast, the scale-invariant network has eigenvalues that are geometrically spaced (**b**, middle left) and eigenvectors that consist of the same motif (**b**, middle right, aside from the boundary effect), resulting in the connectivity that consists of excitation from neurons earlier in the sequence to the ones later in the sequence and inhibition in the opposite direction (**b**, left, only neurons in the sequence are shown. Same below.). Consequently, it allows scale-invariant sequential activity (**b** right). **c, d** Same comparison between an instance of a random recurrent network (**c**) and a network constructed from Section 3.2 (scale-invariant network (complex), **d**). The eigenvalues of the connectivity of a random network are not geometrically spaced (**c**, middle left). The eigenvectors do not consist of the same motif (**c**, middle right). Consequently its activity is not scale-invariant (**c**, right). In contrast, the connectivity of the scale-invariant network has eigenvalues whose real and imaginary parts both form geometric series (**d**, middle left). Its eigenvectors also contain a repeating motif (**d**, middle right). Consequently the scale-invariant network supports scale-invariant sequential activity (**d**, right).

single neurons decay with a spectrum of time constants. The time constants increase from sensory to prefrontal areas (Murray et al. (2014)).

The second constraint we derived requires that the eigenvectors of the connectivity matrix need to consist of the same motif localized at different entries. It is hard to imagine how this kind of translation-invariant eigenvectors could arise generically in neural circuits. However the eigenvectors that satisfy this constraint are a special case of "localized eigenvectors", which have been studied extensively first in condensed matter physics and later in theoretical neuroscience. Anderson first argued that the eigenvectors of matrices where interactions are random and concentrated on the diagonal are exponentially localized (Anderson, P. (1958)). Later numerical and analytical studies confirmed that localized eigenvectors indeed arise in neural networks in the presence of a gradient in the strength of the local interactions (Chaudhuri et al. (2014)) as well as global inhibition, as in the case of ring attractor networks (Tanaka and Nelson (2018)). However, it should be pointed out that the procedures described in this paper do not necessarily generate local interactions where the elements M_{ij} decays with $|i-j|$, as can be seen in Figure 4b, d. Furthermore, our constraint on the eigenvectors is more stringent than only requiring them to be localized — the different localized "patches" need to be the same as well.

In a recent study, Cueva et al. (2019) showed that the cumulative dimensionality of the population activity during working memory increases with a decreasing speed (Cueva et al. (2019)). This is consistent with the network activity generated by linear networks that satisfy the two constraints we derived. Translation-invariant eigenvectors ensure that each eigenmode contributes one unique dimension to the activity. Geometrically spaced eigenvalues ensure that one eigenmode would be suppressed per unit time on a logarithmic scale. Therefore the cumulative dimensionality for the scale-invariant sequential activity described by our model would increase with the logarithm of time. Furthermore, notice that any affine transformation on the state space would not change the cumulative dimensionality. Therefore the same relationship for cumulative dimensionality would hold even if the eigenvectors are not translation-invariant. A geometrically spaced eigenvalues is sufficient to generate linear dynamics whose cumulative dimensionality increases with the logarithm of time.

Goldman (Goldman (2009)) proposed a class of linear recurrent network models that are able to generate sequential activity. In these networks, the feedforward dynamics are constructed by building up feedforward interactions between orthogonal Schur modes and are hidden in the collective network dynamics (Goldman (2009)). Thanks to the hidden feedforward dynamics, the networks can sustain its activity far longer than the timescale constrained by the eigenvalue spectrum of the network. The example studied in Section 4.1 is the simplest model proposed in Goldman (2009). Although this simple model does not allow scale-invariant sequential activity as shown in Section 4.1, it does not indicate that any network constructed in the way in Goldman (2009) cannot generate scale-invariant activity. It is just that eigenvalue decomposition provides a much more clear view to look at the problem of scale-invariance.

Shankar studied a model where the activity of individual nodes can be effectively described by a filtered input through a set of scale-invariant kernel functions with different timescales (Shankar (2015)). It was shown that if this transformation were to be implemented by a network involving only local (possibly non-linear) interactions between nodes corresponding to similar timescales, the forms of the kernel functions are strongly constrained: the only possible form of the network activity are given by a linear combination of inverse Laplace transforms (see Section 3.3). In this work we considered a slightly different problem: the interactions between different nodes were constrained to be linear, but non-local interactions were also allowed since the connectivity matrix was not constrained to be local. In this case, the activity given by the inverse Laplace transform (Equation 11) covers a subspace of all possible solutions. It is still noticeable that these two related approaches converge on similar results.

Acknowledgements

This work was supported by ONR MURI award N00014-16-1-2832, ONR DURIP award N00014-17-1-2304 and NIBIB R01EB022864. .

References

Amir, A., Oreg, Y., and Imry, Y. (2012). On relaxations and aging of various glasses. *Proceedings of the National Academy of Sciences*, 109(6):1850–1855.

Anderson, P., W. (1958). Absence of Diffusion in Certain Random Lattices. *Physical Review*, 109(5):1492–1506.

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):366–72.

Buonomano, D. V. and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113–25.

Chance, F. S., Abbott, L. F., and Reyes, A. D. (2002). Gain modulation from background synaptic input. *Neuron*, 35(4):773–82.

Chaudhuri, R., Bernacchia, A., and Wang, X. J. (2014). A diversity of localized timescales in network activity. *eLife*, 3:e01239.

Cueva, C. J., Marcos, E., Saez, A., Genovesio, A., Jazayeri, M., Romo, R., Salzman, C. D., Shadlen, M. N., and Fusi, S. (2019). Low dimensional dynamics for working memory and time encoding. *bioRxiv*.

Egorov, A. V., Hamam, B. N., Fransén, E., Hasselmo, M. E., and Alonso, A. A. (2002). Graded persistent activity in entorhinal cortex neurons. *Nature*, 420(6912):173–8.

Fransén, E., Alonso, A. A., and Hasselmo, M. E. (2002). Simulations of the role of the muscarinic-activated calcium-sensitive nonspecific cation current INCM in entorhinal neuronal activity during delayed matching tasks. *Journal of Neuroscience*, 22(3):1081–1097.

Gibbon, J., Baldock, M., Locurto, C., Gold, L., and Terrace, H. (1977). Trial and intertrial durations in autoshaping. *Journal of Experimental psychology: Animal behavior processes*, 3(3):264.

Girko, V. L. (1985). Circular law. *Theory of Probability & Its Applications*, 29(4):694–706.

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., and Gretz, A. L. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, 6:355–369.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, 61(4):621–634.

Gütig, R. and Sompolinsky, H. (2009). Time-warp-invariant neuronal processing. *PLoS Biology*, 7(7).

Harvey, C. D., Coen, P., and Tank, D. W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62–8.

Hopfield, J. J. and Brody, C. D. (2000). What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3):1–6.

Howard, M. W., Youker, T. E., and Venkatadass, V. (2008). The persistence of memory: Contiguity effects across several minutes. *Psychonomic Bulletin & Review*, 15(PMC2493616):58–63.

Liu, Y., Tiganj, Z., Hasselmo, M. E., and Howard, M. W. (2018). A neural microcircuit model for a scalable scale-invariant representation of time. *Hippocampus*.

Loewenstein, Y. and Sompolinsky, H. (2003). Temporal integration by calcium dynamics in a model neuron. *Nature Neuroscience*, 6(9):961–7.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., and Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, 71(4):737 – 749.

Machens, C. K., Romo, R., and Brody, C. D. (2010). Functional, But Not Anatomical, Separation of "What" and "When" in Prefrontal Cortex. *Journal of Neuroscience*, 30(1):350–360.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, 319(5869):1543–1546.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64:482–488.

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., et al. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience*, 17(12):1661–1663.

Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, 321(5894):1322–7.

Post, E. (1930). Generalized differentiation. *Transactions of the American Mathematical Society*, 32:723–781.

Rajan, K. and Abbott, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Physical Review Letters*, 97(18):188104.

Rajan, K., Harvey, C. D., and Tank, D. W. (2016). Recurrent Network Models of Sequence Generation and Memory. *Neuron*, 90(1):128–142.

Rakitin, B. C., Gibbon, J., Penny, T. B., Malapani, C., Hinton, S. C., and Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 24:15–33.

Salz, D. M., Tiganj, Z., Khasnabish, S., Kohley, A., Sheehan, D., Howard, M. W., and Eichenbaum, H. (2016). Time cells in hippocampal area CA3. *Journal of Neuroscience*, 36:7476–7484.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.

Shankar, K. H. (2015). Generic construction of scale-invariantly coarse grained memory. *Lecture Notes in Artificial intelligence*, 8955:175–184.

Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, 24(1):134–193.

Shankar, K. H. and Howard, M. W. (2013). Optimally fuzzy temporal memory. *Journal of Machine Learning Research*, 14:3753–3780.

Sompolinsky, H., Crisanti, A., and Sommers, H.-J. (1988). Chaos in random neural networks. *Physical review letters*, 61(3):259.

Stokes, M. G. (2015). 'activity-silent'working memory in prefrontal cortex: a dynamic coding framework. *Trends in cognitive sciences*, 19(7):394–405.

Tanaka, H. and Nelson, D. R. (2018). Non-hermitian quasi-localization and ring attractor neural networks.

Tank, D. and Hopfield, J. (1987). Neural computation by concentrating information in time. *Proceedings of the National Academy of Sciences*, 84(7):1896–1900.

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., and Howard, M. W. (2018). Compressed timeline of recent experience in monkey lateral prefrontal cortex. *Journal of cognitive neuroscience*, pages 1–16.

Tiganj, Z., Hasselmo, M. E., and Howard, M. W. (2015). A simple biophysically plausible model for long time constants in single neurons. *Hippocampus*, 25(1):27–37.

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., and Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, 561(7721):57.

Voelker, A. R. and Eliasmith, C. (2018). Improving spiking dynamical networks: Accurate delays, higher-order synapses, and time cells. *Neural computation*, 30(3):569–609.

Wang, J., Narain, D., Hosseini, E. A., and Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nature neuroscience*, 21(1):102.