

Topographic Organization of Receptive Fields in Recursive Self-Organizing Map

Peter Tiño

*School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK*

Igor Farkaš

*Faculty of Mathematics, Physics and Informatics
Comenius University
Mlynská dolina, 842 48 Bratislava, Slovak Republic*

Jort van Mourik

*Neural Computing Research Group
Aston University
Aston Triangle, Birmingham B4 7ET, UK*

Abstract

Recently, there has been an outburst of interest in extending topographic maps of vectorial data to more general data structures, such as sequences or trees. However, at present, there is no general consensus as to how best to process sequences using topographic maps and this topic remains a very active focus of current neurocomputational research. The representational capabilities and internal representations of the models are not well understood. We rigorously analyze a generalization of the Self-Organizing Map (SOM) for processing sequential

data, Recursive SOM (RecSOM [29]), as a non-autonomous dynamical system consisting of a set of fixed input maps. We argue that contractive fixed input maps are likely to produce Markovian organizations of receptive fields on the RecSOM map. We derive bounds on parameter β (weighting the importance of importing past information when processing sequences) under which contractiveness of the fixed input maps is guaranteed. Some generalizations of SOM contain a dynamic module responsible for processing temporal contexts as an integral part of the model. We show that Markovian topographic maps of sequential data can be produced using a simple fixed (non-adaptable) dynamic module externally feeding a standard topographic model designed to process static vectorial data of fixed dimensionality (e.g. SOM). We elaborate upon the importance of non-Markovian organizations in topographic maps of sequential data.

1 Introduction

In its original form the self-organizing map (SOM) [15] is a nonlinear projection method that maps a high-dimensional metric vector space onto a two-dimensional regular grid in a topologically ordered fashion [16]. Each grid point has an associated codebook vector representing a local subset (Voronoi compartment) of the data space. Neighboring grid points represent neighboring regions of the data space. Given a collection of possibly high-dimensional data points, by associating each point with its codebook representative (and so in effect with its corresponding grid point) a two-dimensional topographic map of the data collection is obtained. Locations of the codebook vectors in the data space are adapted to the layout of data points in an unsupervised learning process. Both competitive learning¹ and co-operative learning² are employed.

¹for each data point there is a competition among the codebook vectors for the right to represent it

²not only the codebook vector that has won the competition to represent a data point is allowed to adapt itself to that point, but so are, albeit to a lesser degree, codebook vectors associated with grid locations topologically close to the winner

Many modifications of the standard SOM have been proposed in the literature (e.g. [31, 18]). Formation of topographic maps via self-organization constitutes an important paradigm in machine learning with many successful applications e.g. in data and web-mining.

Most approaches to topographic map formation operate on the assumption that the data points are members of a finite-dimensional vector space of a fixed dimension. Recently, there has been an outburst of interest in extending topographic maps to more general data structures, such as sequences or trees.

Several modifications of SOM to sequences and/or tree structures have been proposed in the literature ([3] and [7] review most of the approaches). Modified versions of SOM that have enjoyed a great deal of interest equip SOM with *additional feed-back connections* that allow for natural processing of recursive data types. No prior notion of metric on the structured data space is imposed, instead, the similarity measure on structures evolves through parameter modification of the feedback mechanism and recursive comparison of constituent parts of the structured data. Typical examples of such models are Temporal Kohonen Map [2], recurrent SOM [17], feedback SOM [11], recursive SOM [29], merge SOM [23] and SOM for structured data [5]. Other alternatives for constructing topographic maps of structured data have been suggested e.g. in [12, 20, 30, 22].

At present, there is no general consensus as to how best to process sequences with SOMs and this topic remains a very active focus of current neurocomputational research [3, 22, 6]. As pointed out in [6], the representational capabilities of the models are hardly understood. The internal representation of structures within the models is unclear and it is debatable as to which model of recursive unsupervised maps can represent the temporal context of time series in the best way. The first major steps towards the much needed mathematical characterization and analysis of such models were taken in [7, 6]. The authors present the recursive models of unsupervised maps in a unifying framework and study such models from the point of view of internal representations, noise tolerance and topology preservation.

In this paper we continue with the task of mathematical characterization and theoretical analysis of the hidden ‘build-in’ architectural biases for topographic organizations of structured data in the recursive unsupervised maps. Our starting position is viewing such models as non-autonomous dynamical systems with internal dynamics driven by a stream of external inputs. In the line of our recent research, we study the organization of the non-autonomous dynamics on the basis of dynamics of individual fixed-input maps [28]. Recently, we have shown how contractive behavior of the individual fixed-input maps translates to non-autonomous dynamics that organizes the state space in a Markovian fashion: sequences with similar most recent entries tend to have close state-space representations. Longer shared histories of the recently observed items result in closer state-space representations [28, 8, 26].

We concentrate on the Recursive SOM (RecSOM) [29], because RecSOM transcends the simple local recurrence of leaky integrators of earlier models and it has been demonstrated that it can represent much richer dynamical behavior [6].

By studying RecSOM as a non-autonomous dynamical system, we attempt to answer the following questions: Is the architecture of RecSOM naturally biased towards Markovian representations of input streams? If so, under what conditions may Markovian representations occur? How natural are such conditions, i.e. can Markovian organizations of the topographic maps be expected under widely-used architectures and (hyper)parameter settings in RecSOM? What can be gained by having a trainable recurrent part in RecSOM, i.e. how does RecSOM compare with a much simpler setting of SOM operating on a simple *non-trainable* iterative function system with Markovian state-space organization [25]?

The paper has the following organization: We introduce the RecSOM model in section 2 and analyze it rigorously as a non-autonomous dynamical system in section 3. The experiments in section 4 are followed by a discussion in section 5. Section 6 concludes the paper by summarizing the key messages of this study.

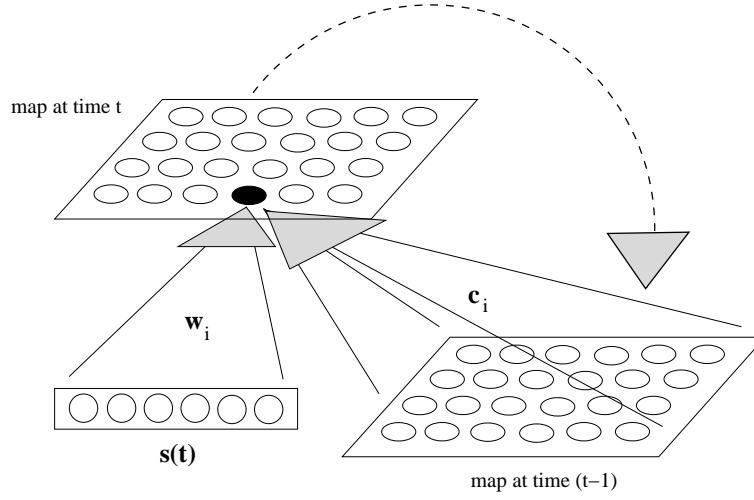


Figure 1: Recursive SOM architecture [29]. The original SOM algorithm is used for both input vector $\mathbf{s}(t)$ and for the context represented as the map activation $\mathbf{y}(t-1)$ from the previous time step. Solid lines represent trainable connections, dashed line represents one-to-one copy of the activity vector \mathbf{y} . The network learns to associate the current input with previous activity states. This way each neuron responds to a sequence of inputs.

2 Recursive Self-Organizing Map - RecSOM

The architecture of the RecSOM model [29] is shown in figure 1. Each neuron $i \in \{1, 2, \dots, N\}$ in the map has two weight vectors associated with it:

- $\mathbf{w}_i \in \mathbb{R}^n$ – linked with an n -dimensional input $\mathbf{s}(t)$ feeding the network at time t
- $\mathbf{c}_i \in \mathbb{R}^N$ – linked with the context

$$\mathbf{y}(t-1) = (y_1(t-1), y_2(t-1), \dots, y_N(t-1))$$

containing map activations $y_i(t-1)$ from the previous time step.

The output of a unit i at time t is computed as

$$y_i(t) = \exp(-d_i(t)), \quad (1)$$

where³

$$d_i(t) = \alpha \cdot \|\mathbf{s}(t) - \mathbf{w}_i\|^2 + \beta \cdot \|\mathbf{y}(t-1) - \mathbf{c}_i\|^2. \quad (2)$$

In eq. (2), $\alpha > 0$ and $\beta > 0$ are model parameters that respectively influence the effect of the input and the context upon neuron's profile. Both weight vectors can be updated using the same form of learning rule [29]:

$$\Delta \mathbf{w}_i = \gamma \cdot h_{ik} \cdot (\mathbf{s}(t) - \mathbf{w}_i), \quad (3)$$

$$\Delta \mathbf{c}_i = \gamma \cdot h_{ik} \cdot (\mathbf{y}(t-1) - \mathbf{c}_i), \quad (4)$$

where k is an index of the best matching unit at time t , $k = \operatorname{argmin}_{i \in \{1, 2, \dots, N\}} d_i(t)$, and $0 < \gamma < 1$ is the learning rate. Note that the best matching ('winner') unit can be equivalently defined as the unit k of the highest activation $y_k(t)$:

$$k = \operatorname{argmax}_{i \in \{1, 2, \dots, N\}} y_i(t). \quad (5)$$

Neighborhood function h_{ik} is a Gaussian (of width σ) on the distance $d(i, k)$ of units i and k in the map:

$$h_{ik} = e^{-\frac{d(i, k)^2}{\sigma^2}}. \quad (6)$$

The 'neighborhood width', σ , linearly decreases in time to allow for forming topographic representation of input sequences.

3 Contractive fixed-input dynamics in RecSOM

In this section we wish to answer the following principal question: Given a fixed RecSOM input \mathbf{s} , under what conditions will the mapping $\mathbf{y}(t) \mapsto \mathbf{y}(t+1)$ become a contraction, so that the *autonomous* RecSOM dynamics is dominated by a unique attractive fixed point? As we shall see, contractive fixed-input dynamics of RecSOM can lead to maps with Markovian representations of temporal contexts.

Under a fixed input vector $\mathbf{s} \in \mathbb{R}^n$, the time evolution (2) becomes

$$d_i(t+1) = \alpha \cdot \|\mathbf{s} - \mathbf{w}_i\|^2 + \beta \cdot \|(e^{-d_1(t)}, e^{-d_2(t)}, \dots, e^{-d_N(t)}) - \mathbf{c}_i\|^2. \quad (7)$$

³ $\|\cdot\|$ denotes the Euclidean norm

After applying a one-to-one coordinate transformation $y_i = e^{-d_i}$, eq. (7) reads

$$y_i(t+1) = e^{-\alpha\|\mathbf{s}-\mathbf{w}_i\|^2} \cdot e^{-\beta\|\mathbf{y}(t)-\mathbf{c}_i\|^2}, \quad (8)$$

where

$$\mathbf{y}(t) = (y_1(t), y_2(t), \dots, y_N(t)) = (e^{-d_1(t)}, e^{-d_2(t)}, \dots, e^{-d_N(t)}).$$

We denote the Gaussian kernel of inverse variance $\eta > 0$, acting on \mathbb{R}^N , by $G_\eta(\cdot, \cdot)$, i.e. for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$,

$$G_\eta(\mathbf{u}, \mathbf{v}) = e^{-\eta\|\mathbf{u}-\mathbf{v}\|^2}. \quad (9)$$

The system of equations (8) can be written in a vector form as

$$\mathbf{y}(t+1) = \mathbf{F}_\mathbf{s}(\mathbf{y}(t)) = (F_{\mathbf{s},1}(\mathbf{y}(t)), \dots, F_{\mathbf{s},N}(\mathbf{y}(t))), \quad (10)$$

where

$$F_{\mathbf{s},i}(\mathbf{y}) = G_\alpha(\mathbf{s}, \mathbf{w}_i) \cdot G_\beta(\mathbf{y}, \mathbf{c}_i), \quad i = 1, 2, \dots, N. \quad (11)$$

Recall that given a fixed input \mathbf{s} , we aim to study the conditions under which the map $\mathbf{F}_\mathbf{s}$ becomes a contraction. Then, by the Banach Fixed Point theorem, the autonomous RecSOM dynamics $\mathbf{y}(t+1) = \mathbf{F}_\mathbf{s}(\mathbf{y}(t))$ will be dominated by a unique attractive fixed point $\mathbf{y}_\mathbf{s} = \mathbf{F}_\mathbf{s}(\mathbf{y}_\mathbf{s})$.

A mapping $\mathbf{F} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is said to be a contraction with contraction coefficient $\rho \in [0, 1)$, if for any $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$,

$$\|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{y}')\| \leq \rho \cdot \|\mathbf{y} - \mathbf{y}'\|. \quad (12)$$

\mathbf{F} is a contraction if there exists $\rho \in [0, 1)$ so that \mathbf{F} is a contraction with contraction coefficient ρ .

Lemma 3.1 *Consider $\mathbf{y}, \mathbf{y}', \mathbf{c} \in \mathbb{R}^N$, $\mathbf{y} \neq \mathbf{y}'$. Denote by $\Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}')$ the $(N-1)$ -dimensional hyperplane orthogonal to $(\mathbf{y} - \mathbf{y}')$ and containing \mathbf{c} . Let $\tilde{\mathbf{c}}$ be the intersection of $\Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}')$ with the line $\omega(\mathbf{y}, \mathbf{y}')$ passing through \mathbf{y}, \mathbf{y}' (see figure 2). Then, for any $\beta > 0$,*

$$\max_{\mathbf{u} \in \Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}')} \{|G_\beta(\mathbf{y}, \mathbf{u}) - G_\beta(\mathbf{y}', \mathbf{u})|\} = |G_\beta(\mathbf{y}, \tilde{\mathbf{c}}) - G_\beta(\mathbf{y}', \tilde{\mathbf{c}})|.$$

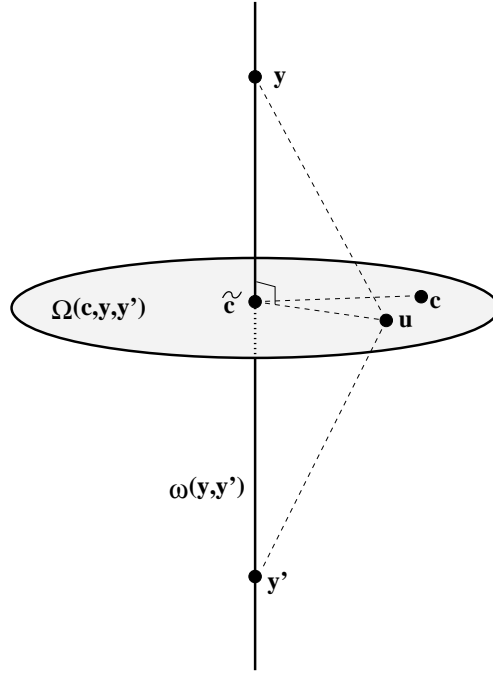


Figure 2: Illustration for the proof of Lemma 3.1. The line $\omega(\mathbf{y}, \mathbf{y}')$ passes through $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$. The $(N-1)$ -dimensional hyperplane $\Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}')$ is orthogonal to $\omega(\mathbf{y}, \mathbf{y}')$ and contains the point $\mathbf{c} \in \mathbb{R}^N$. $\tilde{\mathbf{c}}$ is the orthogonal projection of \mathbf{c} onto $\omega(\mathbf{y}, \mathbf{y}')$, i.e. $\Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}') \cap \omega(\mathbf{y}, \mathbf{y}') = \{\tilde{\mathbf{c}}\}$.

Proof: For any $\mathbf{u} \in \Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}')$,

$$\|\mathbf{y} - \mathbf{u}\|^2 = \|\mathbf{y} - \tilde{\mathbf{c}}\|^2 + \|\mathbf{u} - \tilde{\mathbf{c}}\|^2$$

and

$$\|\mathbf{y}' - \mathbf{u}\|^2 = \|\mathbf{y}' - \tilde{\mathbf{c}}\|^2 + \|\mathbf{u} - \tilde{\mathbf{c}}\|^2.$$

So,

$$\begin{aligned} |G_\beta(\mathbf{y}, \mathbf{u}) - G_\beta(\mathbf{y}', \mathbf{u})| &= |\exp\{-\beta\|\mathbf{y} - \mathbf{u}\|^2\} - \exp\{-\beta\|\mathbf{y}' - \mathbf{u}\|^2\}| \\ &= \exp\{-\beta\|\mathbf{u} - \tilde{\mathbf{c}}\|^2\} \cdot |\exp\{-\beta\|\mathbf{y} - \tilde{\mathbf{c}}\|^2\} - \exp\{-\beta\|\mathbf{y}' - \tilde{\mathbf{c}}\|^2\}| \\ &\leq |\exp\{-\beta\|\mathbf{y} - \tilde{\mathbf{c}}\|^2\} - \exp\{-\beta\|\mathbf{y}' - \tilde{\mathbf{c}}\|^2\}|, \end{aligned}$$

with equality if and only if $\mathbf{u} = \tilde{\mathbf{c}}$.

Q.E.D.

Lemma 3.2 Consider any $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$, $\mathbf{y} \neq \mathbf{y}'$ and the line $\omega(\mathbf{y}, \mathbf{y}')$ passing through \mathbf{y}, \mathbf{y}' . Let $\bar{\omega}(\mathbf{y}, \mathbf{y}')$ be the line $\omega(\mathbf{y}, \mathbf{y}')$ without the segment connecting \mathbf{y} and \mathbf{y}' , i.e.

$$\bar{\omega}(\mathbf{y}, \mathbf{y}') = \{\mathbf{y} + \kappa \cdot (\mathbf{y} - \mathbf{y}') \mid \kappa \in (-\infty, -1] \cup [0, \infty)\}$$

Then, for any $\beta > 0$,

$$\operatorname{argmax}_{\mathbf{c} \in \omega(\mathbf{y}, \mathbf{y}')} \{|G_\beta(\mathbf{y}, \mathbf{c}) - G_\beta(\mathbf{y}', \mathbf{c})|\} \in \bar{\omega}(\mathbf{y}, \mathbf{y}').$$

Proof: For $0 < \kappa \leq \frac{1}{2}$, consider two points

$$\mathbf{c}(-\kappa) = \mathbf{y} - \kappa \cdot (\mathbf{y} - \mathbf{y}') \quad \text{and} \quad \mathbf{c}(\kappa) = \mathbf{y} + \kappa \cdot (\mathbf{y} - \mathbf{y}').$$

Let $\delta = \|\mathbf{y} - \mathbf{y}'\|$. Then,

$$G_\beta(\mathbf{y}, \mathbf{c}(-\kappa)) = G_\beta(\mathbf{y}, \mathbf{c}(\kappa)) = e^{-\beta\delta^2\kappa^2}$$

and

$$G_\beta(\mathbf{y}', \mathbf{c}(\kappa)) = e^{-\beta\delta^2(1+\kappa)^2} < e^{-\beta\delta^2(1-\kappa)^2} = G_\beta(\mathbf{y}', \mathbf{c}(-\kappa)).$$

Hence,

$$G_\beta(\mathbf{y}, \mathbf{c}(\kappa)) - G_\beta(\mathbf{y}', \mathbf{c}(\kappa)) > G_\beta(\mathbf{y}, \mathbf{c}(-\kappa)) - G_\beta(\mathbf{y}', \mathbf{c}(-\kappa)).$$

A symmetric argument can be made for the case

$$\mathbf{c}(-\kappa) = \mathbf{y}' - \kappa \cdot (\mathbf{y}' - \mathbf{y}), \quad \mathbf{c}(\kappa) = \mathbf{y}' + \kappa \cdot (\mathbf{y}' - \mathbf{y}), \quad 0 < \kappa \leq \frac{1}{2}.$$

It follows that for every⁴ $\mathbf{c}_- \in \omega(\mathbf{y}, \mathbf{y}') \setminus \bar{\omega}(\mathbf{y}, \mathbf{y}')$ in between the points \mathbf{y} and \mathbf{y}' , there exist a $\mathbf{c}_+ \in \bar{\omega}(\mathbf{y}, \mathbf{y}')$ such that

$$|G_\beta(\mathbf{y}, \mathbf{c}_+) - G_\beta(\mathbf{y}', \mathbf{c}_+)| > |G_\beta(\mathbf{y}, \mathbf{c}_-) - G_\beta(\mathbf{y}', \mathbf{c}_-)|.$$

Q.E.D.

For $\beta > 0$, define a function $H_\beta : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}$,

$$H_\beta(\kappa, \delta) = e^{\beta\delta^2(2\kappa+1)} - \frac{1}{\kappa} - 1. \tag{13}$$

⁴ $A \setminus B$ is the set of elements in A not contained in B

Theorem 3.3 Consider $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$, $\|\mathbf{y} - \mathbf{y}'\| = \delta > 0$. Then, for any $\beta > 0$,

$$\operatorname{argmax}_{\mathbf{c} \in \mathbb{R}^N} \{|G_\beta(\mathbf{y}, \mathbf{c}) - G_\beta(\mathbf{y}', \mathbf{c})|\} \in \{\mathbf{c}_{\beta,1}(\delta), \mathbf{c}_{\beta,2}(\delta)\},$$

where

$$\mathbf{c}_{\beta,1}(\delta) = \mathbf{y} + \kappa_\beta(\delta) \cdot (\mathbf{y} - \mathbf{y}'), \quad \mathbf{c}_{\beta,2}(\delta) = \mathbf{y}' + \kappa_\beta(\delta) \cdot (\mathbf{y}' - \mathbf{y})$$

and $\kappa_\beta(\delta) > 0$ is implicitly defined by

$$H_\beta(\kappa_\beta(\delta), \delta) = 0. \quad (14)$$

Proof: By Lemma 3.1, when maximizing $|G_\beta(\mathbf{y}, \mathbf{c}) - G_\beta(\mathbf{y}', \mathbf{c})|$, we should locate \mathbf{c} on the line $\omega(\mathbf{y}, \mathbf{y}')$ passing through \mathbf{y} and \mathbf{y}' . By Lemma 3.2, we should concentrate only on $\bar{\omega}(\mathbf{y}, \mathbf{y}')$, i.e. on points outside the line segment connecting \mathbf{y} and \mathbf{y}' .

Consider points on the line segment

$$\{\mathbf{c}(\kappa) = \mathbf{y} + \kappa \cdot (\mathbf{y} - \mathbf{y}') \mid \kappa \geq 0\}.$$

Parameter $\kappa > 0$, such that $\mathbf{c}(\kappa)$ maximizes $|G_\beta(\mathbf{y}, \mathbf{c}) - G_\beta(\mathbf{y}', \mathbf{c})|$, can be found by maximizing

$$g_{\beta,\delta}(\kappa) = e^{-\beta\delta^2\kappa^2} - e^{-\beta\delta^2(\kappa+1)^2}. \quad (15)$$

Setting the derivative of $g_{\beta,\delta}(\kappa)$ (with respect to κ) to zero results in

$$e^{-\beta\delta^2(\kappa+1)^2}(\kappa+1) - e^{-\beta\delta^2\kappa^2}\kappa = 0, \quad (16)$$

which is equivalent to

$$e^{-\beta\delta^2(2\kappa+1)} = \frac{\kappa}{\kappa+1}. \quad (17)$$

Note that κ in (17) cannot be zero, as for finite positive β and δ , $e^{-\beta\delta^2(2\kappa+1)} > 0$.

Hence, it is sufficient to concentrate only on the line segment

$$\{\mathbf{c}(\kappa) = \mathbf{y} + \kappa \cdot (\mathbf{y} - \mathbf{y}') \mid \kappa > 0\}.$$

It is easy to see that $\kappa_\beta(\delta) > 0$ satisfying (17) also satisfies $H_\beta(\kappa_\beta(\delta), \delta) = 0$. Moreover, for a given $\beta > 0$, $\delta > 0$, there is a *unique* $\kappa_\beta(\delta) > 0$ given by

$H_\beta(\kappa_\beta(\delta), \delta) = 0$. In other words, the function $\kappa_\beta(\delta)$ is one-to-one. To see this, note that $e^{\beta\delta^2(2\kappa+1)}$ is an increasing function of $\kappa > 0$ with range $(e^{\beta\delta^2}, \infty)$, while $1 + \frac{1}{\kappa}$ is a decreasing function of $\kappa > 0$ with range $(\infty, 1)$.

The second derivative of $g_{\beta,\delta}(\kappa)$ is (up to a positive scaling constant $\frac{1}{2\beta\delta^2}$) equal to:

$$e^{-\beta\delta^2(\kappa+1)^2} [1 - 2\beta\delta^2(\kappa+1)^2] - e^{-\beta\delta^2\kappa^2} [1 - 2\beta\delta^2\kappa^2] \quad (18)$$

which can be rearranged as

$$\left[e^{-\beta\delta^2(\kappa+1)^2} - e^{-\beta\delta^2\kappa^2} \right] - 2\beta\delta^2 \left[e^{-\beta\delta^2(\kappa+1)^2} (\kappa+1)^2 - e^{-\beta\delta^2\kappa^2} \kappa^2 \right]. \quad (19)$$

The first term in (19) is negative, as for $\kappa > 0$, $e^{-\beta\delta^2(\kappa+1)^2} < e^{-\beta\delta^2\kappa^2}$. We will show that the second term, evaluated at $\kappa_\beta(\delta) = K$, is also negative. To that end, note that by (16),

$$e^{-\beta\delta^2(K+1)^2} (K+1)^2 - e^{-\beta\delta^2K^2} K(K+1) = 0.$$

But because $e^{-\beta\delta^2K^2} K > 0$, we have

$$e^{-\beta\delta^2(K+1)^2} (K+1)^2 - e^{-\beta\delta^2K^2} K^2 > 0,$$

and so

$$-2\beta\delta^2 \left[e^{-\beta\delta^2(K+1)^2} (K+1)^2 - e^{-\beta\delta^2K^2} K^2 \right]$$

is negative.

Because the second derivative of $g_{\beta,\delta}(\kappa)$ at the extremum point $\kappa_\beta(\delta)$ is negative, the unique solution $\kappa_\beta(\delta)$ of $H_\beta(\kappa_\beta(\delta), \delta) = 0$ yields the point $\mathbf{c}_{\beta,1}(\delta) = \mathbf{y} + \kappa_\beta(\delta) \cdot (\mathbf{y} - \mathbf{y}')$ that maximizes $|G_\beta(\mathbf{y}, \mathbf{c}) - G_\beta(\mathbf{y}', \mathbf{c})|$.

Arguments concerning the point $\mathbf{c}_{\beta,2}(\delta) = \mathbf{y}' + \kappa_\beta(\delta) \cdot (\mathbf{y}' - \mathbf{y})$ can be made along the same lines by considering points on the line segment

$$\{\mathbf{c}(\kappa) = \mathbf{y}' + \kappa \cdot (\mathbf{y}' - \mathbf{y}) \mid \kappa > 0\}.$$

Q.E.D.

Lemma 3.4 For all $k > 0$,

$$\frac{(1+2k)}{2(1+k)^2} < \log\left(\frac{1+k}{k}\right) < \frac{(1+2k)}{2k^2}. \quad (20)$$

Proof: Consider the functions

$$\begin{aligned} \Delta_1(k) &= \frac{(1+2k)}{2k^2} - \log\left(\frac{1+k}{k}\right) \\ \Delta_2(k) &= \log\left(\frac{1+k}{k}\right) - \frac{(1+2k)}{2(1+k)^2} \end{aligned}$$

We find that

$$\begin{aligned} \lim_{k \rightarrow 0} \Delta_1(k) &\simeq 1/2k^2 + \log(k) > 0, & \lim_{k \rightarrow \infty} \Delta_1(k) &\simeq 1/k^2 > 0 \\ \lim_{k \rightarrow 0} \Delta_2(k) &\simeq -\log(k) > 0, & \lim_{k \rightarrow \infty} \Delta_2(k) &\simeq 1/k^2 > 0. \end{aligned}$$

Since

$$\begin{aligned} \Delta'_1(k) &= -\frac{(1+2k)}{k^3(1+k)} < 0, \\ \Delta'_2(k) &= -\frac{(1+2k)}{k(1+k)^3} < 0, \end{aligned}$$

both functions $\Delta_1(k)$ and $\Delta_2(k)$ are monotonically decreasing positive functions of $k > 0$. *Q.E.D.*

Lemma 3.5 For $\beta > 0$, consider a function $D_\beta : \mathbb{R}^+ \rightarrow (0, 1)$,

$$D_\beta(\delta) = g_{\beta,\delta}(\kappa_\beta(\delta)), \quad (21)$$

where $g_{\beta,\delta}(\kappa)$ is defined in (15) and $\kappa_\beta(\delta)$ is implicitly defined by (13) and (14).

Then, D_β has the following properties:

1. $D_\beta > 0$,
2. $\lim_{\delta \rightarrow 0^+} D_\beta(\delta) = 0$,
3. D_β is a continuous monotonically increasing concave function of δ .
4. $\lim_{\delta \rightarrow 0^+} \frac{dD_\beta(\delta)}{d\delta} = \sqrt{\frac{2\beta}{e}}$.

Proof:

To simplify the presentation, we do not write subscript β when referring to quantities such as D_β , $\kappa_\beta(\delta)$ etc.

1. Since $\kappa(\delta) > 0$ for any $\delta > 0$,

$$D(\delta) = e^{-\beta\delta^2\kappa(\delta)^2} - e^{-\beta\delta^2(\kappa(\delta)+1)^2} > 0.$$

2. Even though the function $\kappa(\delta)$ is known only implicitly through (13) and (14), the inverse function, $\delta(\kappa)$, can be obtained explicitly from (13)–(14) as

$$\delta(\kappa) = \sqrt{\frac{\log(\frac{1+\kappa}{\kappa})}{\beta(1+2\kappa)}}. \quad (22)$$

Now, $\delta(\kappa)$ is a monotonically decreasing function. This is easily verified, as the derivative of $\delta(\kappa)$,

$$\delta'(\kappa) = -\frac{(1+2\kappa+2\kappa(1+\kappa)\log(\frac{1+\kappa}{\kappa}))}{2\kappa(1+\kappa)(1+2\kappa)^2\sqrt{\frac{\beta\log(\frac{1+\kappa}{\kappa})}{(1+2\kappa)}}} \quad (23)$$

is negative for all $\kappa > 0$.

Both $\kappa(\delta)$ and $\delta(\kappa)$ are one-to-one (see also proof of theorem 3.3). Moreover, $\delta(\kappa) \rightarrow 0$ as $\kappa \rightarrow \infty$, meaning that $\kappa(\delta) \rightarrow \infty$ as $\delta \rightarrow 0^+$. Hence, $\lim_{\delta \rightarrow 0^+} D(\delta) = 0$.

3. Since $\delta(\kappa)$ is a continuous function of κ , $\kappa(\delta)$ is continuous in δ . Because $e^{-\beta\delta^2\kappa(\delta)^2} - e^{-\beta\delta^2(\kappa(\delta)+1)^2}$ is continuous in $\kappa(\delta)$, $D(\delta)$ is a continuous function of δ .

Because of the relationship between δ and $\kappa(\delta)$, we can write the derivatives $\frac{dD(\delta)}{d\delta}$ and $\frac{d^2D(\delta)}{d\delta^2}$ *explicitly*, changing the independent variable from δ to κ . Instead of $D(\delta)$, we will work with the corresponding function of κ , $\mathcal{D}(\kappa)$, such that

$$D(\delta) = \mathcal{D}(\kappa(\delta)). \quad (24)$$

Given a $\kappa > 0$ (uniquely determining $\delta(\kappa)$), we have (after some manipulations),

$$D(\delta(\kappa)) = \mathcal{D}(\kappa) = \frac{1}{(1+\kappa)} \left(\frac{\kappa}{1+\kappa} \right)^{\frac{\kappa^2}{(1+2\kappa)}}. \quad (25)$$

Since $\delta(\kappa)$ and $\kappa(\delta)$ are inverse functions of each other, their first- and second-order derivatives are related through

$$\kappa'(\delta) = \frac{1}{\delta'(k)}, \quad (26)$$

$$\kappa''(\delta) = \frac{-\delta''(k)}{(\delta'(k))^3}, \quad (27)$$

where $k = \kappa(\delta)$.

Furthermore, we have that

$$D' = \frac{dD(\delta)}{d\delta} = \frac{d\mathcal{D}(\kappa)}{d\kappa} \frac{d\kappa(\delta)}{d\delta} = \mathcal{D}' \kappa' \quad (28)$$

and

$$\begin{aligned} D'' = \frac{d^2 D}{d\delta^2} &= \frac{d}{d\delta} \left(\frac{d\mathcal{D}}{d\kappa} \frac{d\kappa}{d\delta} \right) \\ &= \frac{d^2 \mathcal{D}}{d\kappa^2} \left(\frac{d\kappa}{d\delta} \right)^2 + \frac{d\mathcal{D}}{d\kappa} \frac{d^2 \kappa}{d\delta^2} = \mathcal{D}'' \kappa'^2 + \mathcal{D}' \kappa''. \end{aligned} \quad (29)$$

Using (25)–(29), we arrive at derivatives of $D(\delta)$ with respect to δ , expressed as functions of k^5 :

$$\frac{dD}{d\delta}(k) = \frac{\mathcal{D}'(k)}{\delta'(k)}, \quad (30)$$

$$\frac{d^2 D}{d\delta^2}(k) = \frac{1}{\delta'(k)^3} (\mathcal{D}''(k) \delta'(k) - \mathcal{D}'(k) \delta''(k)). \quad (31)$$

The derivatives (30) and (31) can be calculated explicitly, and evaluated for all $k > 0$. After simplification, $\frac{dD}{d\delta}(k)$ and $\frac{d^2 D}{d\delta^2}(k)$ read

$$2\beta(1+k) \left(\frac{1+k}{k} \right)^{\frac{-(1+k)^2}{(1+2k)}} \sqrt{\frac{\log(\frac{1+k}{k})}{\beta(1+2k)}} \quad (32)$$

and

$$2\beta \frac{(1+k)}{(1+2k)} \left(\frac{1+k}{k} \right)^{\frac{-(1+k)^2}{(1+2k)}} \frac{[1+2k-2k^2 \log(\frac{1+k}{k})] [1+2k-2(1+k)^2 \log(\frac{1+k}{k})]}{[1+2k+2k(1+k) \log(\frac{1+k}{k})]}, \quad (33)$$

respectively.

⁵ $k > 0$ is related to δ through $k = \kappa(\delta)$

Clearly, $\frac{dD}{d\delta}(k) > 0$ for all $\beta > 0$ and $k > 0$. To show that $\frac{d^2D}{d\delta^2}(k) < 0$, recall that by lemma 3.4,

$$\frac{(1+2k)}{2(1+k)^2} < \log\left(\frac{1+k}{k}\right) < \frac{(1+2k)}{2k^2}$$

for all $k > 0$, and so

$$\begin{aligned} 1+2k-2k^2 \log\left(\frac{1+k}{k}\right) &> 0, \\ 1+2k-2(1+k)^2 \log\left(\frac{1+k}{k}\right) &< 0. \end{aligned}$$

All the other factors in (33) are positive.

4. Considering only the leading terms as $\delta \rightarrow 0$ ($k \rightarrow \infty$), we have

$$\lim_{\delta \rightarrow 0^+} \frac{dD}{d\delta}(k) \simeq \sqrt{\frac{2\beta}{e}} + \mathcal{O}\left(\frac{1}{k^2}\right),$$

and so

$$\lim_{\delta \rightarrow 0^+} \frac{dD_\beta(\delta)}{d\delta} = \sqrt{\frac{2\beta}{e}}.$$

Q.E.D.

Denote by $\mathbf{G}_\alpha(\mathbf{s})$ the collection of activations coming from the feed-forward part of RecSOM,

$$\mathbf{G}_\alpha(\mathbf{s}) = (G_\alpha(\mathbf{s}, \mathbf{w}_1), G_\alpha(\mathbf{s}, \mathbf{w}_2), \dots, G_\alpha(\mathbf{s}, \mathbf{w}_N)). \quad (34)$$

Theorem 3.6 *Consider an input $\mathbf{s} \in \mathbb{R}^M$. If for some $\rho \in [0, 1)$,*

$$\beta \leq \rho^2 \frac{e}{2} \|\mathbf{G}_\alpha(\mathbf{s})\|^{-2}, \quad (35)$$

then the mapping $\mathbf{F}_\mathbf{s}$ (eqs. (10) and (11)) is a contraction with contraction coefficient ρ .

Proof: Recall that $\mathbf{F}_\mathbf{s}$ is a contractive mapping with contraction coefficient $0 \leq \rho < 1$ if for any $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$,

$$\|\mathbf{F}_\mathbf{s}(\mathbf{y}) - \mathbf{F}_\mathbf{s}(\mathbf{y}')\| \leq \rho \cdot \|\mathbf{y} - \mathbf{y}'\|.$$

This is equivalent to saying that for any \mathbf{y}, \mathbf{y}' ,

$$\|\mathbf{F}_s(\mathbf{y}) - \mathbf{F}_s(\mathbf{y}')\|^2 \leq \rho^2 \cdot \|\mathbf{y} - \mathbf{y}'\|^2,$$

which can be rephrased as

$$\sum_{i=1}^N G_{2\alpha}(\mathbf{s}, \mathbf{w}_i) \cdot (G_\beta(\mathbf{y}, \mathbf{c}_i) - G_\beta(\mathbf{y}', \mathbf{c}_i))^2 \leq \rho^2 \cdot \|\mathbf{y} - \mathbf{y}'\|^2, \quad (36)$$

For given \mathbf{y}, \mathbf{y}' , $\|\mathbf{y} - \mathbf{y}'\| = \delta > 0$, let us consider the worst case scenario with respect to the position of the context vectors \mathbf{c}_i , so that the bound (36) still holds. By theorem 3.3, when maximizing the left hand side of (36), we should locate \mathbf{c}_i on the line passing through \mathbf{y} and \mathbf{y}' , at either

$$\mathbf{c}_{\beta,1}(\delta) = \mathbf{y} + \kappa_\beta(\delta) \cdot (\mathbf{y} - \mathbf{y}'),$$

or

$$\mathbf{c}_{\beta,2}(\delta) = \mathbf{y}' + \kappa_\beta(\delta) \cdot (\mathbf{y}' - \mathbf{y}),$$

where $\kappa_\beta(\delta)$ is implicitly defined by $H_\beta(\kappa_\beta(\delta), \delta) = 0$. In that case, we have

$$|G_\beta(\mathbf{y}, \mathbf{c}_{\beta,j}(\delta)) - G_\beta(\mathbf{y}', \mathbf{c}_{\beta,j}(\delta))| = D_\beta(\delta), \quad j = 1, 2.$$

Since $D_\beta(\delta)$ is a continuous concave function on $\delta > 0$ and $\lim_{\delta \rightarrow 0^+} D_\beta(\delta) = 0$, with $\lim_{\delta \rightarrow 0^+} \frac{dD_\beta(\delta)}{d\delta} = \sqrt{\frac{2\beta}{e}}$, we have the following upper bound:

$$D_\beta(\delta) \leq \delta \sqrt{\frac{2\beta}{e}}. \quad (37)$$

Applying (37) to (36), we get that if

$$\delta^2 \frac{2\beta}{e} \sum_{i=1}^N G_{2\alpha}(\mathbf{s}, \mathbf{w}_i) \leq \rho^2 \delta^2, \quad (38)$$

then \mathbf{F}_s will be a contraction with contraction coefficient ρ .

Inequality (38) is equivalent to

$$\frac{2\beta}{e} \|\mathbf{G}_\alpha(\mathbf{s})\|^2 \leq \rho^2. \quad (39)$$

Q.E.D.

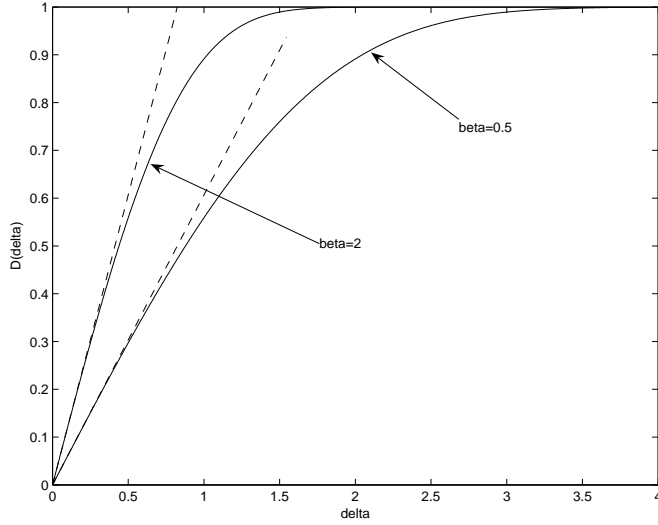


Figure 3: Functions $D_\beta(\delta)$ for $\beta = 0.5$ and $\beta = 2$ (solid lines). Also shown (dashed lines) are the linear upper bounds (37).

Corollary 3.7 *Consider a RecSOM fed by a fixed input \mathbf{s} . Define*

$$\Upsilon(\mathbf{s}) = \frac{e}{2} \|\mathbf{G}_\alpha(\mathbf{s})\|^{-2}. \quad (40)$$

Then, if $\beta < \Upsilon(\mathbf{s})$, $\mathbf{F}_\mathbf{s}$ is a contractive mapping.

We conclude the section by mentioning that we empirically verified validity of the analytical bound (37) for a wide range of values of β , $10^{-2} \leq \beta \leq 5$. For each β , the values of $\kappa_\beta(\delta)$ were numerically calculated on a fine grid of δ -values from the interval $(0, 6)$. These values were then used to plot functions $D_\beta(\delta)$ and to numerically estimate the limit of the first derivative of $D_\beta(\delta)$ as $\delta \rightarrow 0^+$. Numerically determined values matched perfectly the analytical calculations. As an illustration, we show in figure 3 functions $D_\beta(\delta)$ for $\beta = 0.5$ and $\beta = 2$ (solid lines). Also shown (dashed lines) are the linear upper bounds (37).

4 Experiments

In this section we demonstrate and analyze (using the results of section 3) the potential of RecSOM for creating Markovian context representations on three

types of sequences of different nature and complexity: stochastic automaton, laser data and natural language. The first and the third data sets were also used in [29].

In order to get an insight about the benefit of having a *trainable* recurrent part in RecSOM, we also compare RecSOM with standard SOM operating on Markovian suffix-based vector representations of fixed dimensionality obtained from a simple *non-trainable* iterative function system [25].

4.1 Stochastic automaton

The first input series was a binary sequence of 300,000 symbols generated by a first-order Markov chain over the alphabet $\{a, b\}$, with transition probabilities $P(a|b) = 0.3$ and $P(b|a) = 0.4$ [29]. Attempting to replicate Voegtlin’s results, we used RecSOM with 10×10 neurons and one-dimensional coding of input symbols: $a = 0$, $b = 1$. We chose RecSOM parameters from the stable region on the stability map evaluated by Voegtlin for this particular stochastic automaton [29]: $\alpha = 2$ and $\beta = 1$. The learning rate was set to $\gamma = 0.1$. To allow for map ordering, we allow the neighborhood width, σ (see eq. (6)), to linearly decrease from 5.0 to 0.5 during the first 200,000 iterations (ordering phase), and then keep it constant over the next 100,000 iterations (fine-tuning phase)⁶.

Weights of the trained RecSOM (after 300,000 iterations) are shown in figure 4. Input weights (left) are topographically organized in two regions, representing the two input symbols. Context weights of all neurons have a unimodal character and are topographically ordered with respect to the peak position (mode).

We have also constructed a map of the neurons’ receptive fields (RFs)

⁶Voegtlin did not consider reducing the neighborhood size. However, we found that the decreasing neighborhood width was crucial for topographic ordering. Initially small σ did not lead to global ordering of weights. This should not be surprising, since for $\sigma = 0.5$ (used in [29]), the value h_{ik} of the neighborhood function for the nearest neighbor is only $\exp(-1/.5^2) = 0.0183$ (considering a squared grid of neurons with mesh size 1). Decreasing σ is also important in standard SOM.

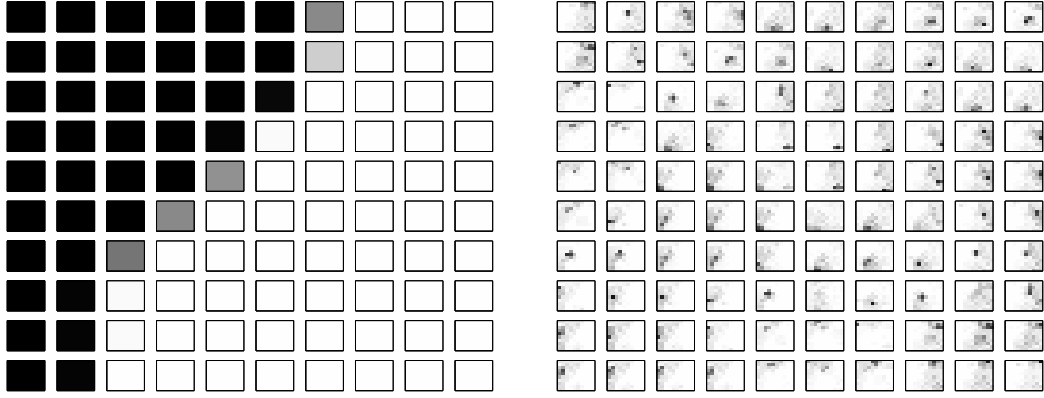


Figure 4: Converged input (left) and context (right) weights after training on the stochastic automaton. All values are between 0 (code of 'a' – white) and 1 (code of 'b' – black). Input weights can be clustered into two groups corresponding to the two input symbols, with a few intermediate units at the boundary. Topographic organization of the context weights is also clearly evident.

(shown in figure 5). Following [29], RF of a neuron is defined as the common suffix of all sequences for which that neuron becomes the best-matching unit. It is evident that the RFs are topographically ordered with respect to the most recent symbols. This map is consistent with the input weights of the neurons (left part of figure 4), when considering only the last symbol.

For these RFs, we computed the quantizer depth (according to [29]), which quantifies the amount of memory captured by the map. It is defined as

$$\bar{n} = \sum_{i=1}^N p_i n_i, \quad (41)$$

where p_i is the probability of the RF of neuron i , and n_i is its length. Due to the constraints imposed by topographic organization of RFs, the quantizer depth of the map, $\bar{n} = 6.18$, is smaller than that of the theoretically optimal (unconstrained) quantizer computed by Voegtlin as $\bar{n} = 7.08$ [29].

The RecSOM model can be considered a nonautonomous dynamical system driven by the external input stream (in this case, sequences over an alphabet of two input symbols 'a' and 'b'). In order to investigate the fixed-input

baaab	aaaab	aaaab	aaaab	bab	aabab		bbaaa	bbaaa	bbaaa
abaab	aaaab	aaaab	bbaab	bbaab	babab		bbaaa	babaa	babaa
ababb	b	bbab	bbbab	aaaab	aaaab	aaaaa	aaaaa	aabaa	aabaa
baabb	aaabb	bbab	abbbb	baaab	baaaa	aaaaa	aaaaa	aaaaa	aaaaa
aaabb	aaabb	abbbb	abbbb		baaaa	aaaaa	aaaaa	aaaaa	aaaaa
bbabb	bbbbb	bbbbb		bba		abbaa		aaaaa	aaaaa
bbbbb	bbbbb		bbbba	abbba	abbaa	bbbaa	bbbaa	aaaaa	aaaaa
babbb	bbbbb	bbbba	bbbba	bbbba		bbbaa	bbbaa		aaaaa
aabbb	babbb	babba	babba	bbaba	aaaba	baaba	abaaa	abaaa	abaaa
aabbb	bb	aabba	aabba	baaba	aaaba	baba	baaaa	baaaa	baaaa

Figure 5: Receptive fields of RecSOM trained on the stochastic two-state automaton. Topographic organization is observed with respect to most recent symbols (only 5 symbols are shown for clarity).

dynamics (10) of the mappings⁷ \mathbf{F}_a and \mathbf{F}_b for symbols 'a' and 'b', respectively, we randomly (with uniform distribution) initialized context activations $\mathbf{y}(0)$ in 10,000 different positions within the state space $(0, 1]^N$. For each initial condition $\mathbf{y}(0)$, we checked asymptotic dynamics of the fixed input maps \mathbf{F}_s , $s \in \{a, b\}$, by monitoring L_2 -norm of the activation differences $(\mathbf{y}(t) - \mathbf{y}(t-1))$ and recording the limit set (after 1000 iterations). Both autonomous dynamics

⁷We slightly abuse the mathematical notation here. As arguments of the bounds, we write the actual input symbols, rather than their vector encodings \mathbf{s} .

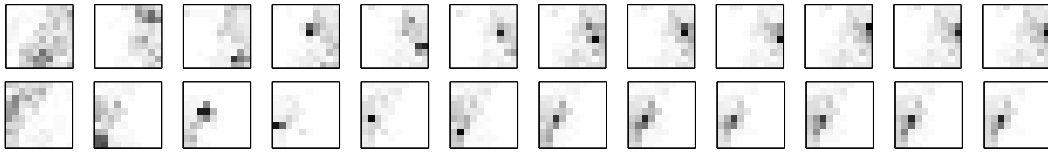


Figure 6: Fixed-input dynamics of RecSOM trained on the stochastic automaton – symbol 'a' (1st row), symbol 'b' (2nd row). The activations settle to a stable unimodal profile after roughly 10 iterations.

settle down in the respective unique attractive fixed points $\mathbf{y}_a = \mathbf{F}_a(\mathbf{y}_a)$ and $\mathbf{y}_b = \mathbf{F}_b(\mathbf{y}_b)$. An example of the fixed-input dynamics is displayed in figure 6. Both autonomous systems settle in the fixed points in roughly 10 iterations. Note the unimodal profile of the fixed points.

It is important to appreciate how the character of the RecSOM fixed-input dynamics (10) for each individual input symbol shapes the overall organization of the map. For each input symbol $s \in \{a, b\}$, the autonomous dynamics $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$ induces a dynamics of the winner units on the map:

$$\begin{aligned} i_s(t) &= \operatorname{argmax}_{i \in \{1, 2, \dots, N\}} y_i(t) \\ &= \operatorname{argmax}_{i \in \{1, 2, \dots, N\}} F_{s,i}(\mathbf{y}(t-1)). \end{aligned} \quad (42)$$

The dynamics (42) is illustrated in figure 7. For each of the 10,000 initial conditions $\mathbf{y}(0)$, we first let the system (10) settle down by preiterating it for 1000 iterations and then mark the map position of the winner units $i_s(t)$ for further 100 iterations. As the fixed-input dynamics for $s \in \{a, b\}$ is dominated by the unique attractive fixed point \mathbf{y}_s , the induced dynamics on the map, (42), settles down in neuron i_s , corresponding to the mode of \mathbf{y}_s :

$$i_s = \operatorname{argmax}_{i \in \{1, 2, \dots, N\}} y_{s,i}. \quad (43)$$

The neuron i_s will be most responsive to input subsequences ending with long blocks of symbols s . As seen in figure 5, receptive fields of other neurons on the map are organized with respect to the closeness of the neurons to the fixed input winners i_a and i_b . Such an organization follows from the attractive fixed point behaviour of the individual maps \mathbf{F}_a , \mathbf{F}_b , and the unimodal character of their fixed points \mathbf{y}_a and \mathbf{y}_b . As soon as symbol s is seen, the mode of the activation profile \mathbf{y} drifts towards the neuron i_s . The more consecutive symbols s we see, the more dominant the attractive fixed point of \mathbf{F}_s becomes and the closer the winner position is to i_s .

This mechanism for creating suffix-based RF organization is reminiscent of the Markovian fractal subsequence representations used in [25] to build Markov models with context dependent length. In the next subsection we compare

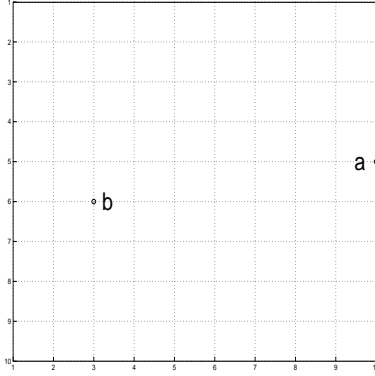


Figure 7: Dynamics of the winning units on the map induced by the fixed-input dynamics $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$, $s \in \{a, b\}$. For both input symbols, the autonomous dynamics of \mathbf{F}_s is dominated by an attractive fixed point \mathbf{y}_s . The induced dynamics on the map settles down in the position of the mode of \mathbf{y}_s .

maps of RecSOM with those obtained using a standard SOM operating on such fractal representations (of fixed dimensionality). Unlike in RecSOM, the dynamic part responsible for processing temporal context is fixed.

Theoretical upper bounds on β guaranteeing the existence of stable activation profiles in the fixed-input RecSOM dynamics were calculated as⁸: $\Upsilon(a) = 0.0226$ and $\Upsilon(b) = 0.0336$. Clearly, a fixed-point (attractive) RecSOM dynamics is obtained for values of β well above the guaranteed theoretical bounds (40).

4.2 IFS sequence representations combined with standard SOM

Previously, we have shown that a simple affine contractive iterative function system (IFS) [1] can be used to transform temporal structure of symbolic sequences into a spatial structure of points in a metric space [25]. The points represent subsequences in a Markovian manner: Subsequences sharing a common suffix are mapped close to each other. Furthermore, the longer is the shared suffix the closer lie the subsequence representations.

The IFS representing sequences over an alphabet \mathcal{A} of A symbols operates

⁸Again, we write the actual input symbols, rather than their vector encodings \mathbf{s} .

on an m -dimensional unit hypercube $[0, 1]^m$, where⁹ $m = \lceil \log_2 A \rceil$. With each symbol $s \in \mathcal{A}$ we associate an affine contraction on $[0, 1]^m$,

$$s(\mathbf{x}) = k\mathbf{x} + (1 - k)\mathbf{t}_s, \quad \mathbf{t}_s \in \{0, 1\}^m, \quad \mathbf{t}_s \neq \mathbf{t}_{s'} \text{ for } s \neq s', \quad (44)$$

with contraction coefficient $k \in (0, \frac{1}{2}]$. The attractor of the IFS (44) is the unique set $K \subseteq [0, 1]^m$, known as the Sierpinski sponge [14], for which $K = \bigcup_{s \in \mathcal{A}} s(K)$ [1].

For a prefix $u = u_1 u_2 \dots u_n$ of a string v over \mathcal{A} and a point $\mathbf{x} \in [0, 1]^m$, the point

$$u(\mathbf{x}) = u_n(u_{n-1}(\dots(u_2(u_1(\mathbf{x})))) = (u_n \circ u_{n-1} \circ \dots \circ u_2 \circ u_1)(\mathbf{x}) \quad (45)$$

constitutes a spatial representation of the prefix u under the IFS (44). Finally, the overall temporal structure of symbols in a (possibly long) sequence v over \mathcal{A} is represented by a collection of the spatial representations $u(\mathbf{x})$ of all its prefixes u , with a convention that $\mathbf{x} = \{\frac{1}{2}\}^m$.

Theoretical properties of such representations were investigated in [24]. The IFS-based Markovian coding scheme can be used to construct generative probabilistic models on sequences analogous to the variable memory length Markov models [25]. Key element of the construction is a quantization of the spatial IFS representations into clusters that group together subsequences sharing potentially long suffixes (densely populated regions of the suffix-organized IFS subsequence representations).

The Markovian layout of the IFS representations of symbolic sequences can also be used for constructing suffix-based topographic maps of symbolic streams in an unsupervised manner. By applying a standard SOM [16] to the IFS representations one may readily obtain topographic maps of Markovian flavour, similar to those obtained by RecSOM. The key difference between RecSOM and IFS+SOM (standard SOM operating on IFS representations) is that the latter approach assumes a fixed non-trainable dynamic part responsible for processing temporal contexts in the input stream. The recursion is not

⁹for $x \in \mathbb{R}$, $\lceil x \rceil$ is the smallest integer y , such that $y \geq x$

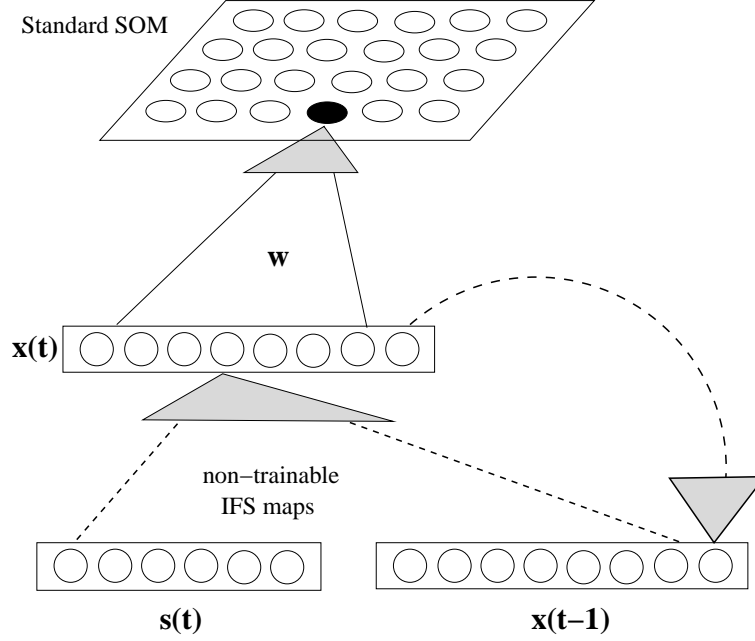


Figure 8: Standard SOM operating on IFS representations of symbolic streams (IFS+SOM model). Solid lines represent trainable feed-forward connections. No learning takes place in the dynamic IFS part responsible for processing temporal contexts in the input stream.

a part of the map itself, but is performed outside the map as a preprocessing step before feeding the standard SOM (see figure 8).

As shown in figure 9, the combination of IFS representations¹⁰ and standard SOM¹¹ leads to a map of RFs similar to that produced by RecSOM. In this case, quantizer depth was $\bar{n} = 4.48$.

The dynamics of SOM activations \mathbf{y} , driven by the IFS dynamics (44) and (45), again induces the dynamics (42) of winning units on the map. Since the IFS maps are affine contractions with fixed points \mathbf{t}_a and \mathbf{t}_b , the dynamics of winner units for both input symbols $s \in \{a, b\}$ settles in the SOM representations of \mathbf{t}_s (see figure 10). Note how fixed points of the induced winning neuron dynamics shape the suffix-based organization of receptive fields in figure 9.

¹⁰IFS coefficient $k = 0.3$

¹¹parameters such as learning rate and schedule for neighborhood width σ were taken from RecSOM

bbbbb	abbbb	abbb	ababb	aaabb	bbab	abaab	aaaab	bbbba	bbbba
bbbb	bbb	aabbb	baabb	b	abbab	baaab	–	bbbba	bbbba
babbb	bb	bbabb	baabb	bbbab	abab	aaaab	abbba	abbba	babba
abb	babb	aabb	abbab	aabab	abaab	bba	ba	bbaba	baaba
bab	abbab	babab	bbaab	abaab	abbba	ababa	ababa	aaaba	bbbba
ab	baab	aab	aaaab	babba	aaaba	bba	abaa	babaa	abbba
baaab	aaab	bbbba	aabba	aaba	abbba	aabaa	bbaaa	baaa	bbaaa
aaaab	bbbba	aabba	aba	a	babaa	aa	abaaa	aaa	baaaa
bbba	babba	baba	aaaba	bbbaa	babaa	bbaaa	baaaa	aaaaa	aaaaa
abbba	abba	ababa	aaaba	bbba	aabaa	abaaa	aaaa	aaaaa	aaaaa

Figure 9: Receptive fields of a standard SOM trained on IFS representations of sequences obtained from the automaton. Suffix-based topographic organization similar to that found in RecSOM is apparent.

4.3 Laser data

In this experiment we trained the RecSOM on a sequence of quantized activity differences of a laser in a chaotic regime. The series of length 8000 was quantized into a symbolic stream over 4 symbols (as in [27, 25, 28]) represented by two-bit binary codes: $a = 00$, $b = 01$, $c = 10$, $d = 11$. RecSOM with 2 inputs and $10 \times 10 = 100$ neurons was trained for 400.000 iterations, using $\alpha = 1$, $\beta = 0.2$ and $\gamma = 0.1$. The neighborhood width σ linearly decreased $5.0 \rightarrow 0.5$ during the first 300.000 iterations and then remained unchanged.

The behavior of the model is qualitatively the same as in the previous experiment. The weights became topographically ordered, as seen in figure 11. Input weights (on the left) can be clustered into four regions corresponding to the four symbols. Context weights of all neurons have a unimodal character and are ordered with respect to the peak position.

The map of RFs is shown in figure 12. Again, it is evident that the RFs

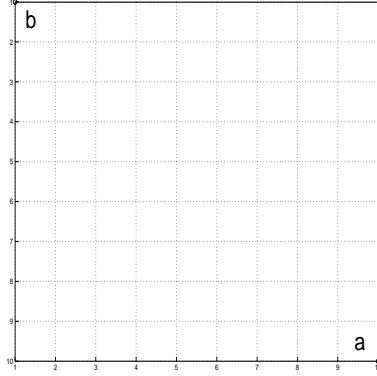


Figure 10: Dynamics of the winning units on the map of IFS+SOM induced by the fixed-input dynamics for symbols 'a' and 'b'.

are topographically ordered with respect to most recent symbols, even though there is a higher variability of RF depths among the neurons. The quantizer depth was $\bar{n} = 4.89$. The RF map is consistent with input weights of the neurons (left part of figure 11), when considering only the last symbol.

By checking the asymptotic regimes of the fixed-input RecSOM dynamics (10) as in the previous experiment, we found out that the fixed-input dynamics are again driven by unique attractive fixed points \mathbf{y}_a , \mathbf{y}_b , \mathbf{y}_c and \mathbf{y}_d . Dynamics of the winning units on the map induced by the fixed-input dynamics $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$, $s \in \{a, b, c, d\}$, is shown in figure 13 (left). The induced dynamics settles down in the mode position of \mathbf{y}_s .

Upper bounds on β guaranteeing the existence of stable activation profiles in the fixed-input RecSOM dynamics were determined as: $\Upsilon(a) = 0.0326$, $\Upsilon(b) = 0.0818$, $\Upsilon(c) = 0.0253$ and $\Upsilon(d) = 0.0743$. Again, we observe contractive behavior for β above the theoretical bounds.

As in the first experiment, we trained a standard SOM on (this time two-dimensional) inputs created by the IFS (44). The map is shown in figure 14. Both RecSOM and the combination of IFS with standard SOM¹² lead to 'similar' maps of RFs. The induced fixed-input dynamics of winning units in IFS+SOM are shown in figure 13 (right). The quantizer depth for IFS+SOM

¹²IFS coefficient $k = 0.3$; learning rate and schedule for the neighborhood width σ were taken from RecSOM

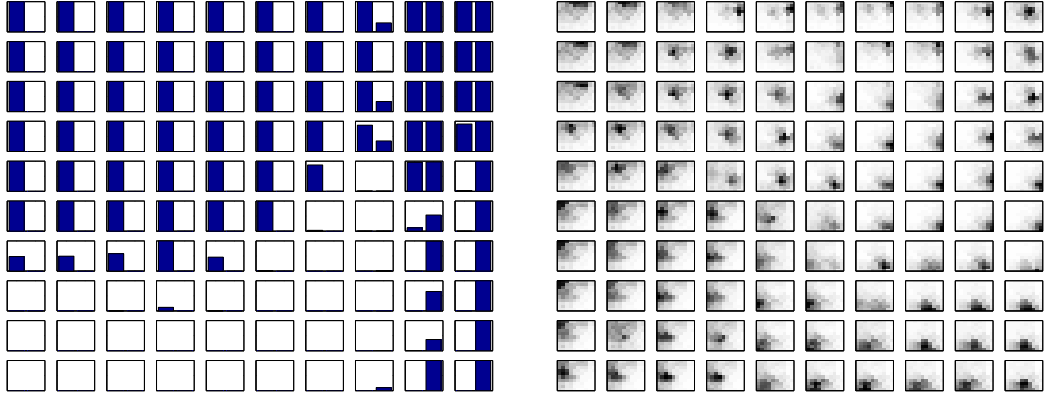


Figure 11: RecSOM weights after training on the laser data. All values are between 0 (white) and 1 (black). *Left:* Input weights (shown as histograms) can be clustered into four groups corresponding to four binary-coded inputs, with a few intermediate neurons at the boundaries. *Right:* Context weights (matrix plots) are topographically organized, with each neuron’s weight profile having a unimodal shape.

was evaluated as $\bar{n} = 5.2$.

4.4 Language

In our last experiment we used a corpus of written English, the novel ”Brave New World” by Aldous Huxley. In the corpus we removed punctuation symbols, upper-case letters were switched to lower-case and the space between words was transformed into a symbol ’-’. The complete data set (after filtering) comprised 356606 symbols. Letters of the Roman alphabet were binary-encoded using 5 bits and presented to the network one at a time. Unlike in [29], we did not reset the context map activations between the words. RecSOM with 400 neurons was trained for two epochs using the following parameter settings: $\alpha = 3$, $\beta = 0.7$, $\gamma = 0.1$ and $\sigma : 10 \rightarrow 0.5$. Radius σ reached its final value at the end of the first epoch and then remained constant to allow for fine-tuning of the weights. The map of RFs is displayed in figure 15 (quantizer depth $\bar{n} = 1.91$).

Figure 16 illustrates asymptotic regimes of the fixed-input RecSOM dynam-

badcc	dcc		abddc	abadc		bcdc		bdd	bcd
bcdcc		abacc	bcc			bddc	abddc	aabdd	
bddcc	aaccc	acc	aaacc		bc				aabad
bccc	aaccc			aabac		caabc		aabbd	
bccc	ccc			ac	caaac		aaba	bd	caabb
dccc		cccc	cccc	ccccc	c		caaba		caabb
			dcccc		accaa	aaaa	a	caaab	b
dcca		dccca		dccaa	ccaa		ccaaa		ccaab
cca	cca		cccca	cccaa	cccaa	aa	ccaaa		ccaab
ccca		cccca	cccca	cccaa	cccaa	cccaa		cccab	ccaab

Figure 12: Receptive fields of RecSOM trained on the laser data. Topographic organization is observed with respect to most recent symbols (at most 5 symbols are shown for clarity).

ics (10) in terms of map activity differences between consecutive time steps¹³. We observed a variety of behaviors. For some symbols, the activity differences converge to zero (attractive fixed points); for other symbols, the differences level at nonzero values (periodic attractors of period two, e.g. symbols 'i', 't', 'a', '-'). Fixed input RecSOM dynamics for symbol 'o' follows a complicated a-periodic trajectory¹⁴.

¹³Because of the higher dimensionality of the activation space ($N = 400$), we used a different strategy for generating the initial conditions $\mathbf{y}(0)$. We randomly varied only those components $y_i(0)$ of $\mathbf{y}(0)$, which had a potential to give rise to different fixed-input dynamics. Since $0 < y_i(0) \leq 1$ for all $i = 1, 2, \dots, N$, it follows from (11), that these can only be components y_i , for which the constant $G_\alpha(\mathbf{s}, \mathbf{w}_i)$ is not negligibly small. It is sufficient to use a small enough threshold $\theta > 0$, and set $y_i(0) = 0$ if $G_\alpha(\mathbf{s}, \mathbf{w}_i) < \theta$. Such a strategy can significantly reduce the dimension of the search space. We used $\theta = 0.001$ and the number of components of $\mathbf{y}(0)$ involved in generating the initial conditions varied from 31 to 138, depending on the input symbol.

¹⁴A detailed investigation revealed that the same holds for the autonomous dynamics under symbol 'e' (even though this is less obvious by scanning figure 16).

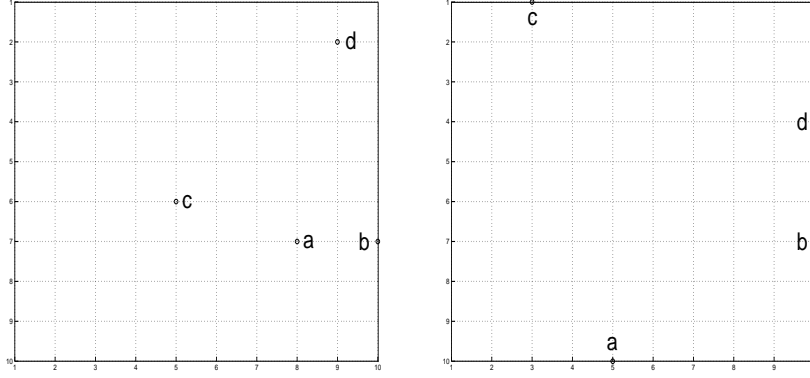


Figure 13: Laser data experiment. Dynamics of the winning units on the RecSOM (left) and IFS+SOM (right) maps induced by the fixed-input dynamics for input symbols 'a', 'b', 'c' and 'd'.

Dynamics of the winner units on the map induced by the fixed-input dynamics of \mathbf{F}_s are shown in figure 17 (left). As before, for symbols s with dynamics $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$ dominated by a single fixed point \mathbf{y}_s , the induced dynamics on the map settles down in the mode position of \mathbf{y}_s . However, some autonomous dynamics $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$ of period two (e.g. $s \in \{n, h, r, p, s\}$) induce a trivial dynamics on the map driven to a single point (grid position). In those cases, the points $\mathbf{y}^1, \mathbf{y}^2$ on the periodic orbit ($\mathbf{y}^1 = \mathbf{F}_s(\mathbf{y}^2), \mathbf{y}^2 = \mathbf{F}_s(\mathbf{y}^1)$) lie within the representation region (Voronoi compartment) of the same neuron. Interestingly enough, the complicated dynamics of \mathbf{F}_o and \mathbf{F}_e translates into aperiodic oscillations between just two grid positions. Still, the suffix based organization of RFs in figure 15 is shaped by the underlying collection of the fixed input dynamics of \mathbf{F}_s (illustrated in figure 17 (left) through the induced dynamics on the map).

Theoretical upper bounds on β (eq. (40)) are shown in figure 18. Whenever for an input symbol s the bound $\Upsilon(s)$ is above $\beta = 0.7$ (dashed horizontal line) used to train RecSOM (e.g. symbols 'j', 'q', 'x'), we can be certain that the fixed input dynamics given by the map \mathbf{F}_s will be dominated by an attractive fixed point. For symbols s with $\Upsilon(s) < \beta$, there is a possibility of a more complicated dynamics driven by \mathbf{F}_s . Indeed, the theoretical bounds $\Upsilon(s)$ for all symbols s with asymptotic fixed-input dynamics that goes beyond a single

acccc	cccc	cccc	dccc	bcdcc	bddcc		abcdc	bbcdc	bddc
	aaccc		dccc		bddcc				
aaacc	caacc	ccc	bccc	badcc		abadc	abbdc	aabcd	abbcd
	abacc		aabcc	bcc					bdd
caaac	ccaac	dccca	cabac		caabc	bc	aabad	aabd	abbd
aaaac	–	ccca	cdcca	aabac	aaabc	aabbc		b	
cccca	cccca	dccca	adcca	dcca			cccab	ccabb	caabb
ccca	aacca					ccaba			
	cccaa	ccaa		ccaaa	ccaaa	aaba		caaab	ccaab
dccaa	dccaa	cccaa	accaa	aaaa		caaba		ccaab	ccaab

Figure 14: Receptive fields of the IFS+SOM architecture obtained on the laser data. Topographic organization is observed with respect to the most recent symbols. The map is qualitatively similar to that obtained by RecSOM.

stable sink are below $\beta = 0.7$. Obviously, as seen in the previous experiments, $\Upsilon(s) < \beta$ does not necessarily imply more complicated fixed input dynamics on symbol s .

We also trained a standard SOM with 20×20 neurons on five-dimensional inputs created by the IFS¹⁵ (44). The map is shown in figure 19 (quantizer depth $\bar{n} = 1.69$). The induced dynamics on the map is illustrated in figure 17 (right). The suffix based organization of RFs is shaped by the underlying collection of autonomous attractive IFS dynamics.

¹⁵IFS coefficient $k = 0.3$; learning rate and schedule for the neighborhood width σ were taken from RecSOM

n-	n-	h-	ad-	d-	he-	he-	a-	ag	.		in	ig	.		-th	-th	-th	th	ti
an-	u-	-	l-	nd-	e-	re-	-a-	ao	an	ain	in		.	l		t-h	th	.	.
y-	i-	g-	ng-	ed-	f-	-to-	o-		en	un	-in		al	-al		h	wh		ty
ot-	at-	p-	-a-	n-	on-	m-	o-		-an	n	rn		ul	ll	e-l	e-h	gh	x	y
to	t-	es-	as-	er-	er-	mo	o	-to		-on	ion		.	ol	e-m	m		.	ey
t-	ut-	s-	is-	or-	ero	t-o	o	lo	ho	on	on	oo	.	om	um	im	am	ai	ry
ts	tw	ts-	r-	r-	ro	wo	io	e-o	-o	e-n	on			-m	t-m		si	ai	ri
e-s	he-w	-w	t-w		no	so	tio	-o	ng-o	-o	-n	-l	-h		e-i	di	ei	ni	ui
he-s	e-w	w	nw	ong	no	ak	k	-k	--	-o	.	-l	-h	-i	t-i	-wi	-hi	-li	-thi
ns	rs		ing	ng	nf	e-k	j	e-c	-s		-g	-m	-y		-i	-i	i	li	hi
s	us	uc	e-g	g	if	e-f	e-b	-c	-s	-w	-w	-e	.	-a	-a	n-a	ia	la	ha
is	c	nc		f	of	-f	-f	-b		-u	-u	-d		d-a	t-a	na	da	.	-ha
as	ac	ic	ib	b	.	oc	-v	.	-p	g-t	-t	-d	-e	-q	e-a	a	wa	era	ra
ac		ir	e-r		.	os		-r	-p	-t	s-t		.	ow		sa		ore	re
ar	ar	hr	r	tr	or	op	ov	-v	t-t	d-t	-t	ot	od	.	u	se	we	ere	pe
es	er	her	z	p	e-p	p	av		d-t	n-t	e-t	ot		ou	au	-se	be	ue	me
es	.	her	ter	ap	.	mp	v	st	rt	-st	tt	ut	out	lu	tu	e	e-e	ce	-he
ew		ev	.	q	ea	.	.	at	t	o-t	ent	ont	ind	d	dd	de	te	e	he
the-	e-	e-		em	ec	.		at	-at	ht	-it	nt	-and	rd	e-d		ne	-the	the
he-	e-	eo	.	.	ee	ed	ed	ad	it	it	id	ond	nd	and	ud	ld	le	-the	he

Figure 15: Receptive fields of RecSOM trained on English text.

5 Discussion

5.1 Topographic maps with Markovian flavour

Maps of sequential data obtained by RecSOM often seem to have a Markovian flavor. The neural units become sensitive to recently observed symbols. Suffix-based receptive fields (RFs) of the neurons are topographically organized in connected regions according to the last symbol. Within each of those regions, RFs are again topographically organized with respect to the symbol preceding the last symbol etc. Such a ‘self-similar structure’ is typical of spatial representations of symbolic sequences via contractive (affine) Iterative Function Systems (IFS) [13, 19, 21, 4, 9, 10, 24]. Such IFS can be considered simple non-autonomous dynamical systems driven by an input stream of symbols. Each IFS mapping is a contraction and therefore each fixed-input *autonomous* system has a trivial dynamics completely dominated by an attractive fixed point. However, the *non-autonomous* dynamics of the IFS can be quite complex, depending on the complexity of the input stream (see [24]).

More importantly, it is the attractive character of the individual fixed-input

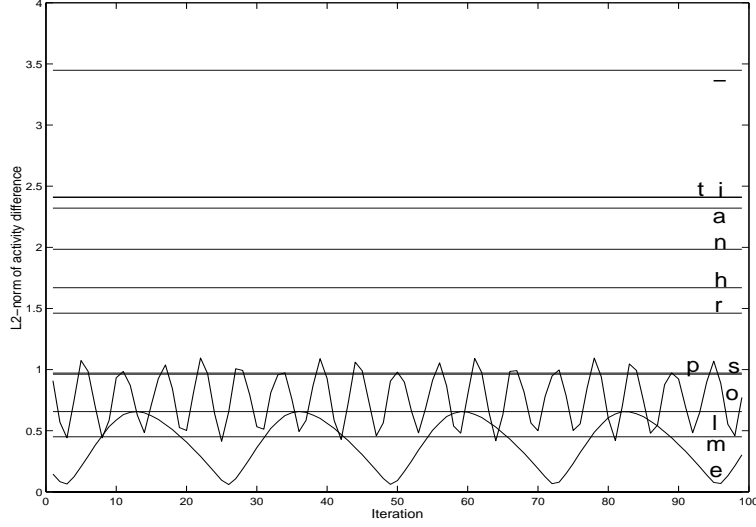


Figure 16: Fixed-input asymptotic dynamics of RecSOM after training on English text. Plotted are L_2 norms of the differences of map activities between the successive iterations. Labels denote the associated input symbols (for clarity, not all labels are shown).

IFS maps that shapes the Markovian organization of the state space. Imagine we feed the IFS with a long string $s_1 \dots s_{p-2} s_{p-1} s_p \dots s_{r-2} s_{r-1} s_r \dots$ over some finite alphabet \mathcal{A} of A symbols. Consider the IFS states at time instances p and r , $p < r$. No matter how far apart the time instances p and r are, if the prefixes $s_{1:p} = s_1 \dots s_{p-2} s_{p-1} s_p$ and $s_{1:r} = s_1 \dots s_{r-2} s_{r-1} s_r$ share a common suffix, the corresponding IFS states (see eqs. (44-45)), $s_{1:p}(\mathbf{x})$ and $s_{1:r}(\mathbf{x})$, will lie close to each other. If $s_{1:p}$ and $s_{1:r}$ share a suffix of length L , then for any initial position $\mathbf{x} \in [0, 1]^m$, $m = \lceil \log_2 A \rceil$,

$$\|s_{1:p}(\mathbf{x}) - s_{1:r}(\mathbf{x})\| \leq k^L \sqrt{m}, \quad (46)$$

where $0 < k < 1$ is the IFS contraction coefficient and \sqrt{m} is the diameter of the IFS state space $[0, 1]^m$. Hence, the longer is the shared suffix between $s_{1:p}$ and $s_{1:r}$, the shorter will be the distance between $s_{1:p}(\mathbf{x})$ and $s_{1:r}(\mathbf{x})$. The IFS translates the suffix structure of a symbolic stream into a spatial structure of points (prefix representations) that can be captured on a two-dimensional map using e.g. a standard SOM, as done in our IFS+SOM model.

Similar arguments can be made for a contractive RecSOM of N neurons.

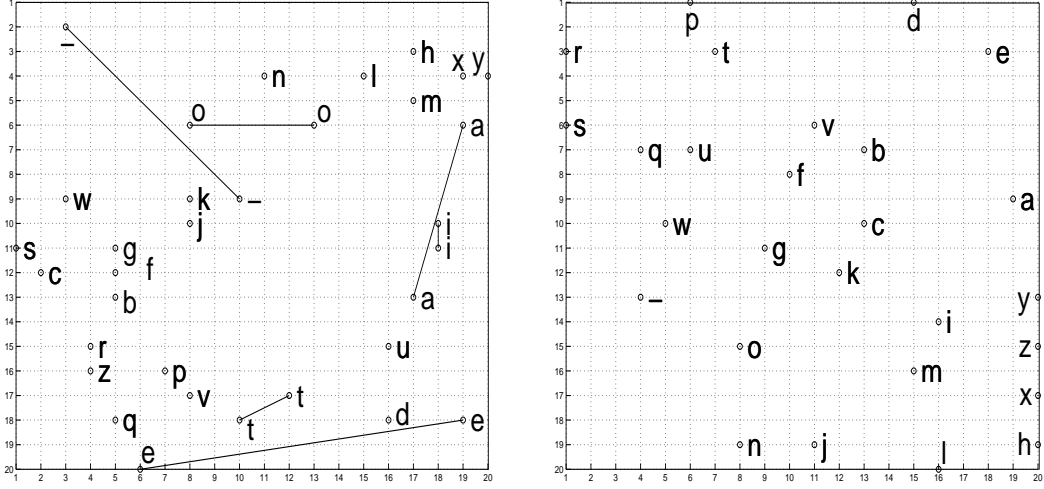


Figure 17: Dynamics of the winning units on the RecSOM (left) and IFS+SOM (right) maps induced by the fixed-input dynamics. The maps were trained on a corpus of written English ("Brave New World" by Aldous Huxley).

Assume that for each input symbol $s \in \mathcal{A}$, the fixed-input RecSOM mapping \mathbf{F}_s (eqs. (10-11)) is a contraction with contraction coefficient ρ_s . Set

$$\rho_{max} = \max_{s \in \mathcal{A}} \rho_s.$$

For a sequence $s_{1:n} = s_1 \dots s_{n-2} s_{n-1} s_n$ over \mathcal{A} and $\mathbf{y} \in (0, 1]^N$, define

$$\begin{aligned} \mathbf{F}_{s_{1:n}}(\mathbf{y}) &= \mathbf{F}_{s_n}(\mathbf{F}_{s_{n-1}}(\dots(\mathbf{F}_{s_2}(\mathbf{F}_{s_1}(\mathbf{y})))))) \\ &= (\mathbf{F}_{s_n} \circ \mathbf{F}_{s_{n-1}} \circ \dots \circ \mathbf{F}_{s_2} \circ \mathbf{F}_{s_1})(\mathbf{y}). \end{aligned} \quad (47)$$

Then, if two prefixes $s_{1:p}$ and $s_{1:r}$ of a sequence $s_1 \dots s_{p-2} s_{p-1} s_p \dots s_{r-2} s_{r-1} s_r \dots$ share a common suffix of length L , we have

$$\|\mathbf{F}_{s_{1:p}}(\mathbf{y}) - \mathbf{F}_{s_{1:r}}(\mathbf{y})\| \leq \rho_{max}^L \sqrt{N}, \quad (48)$$

where \sqrt{N} is the diameter of the RecSOM state space $(0, 1]^N$.

For sufficiently large L , the two activations $\mathbf{y}^1 = \mathbf{F}_{s_{1:p}}(\mathbf{y})$ and $\mathbf{y}^2 = \mathbf{F}_{s_{1:r}}(\mathbf{y})$ will be close enough to have the same location of the mode,¹⁶

$$i_* = \operatorname{argmax}_{i \in \{1, 2, \dots, N\}} y_i^1 = \operatorname{argmax}_{i \in \{1, 2, \dots, N\}} y_i^2,$$

¹⁶or at least mode locations on neighboring grid points of the map

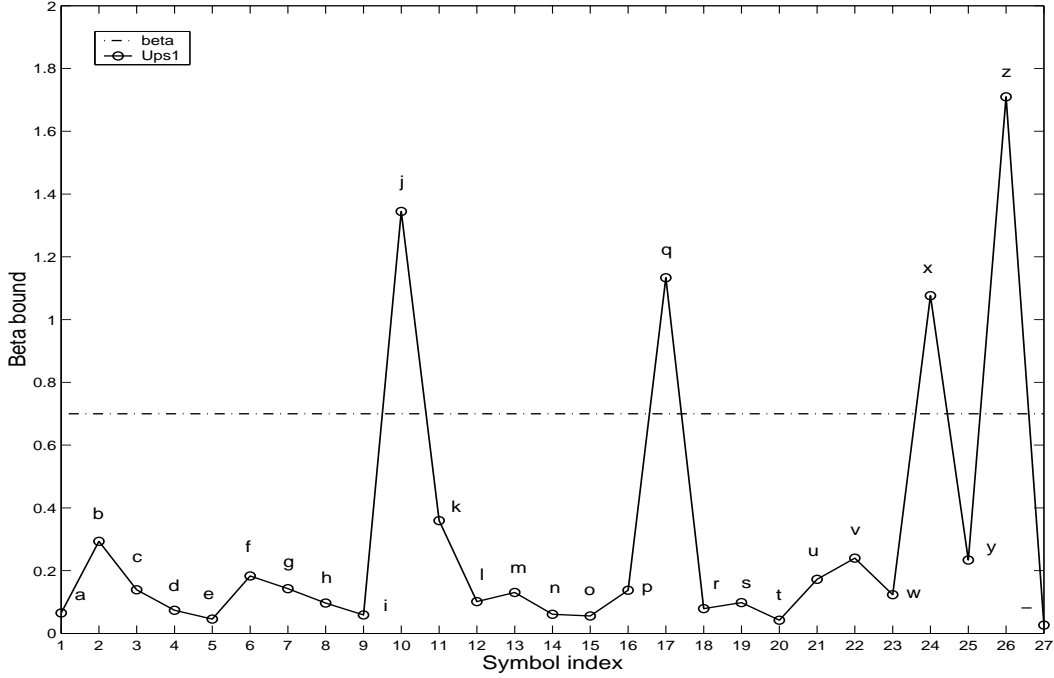


Figure 18: Theoretical bounds on β for RecSOM trained on the English text.

and the two subsequences $s_{1:p}$ and $s_{1:r}$ yield the same best matching unit i_* on the map, *irrespective of the position of the subsequences in the input stream*. All that matters is that the prefixes share a sufficiently long common suffix. We say that such an *organization of RFs on the map has a Markovian flavour*, because it is shaped solely by the suffix structure of the processed subsequences, and it does not depend on the temporal context in which they occur in the input stream. Obviously, one can imagine situations where **(1)** locations of the modes of \mathbf{y}^1 and \mathbf{y}^2 will be distinct, despite a small distance between \mathbf{y}^1 and \mathbf{y}^2 , or where **(2)** the modes of \mathbf{y}^1 and \mathbf{y}^2 coincide, while their distance is quite large. This is the price to be paid for discontinuity of the best-matching-unit operation (5). However, in our extensive experimental studies, we have registered only a negligible number of such cases. Indeed, some of the Markovian RFs in RecSOM maps obtained in the first two experiments over small (two- and four-letter) alphabets were quite deep (up to 10 symbols¹⁷).

¹⁷in some rare cases even deeper

r	r	tr	r	p	p	p	at	t	t	d	nd	d	ed	d		ne	-e	we	ve	
r	r	ur		p	p	t	t	ht	t	d	d	d	d		le	oe	ge	ue	te	
r	r	er	r	-vp	t	t	t	t	nt		-d	d		he	le	me	e	ndae	pe	
r	ar	er		st	st	ut	-t	-t	ot		d		je	e	e	e	de	tube	re	
s	s	s					-t	-t	rot	v	v	b		he	he	ye	fe	be	re	
s	s	s		u	u		t	-t		v		b	b			ke	ce	ce	se	
s	as	s	q	u	u	u		v	yf	f	issf	b	b	a	a				se	
s	s	s		u	-u	u		f	f	f	f	b		a	a	ga	a	la	la	
-s	-s	s	w	w	u	u			f	f	c	c		a	sa	a	ma	a	a	
			w	-w	w		g	g		c	c	c	c		a	-a	-a	a	a	
s-	s-	-		-w	w	g	g	g	o			c	c		-a	-a	-a		za	
s-	s-	-	h-v-			g	g	so	o	a-tk	k	k		i		a		y	y	
s-	-	w-	--	-	g-			o	so	o	rizo	k	-puk	i	i	i	i	ui	y	y
y-	k-	o-	o-	n-		-o	-o	o	o	yo		i	i	i	i	i	i	y	y	
y-	i-	m-	llo-	n-	n-	vo	o	o	o	o		m		i	li	i	i		z	
x-	h-	h-	l-	l-		o	o	o	lo	o		m	m	m		i	bbi	h	wh	
a-	a-	d-	d-	d-		po						-m	m	m	yl		h	h	x	
e-	u-	d-	d-	d-		n	n	on	kn	n	in	ernm		l	l	l	h	-h	h	
e-	e-	e-	d-	t-		n	n	n	n	j	n	l	l	l	dl	l	h	th	h	
e-	e-	e-	p-	t-		n	n	an	an	n	l	-l	l	l	l	l	th	th	h	

Figure 19: Receptive fields of a standard SOM with 20×20 units trained on IFS outputs, obtained on the English text. Topographic organization is observed with respect to the most recent symbols.

5.2 Non-Markovian topographic maps

Periodic (beyond period 1), or aperiodic attractive dynamics of autonomous systems $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$ lead to potentially complicated non-Markovian organizations of RFs on the map. By calculating the RF of a neuron i as the common suffix shared by subsequences yielding i as the best matching unit [29], we always create a suffix based map of RFs. Such RF maps are designed to illustrate the temporal structure learned by RecSOM. Periodic or aperiodic dynamics of \mathbf{F}_s can result in a ‘broken topography’ of RFs: two sequences with the same suffix can be mapped into distinct positions on the map, separated by a region of very different suffix structure. For example, depending on the context, subsequences ending with ‘ee’ can be mapped either near the lower-left, or near the lower-right corners of the RF map in figure 15. Unlike in contractive RecSOM or IFS+SOM models, such context-dependent RecSOM maps embody a potentially unbounded memory structure, because the current

position of the winner neuron is determined by the whole series of processed inputs, and not only by a history of recently seen symbols. Unless we understand the driving mechanism behind such context-sensitive suffix representations, we cannot fully appreciate the meaning of the RF structure of a RecSOM map.

There is a more profound question to be asked: What is the principal motivation behind building topographic maps of sequential data? If the motivation is a better understanding of cortical signal representations (e.g. [30]), then a considerable effort should be devoted to mathematical analysis of the scope of potential temporal representations and conditions for their emergence. If, on the other hand, the primary motivation is data exploration or data preprocessing, then we need to strive for a solid understanding of the way temporal contexts get represented on the map and in what way such representations fit the bill of the task we aim to solve.

There will be situations, where finite memory Markovian context representations are quite suitable. In that case, contractive RecSOM models, and indeed IFS+SOM models as well, may be appropriate candidates. But then the question arises of why exactly there needs to be a *trainable* dynamic part in self-organizing maps generalized to handle sequential data. As demonstrated in the first two experiments, IFS+SOM models can produce informative maps of Markovian context structures without an adaptive recursive submodel. One criterion for assessing the quality of RFs suggested by Voegtlin [29] is the quantizer depth (QD) (eq. (41)). If coding efficiency of induced RFs is a desirable property¹⁸, then RecSOM with Markovian maps may prove to be superior candidates to IFS+SOM models, as shown in the first experiment (QDs of RecSOM and IFS were evaluated as 6.18 and 4.48, respectively), but more theoretical and empirical work is needed to clarify this issue.

For more complicated data sets, like the English language corpus of the third experiment, RF maps beyond simple Markovian organization may be preferable. Yet, it is crucial to understand exactly what structures that are more powerful than Markovian organization of RFs are desired and why. It is

¹⁸Here we mean coding efficiency of RFs constrained by the two-dimensional map structure. Obviously, unconstrained codebooks will always be more optimal.

appealing to notice in the RF map of figure 15 the clearly non-Markovian spatial arrangement into distinct regions of RFs ending with the word-separation symbol ' '. Because of the special role of ' ' and its high frequency of occurrence, it may indeed be desirable to separate endings of words in distinct islands with more refined structure. However, to go beyond mere commenting on empirical observations, one needs to address issues such as

- what properties of the input stream are likely to induce periodic (or aperiodic) fixed input dynamics leading to context-dependent RF representations in SOMs with feedback structures,
- what periods for which symbols are preferable,
- what is the learning mechanism (e.g. sequence of bifurcations of the fixed input dynamics) of creating more complicated context dependent RF maps.

Those are the challenges for our future work.

5.3 Related work

It has been recently observed in [6] that Markovian representations of sequence data occur naturally in topographic maps governed by leaky integration, such as Temporal Kohonen Map [2]. Moreover, under some imposed circumstances, SOM for structured data [5] can represent trees in a Markovian manner by emphasising the topmost parts of the trees. These interesting findings were arrived at by studying pseudometrics in the data structure space induced by the maps. We complement the above results by studying the Rec-SOM map, potentially capable of very complicated dynamic representations, as non-autonomous dynamical systems governed by a collection of fixed input dynamics. Using corollary 3.7, we obtain:

Corollary 5.1 *Provided*

$$\beta < \frac{e}{2N}, \quad (49)$$

irrespective of the input \mathbf{s} , the map $\mathbf{F}_{\mathbf{s}}$ of a RecSOM with N recurrent neurons will be a contraction. For any external input \mathbf{s} , the fixed-input dynamics of such a RecSOM will be dominated by a single attractive fixed point.

Proof: It is sufficient to realize that

$$\|\mathbf{G}_{\alpha}(\mathbf{s})\|^2 = \sum_{i=1}^N e^{-2\alpha\|\mathbf{s}-\mathbf{w}_i\|^2} \leq N.$$

Q.E.D.

Corollary 5.1 states that if parameter β , weighting the importance of importing the past information into processing of sequential data, is smaller than $\frac{e}{2N}$ (N is the number of units on the map), the map is likely to be organized in a clear Markovian manner. The bound $e/(2N)$ may seem rather restrictive, but as argued in [6], the context influence has to be small for time series data to avoid instabilities in the model. Indeed, the RecSOM experiments of [6] (albeit on continuous data) used $N = 10 \times 10 = 100$ units and the map was trained with $\beta = 0.06$, which is only slightly higher than the bound $e/(2N) = 0.0136$. Obviously the bound $e/(2N)$ can be improved by considering other model parameters (Corollary 3.7), as demonstrated in figure 18.

Theoretical results of section 3 also complement Voegtlin’s stability analysis of the the weight adaptation process during training of RecSOM. For $\beta < e/(2N)$, stability of weight updates with respect to small perturbations of the activity profile \mathbf{y} is ensured [29]. Based on our analysis, we conclude that each RecSOM model satisfying Voegtlin’s stability bound on β is contractive, i.e. for *any* input, the fixed input RecSOM dynamics will be dominated by a unique attractive fixed point. This renders the map Markovian quality and training stability.

5.4 Relation between IFS+SOM and recurrent SOM models

Finally, we note that in the test mode (no learning), the IFS+SOM model acts exactly like the recurrent SOM (RSOM) model [17]. Given a sequence $s_1 s_2 \dots$

over a finite alphabet \mathcal{A} , the RSOM model determines the winner neuron at time t by identifying the neuron i with the minimal norm of

$$\mathbf{d}_i(t) = \nu (\mathbf{t}_{s_t} - \mathbf{w}_i) + (1 - \nu) \mathbf{d}_i(t - 1), \quad (50)$$

where $0 < \nu < 1$ is a parameter determining the rate of ‘forgetting the past’, \mathbf{t}_{s_t} is the code of symbol s_t presented at RSOM input at time t and \mathbf{w}_i is the weight vector on connections connecting the inputs with neuron i .

Inputs $\mathbf{x}(t)$ feeding standard SOM in the IFS+SOM model evolve with the IFS dynamics (see (44) and (45))

$$\mathbf{x}(t) = k \mathbf{x}(t - 1) + (1 - k) \mathbf{t}_{s_t}, \quad (51)$$

where $0 < k < 1$ is the IFS contraction coefficient. Best matching unit in SOM is determined by finding the neuron i with the minimal norm of

$$\mathbf{D}_i(t) = \mathbf{x}(t) - \mathbf{w}_i = k \mathbf{x}(t - 1) + (1 - k) \mathbf{t}_{s_t} - \mathbf{w}_i. \quad (52)$$

But $\mathbf{D}_i(t - 1) = \mathbf{x}(t - 1) - \mathbf{w}_i$, and so

$$\mathbf{D}_i(t) = k \mathbf{D}_i(t - 1) + (1 - k) (\mathbf{t}_{s_t} - \mathbf{w}_i), \quad (53)$$

which, after setting $\nu = 1 - k$, leads to

$$\mathbf{D}_i(t) = \nu (\mathbf{t}_{s_t} - \mathbf{w}_i) + (1 - \nu) \mathbf{D}_i(t - 1). \quad (54)$$

Provided $\nu = 1 - k$, the equations (50) and (54) are equivalent.

The key difference between RSOM and IFS+SOM models lies in the training process. While in RSOM, the best matching unit i with minimal norm of $\mathbf{d}_i(t)$ is shifted towards the current input \mathbf{t}_{s_t} , in IFS+SOM the winner unit i with minimal norm of $\mathbf{D}_i(t)$ is shifted towards the (Markovian) IFS code $\mathbf{x}(t)$ coding the whole history of recently seen inputs.

6 Conclusion

We have rigorously analyzed a generalization of the Self-Organizing Map (SOM) for processing sequential data – Recursive SOM (RecSOM [29]) – as a non-autonomous dynamical system consisting of a set of fixed input maps. We

have argued and experimentally demonstrated that contractive fixed input maps are likely to produce Markovian organizations of receptive fields on the RecSOM map. We have derived bounds on the parameter β , weighting the importance of importing the past information into processing of sequential data, that guarantee contractiveness of the fixed input maps.

Generalizations of SOM for sequential data, such as Temporal Kohonen Map [2], recurrent SOM [17], feedback SOM [11], RecSOM [29] and merge SOM [23], contain a dynamic module responsible for processing temporal contexts as an inherent part of the model. We have shown that Markovian topographic maps of sequential data can be produced by a simple fixed (non-adaptable) dynamic module externally feeding the topographic model.

We argue that non-Markovian organizations in topographic maps of sequential data may potentially be very important, but much more empirical and theoretical work is needed to clarify the map formation in SOMs endowed with feedback connections.

References

- [1] M.F. Barnsley. *Fractals everywhere*. Academic Press, New York, 1988.
- [2] G. Chappell and J. Taylor. The temporal kohonen map. *Neural Networks*, 6:441–445, 1993.
- [3] G de A. Barreto, A.F.R. Araújo, and S.C. Kremer. A taxonomy of spatiotemporal connectionist networks revisited: The unsupervised case. *Neural Computation*, 15:1255–1320, 2003.
- [4] A. Fiser, G.E. Tusnady, and I. Simon. Chaos game representation of protein structures. *Journal of Molecular Graphics*, 12(4):302–304, 1994.
- [5] M. Hagenbuchner, A. Sperduti, and A.C. Tsoi. Self-organizing map for adaptive processing of structured data. *IEEE Transactions on Neural Networks*, 14(3):491–505, 2003.

- [6] B. Hammer, A. Micheli, A. Sperduti, and M. Strickert. Recursive self-organizing network models. *Neural Networks*, 17(8-9):1061–1085, 2004.
- [7] B. Hammer, A. Micheli, M. Strickert, and A. Sperduti. A general framework for unsupervised processing of structured data. *Neurocomputing*, 57:3–35, 2004.
- [8] B. Hammer and P. Tiño. Neural networks with small weights implement finite memory machines. *Neural Computation*, 15(8):1897–1926, 2003.
- [9] B.-L. Hao. Fractals from genomes – exact solutions of a biology-inspired problem. *Physica A*, (282):225–246, 2000.
- [10] B.-L. Hao, H.C. Lee, and S. Zhang. Fractals related to long dna sequences and complete genomes. *Chaos, Solitons and Fractals*, 11:825–836, 2000.
- [11] K. Horio and T. Yamakawa. Feedback self-organizing map and its application to spatio-temporal pattern classification. *International Journal of Computational Intelligence and Applications*, 1(1):1–18, 2001.
- [12] D.L. James and R. Miikkulainen. Sardnet: A self-organizing feature map for sequences. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Proceedings of the Advances in Neural Information Processing Systems*, volume 7, pages 577–584. Morgan Kaufmann, 1995.
- [13] J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.
- [14] R. Kenyon and Y. Peres. Measures of full dimension on affine invariant sets. *Ergodic Theory and Dynamical Systems*, 16:307–323, 1996.
- [15] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
- [16] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1479, 1990.

- [17] T. Koskela, M. Varsta and J. Heikkonen, and K. Kaski. Recurrent SOM with local linear models in time series prediction. In *6th European Symposium on Artificial Neural Networks*, pages 167–172, 1998.
- [18] J.A. Lee and M. Verleysen. Self-organizing maps with recursive neighborhood adaptation. *Neural Networks*, 15(8–9):993–1003, 2002.
- [19] J.L. Oliver, P. Bernaola-Galván, J. Guerrero-Garcia, and R. Román Roldan. Entropic profiles of dna sequences through chaos-game-derived images. *Journal of Theor. Biology*, (160):457–470, 1993.
- [20] J.C. Principe, N. Euliano, and S. Garani. Principles and networks for self-organization in space-time. *Neural Networks*, 15(8–9):1069–1083, 2002.
- [21] R. Roman-Roldan, P. Bernaola-Galvan, and J.L. Oliver. Entropic feature for sequence pattern through iteration function systems. *Pattern Recognition Letters*, 15:567–573, 1994.
- [22] R. Schulz and J.A. Reggia. Temporally asymmetric learning supports sequence processing in multi-winner self-organizing maps. *Neural Computation*, 16:535–561, 2004.
- [23] M. Strickert and B. Hammer. Neural gas for sequences. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 53–57, 2003.
- [24] P. Tiño. Multifractal properties of Hao’s geometric representations of DNA sequences. *Physica A: Statistical Mechanics and its Applications*, 304(3–4):480–494, 2002.
- [25] P. Tiño and G. Dorffner. Predicting the future of discrete sequences from fractal representations of the past. *Machine Learning*, 45(2):187–218, 2001.
- [26] P. Tiño and B. Hammer. Architectural bias in recurrent neural networks: Fractal analysis. *Neural Computation*, 15(8):1931–1957, 2004.

- [27] P. Tiño and M. Koteles. Extracting finite state representations from recurrent neural networks trained on chaotic symbolic sequences. *IEEE Transactions on Neural Networks*, 10(2):284–302, 1999.
- [28] P. Tiño, M. Čerňanský, and L. Beňušková. Markovian architectural bias of recurrent neural networks. *IEEE Transactions on Neural Networks*, 15(1):6–15, 2004.
- [29] T. Voegtlin. Recursive self-organizing maps. *Neural Networks*, 15(8–9):979–992, 2002.
- [30] J.C. Wiemer. The time-organized map algorithm: Extending the self-organizing map to spatiotemporal signals. *Neural Computation*, 16:1143–1171, 2003.
- [31] H. Yin. ViSOM - a novel method for multivariate data projection and structure visualisation. *IEEE Transactions on Neural Networks*, 13(1):237–243, 2002.