Dynamics and Topographic Organization of Recursive Self-Organizing Maps

Peter Tiňo
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
and
Institute of Measurement Science, Slovak Academy of Sciences
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) (Voegtlin, 2002), 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). However, by allowing trainable

feedback connections one can obtain Markovian maps with superior memory depth and topography preservation. 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) (Kohonen, 1982) is a nonlinear projection method that maps a high-dimensional metric vector space onto a two-dimensional regular grid in a topologically ordered fashion (Kohonen, 1990). 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 learning[1] and co-operative learning[2] are employed. Many modifications of the standard SOM have been proposed in the literature (e.g. Yin, 2002; Lee & Verleysen, 2002). 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 - (Barreto, Araújo & Kremer, 2003) and (Hammer et al., 2004) 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 (Chappell & Taylor, 1993), recurrent SOM (Koskela et al., 1998), feedback SOM (Horio & Yamakawa, 2001), recursive SOM (Voegtlin, 2002), merge SOM (Strickert & Hammer, 2003) and SOM for structured data (Hagenbuchner, Sperduti, & Tsoi, 2003). Other alternatives for

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

[2] not 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

constructing topographic maps of structured data have been suggested e.g. in (James & Miikkulainen, 1995; Principe, Euliano & Garani, 2002; Wiemer, 2003; Schulz & Reggia, 2004).

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 (Barreto, Araújo & Kremer, 2003; Schulz & Reggia, 2004; Hammer et al., 2004). As pointed out in (Hammer et al., 2004), 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 a much needed mathematical characterization and analysis of such models were taken in (Hammer et al., 2004; Hammer et al., 2004a). 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 (Tiňo, Čerňanský & Beňušková, 2004). 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 (Tiňo, Čerňanský & Beňušková, 2004; Hammer & Tiňo, 2003; Tiňo & Hammer, 2003).

We concentrate on the Recursive SOM (RecSOM) (Voegtlin, 2002), 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 (Hammer et al., 2004).

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 (Tiňo & Dorffner, 2001)?

The paper has the following organization: We introduce the RecSOM model in section

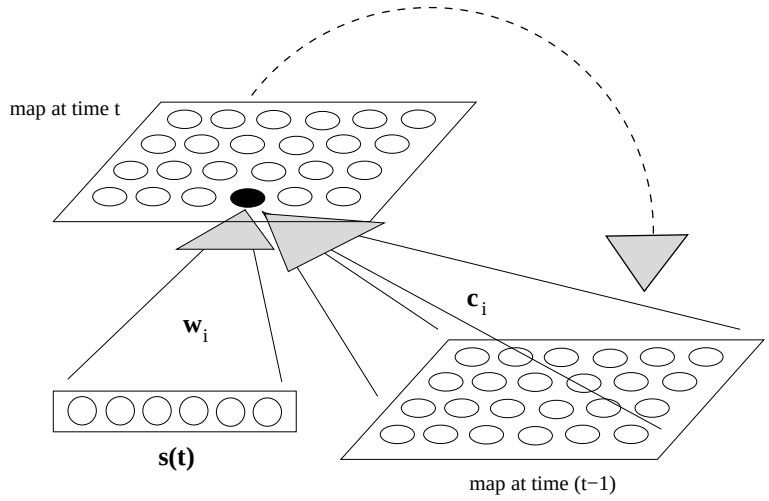

Figure 1: Recursive SOM architecture. The original SOM algorithm is used for both input vector $\mathbf{s}$(t) and for the context represented as the map activation $\mathbf{y}$(t-1) from the previous time step. Solid lines represent trainable connections, dashed line represents one-to-one copy of the 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 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.

2 Recursive Self-Organizing Map - RecSOM

The architecture of the RecSOM model (Voegtlin, 2002) is shown in figure 1. Each neuron $i \in \{1, 2, ..., N\}$ 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

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

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

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

$$y_i(t) = \exp(-d_i(t)), \tag{1}$$

where[3]

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

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

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 (Voegtlin, 2002):

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

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

where k is an index of the best matching unit at time t, $k = \mathrm{argmin}_{i \in \{1,2,...,N\}} d_i(t)$, 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 $y_k(t)$:

$$k = \underset{i \in \{1,2,...,N\}}{\mathrm{argmax}} \; y_i(t). \tag{5}$$

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

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

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

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

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

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

where

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

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

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

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

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

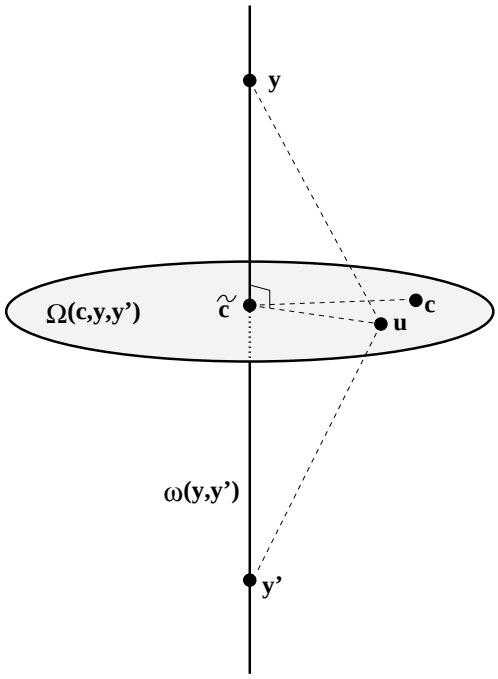

Figure 2: Illustration for the proof of Lemma 3.1. The line $\omega(\mathbf{y}, \mathbf{y}')$ passes through $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$. The $(N-1)$-dimensional hyperplane $\Omega(\mathbf{c}, \mathbf{y}, \mathbf{y}')$ is orthogonal to $\omega(\mathbf{y}, \mathbf{y}')$ and contains the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

where

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

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

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

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

$\mathbf{F}$ 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 three N-dimensional points $\mathbf{y}, \mathbf{y}', \mathbf{c} \in \mathbb{R}^N$, $\mathbf{y} \neq \mathbf{y}'$. 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 $\beta > 0$,

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

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 $\overline{\omega}(\mathbf{y},\mathbf{y}')$ be the line $\omega(\mathbf{y},\mathbf{y}')$ without the segment connecting $\mathbf{y}$ and $\mathbf{y}'$, i.e.*

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

Then, for any $\beta > 0$,

$$\text{argmax}_{\mathbf{c}\in\omega(\mathbf{y},\mathbf{y}')} \left\{|G_\beta(\mathbf{y},\mathbf{c}) - G_\beta(\mathbf{y}',\mathbf{c})|\right\} \in \overline{\omega}(\mathbf{y},\mathbf{y}').$$

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 every[4] $\mathbf{c}_- \in \omega(\mathbf{y},\mathbf{y}') \setminus \overline{\omega}(\mathbf{y},\mathbf{y}')$ in between the points $\mathbf{y}$ and $\mathbf{y}'$, there exist a $\mathbf{c}_+ \in \overline{\omega}(\mathbf{y},\mathbf{y}')$ 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 $H_\beta : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$,

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

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

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

where

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

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

$$H_\beta(\kappa_\beta(\delta),\delta) = 0. \tag{14}$$

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 $\overline{\omega}(\mathbf{y},\mathbf{y}')$, i.e. on points outside the line segment connecting $\mathbf{y}$ and $\mathbf{y}'$.

Consider points on the line segment

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

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

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

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

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

which is equivalent to

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

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

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

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

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:

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

which can be rearranged as

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

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

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

Q.E.D.

Lemma 3.4 *For all $k > 0$,*

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

Proof: Consider the functions

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

We find that

$$\lim_{k\to 0} \Delta_1(k) \simeq 1/2k^2 + \log(k) > 0, \qquad \lim_{k\to\infty} \Delta_1(k) \simeq 1/k^2 > 0$$
$$\lim_{k\to 0} \Delta_2(k) \simeq -\log(k) > 0, \qquad \lim_{k\to\infty} \Delta_2(k) \simeq 1/k^2 > 0.$$

Since

$$\Delta_1'(k) = -\tfrac{(1+2k)}{k^3(1+k)} < 0,$$
$$\Delta_2'(k) = -\tfrac{(1+2k)}{k(1+k)^3} < 0,$$

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 $D_\beta : \mathbb{R}^+ \to (0,1)$,*

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

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

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

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

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

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

is negative for all $\kappa > 0$.

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

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

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

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

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

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

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

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

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

   Furthermore, we have that

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

   and

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

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

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

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

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

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

and

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

respectively.

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

$$1+2k-2k^2\log\left(\frac{1+k}{k}\right) > 0,$$

$$1+2k-2(1+k)^2\log\left(\frac{1+k}{k}\right) < 0.$$

All the other factors in (33) are positive.

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

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

and so

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

Q.E.D.

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

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

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

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

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

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

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

$$\|\mathbf{F_S}(\mathbf{y}) - \mathbf{F_S}(\mathbf{y}')\| \leq \rho \cdot \|\mathbf{y} - \mathbf{y}'\|.$$

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

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

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 $\lim_{\delta \to 0^+} D_\beta(\delta) = 0$, with $\lim_{\delta \to 0^+} \frac{dD_\beta(\delta)}{d\delta} = \sqrt{\frac{2\beta}{e}}$, we have the following upper bound:

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

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

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

then $\mathbf{F_S}$ will be a contraction with contraction coefficient ρ.

Inequality (38) is equivalent to

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

Q.E.D.

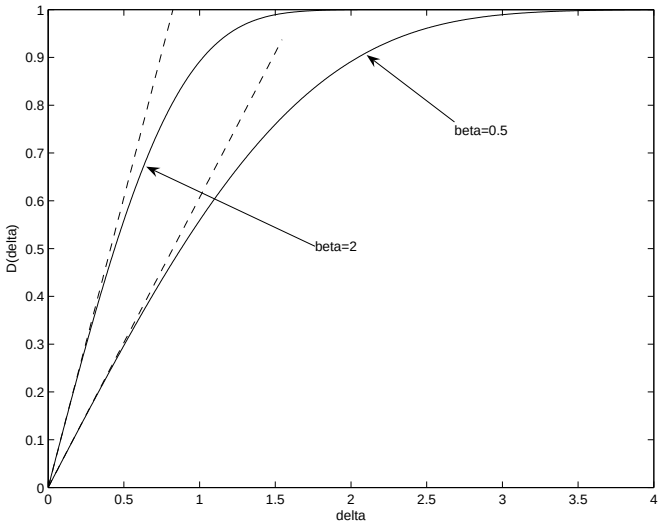

Figure 3: Functions $D_\beta(\delta)$ for $\beta = 0.5$ and $\beta = 2$ (solid lines). Also shown (dashed lines) are the linear upper bounds (37).

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

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

Then, if $\beta < \Upsilon(\mathbf{s})$, $\mathbf{F_S}$ *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 $\delta \to 0^+$. 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 (Voegtlin, 2002).

Following Voegtlin (2002), in order to represent the trained maps, we calculate for each unit in the map its receptive field (RF). Receptive field of a neuron is the common

suffix of all sequences for which that neuron becomes the best-matching unit (Voegtlin, 2002). Voegtlin (2002) also suggests to measure the amount of memory captured by the map through quantizer depth,

$$\mathrm{QD} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{p_i} \ell_\mathrm{i}, \tag{41}$$

where p_i is the probability of the RF of neuron i and ℓ_i is its length.

To assess maps from the point of view of topography preservation, we introduce a measure that aims to quantify the maps' topographic order. For each unit in the map we first calculate the length of the longest common suffix shared by RFs of that unit and its immediate topological neighbors. In other words, for each unit i on the grid, we create a set of strings $\mathcal{R}_i$ consisting of RF of unit i and RFs of its four neighbors on the grid[6]. The length of the longest common suffix of the strings in $\mathcal{R}_i$ is denoted by $\ell(\mathcal{R}_i)$. The topography preservation measure[7] TP is the average of such shared RF suffix lengths over all units in the map,

$$\mathrm{TP} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \ell(\mathcal{R}_\mathrm{i}). \tag{42}$$

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 (Tiňo & Dorffner, 2001).

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 $P(b|a) = 0.4$ (Voegtlin, 2002). 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 (Voegtlin, 2002): $\alpha = 2$ 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)[8].

[6] neurons at the grid boundary have less than four nearest neighbors

[7] It should be noted that quantifying topography preservation in recursive extensions of SOM is not as straightforward as in traditional SOM (Hammer et al., 2004). The proposed TP measure quantifies the degree of local conservation of suffix based RFs across the map.

[8] Voegtlin 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 $\sigma = 0.5$

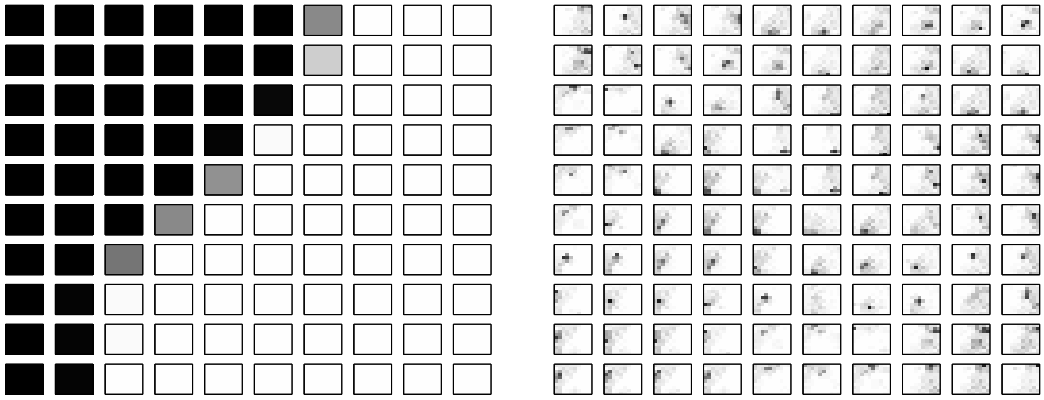

Figure 4: Converged input (left) and context (right) weights after training on the stochastic automaton. All values are between 0 (code of 'a' – white) and 1 (code of 'b' – black). Input weights can be clustered into two groups corresponding to the two input symbols, with a few intermediate units at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

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).

Receptive fields (RFs) of all units of the map are shown in figure 5. For each unit $i \in \{1, 2, ..., N\}$, its RF is shaded according to the local topography preservation measure $\ell(\mathcal{R}_i)$ (see (42))[9]. 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.

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 dynamics (10) of the mappings[10] $\mathbf{F}_a$ 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 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

(used in (Voegtlin, 2002)), 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.

[9]We thank one of the anonymous reviewers for this suggestion.

[10]We slightly abuse the mathematical notation here by indexing the fixed input RecSOM maps $\mathbf{F}$ with the actual input symbols, rather than their vector encodings $\mathbf{s}$.

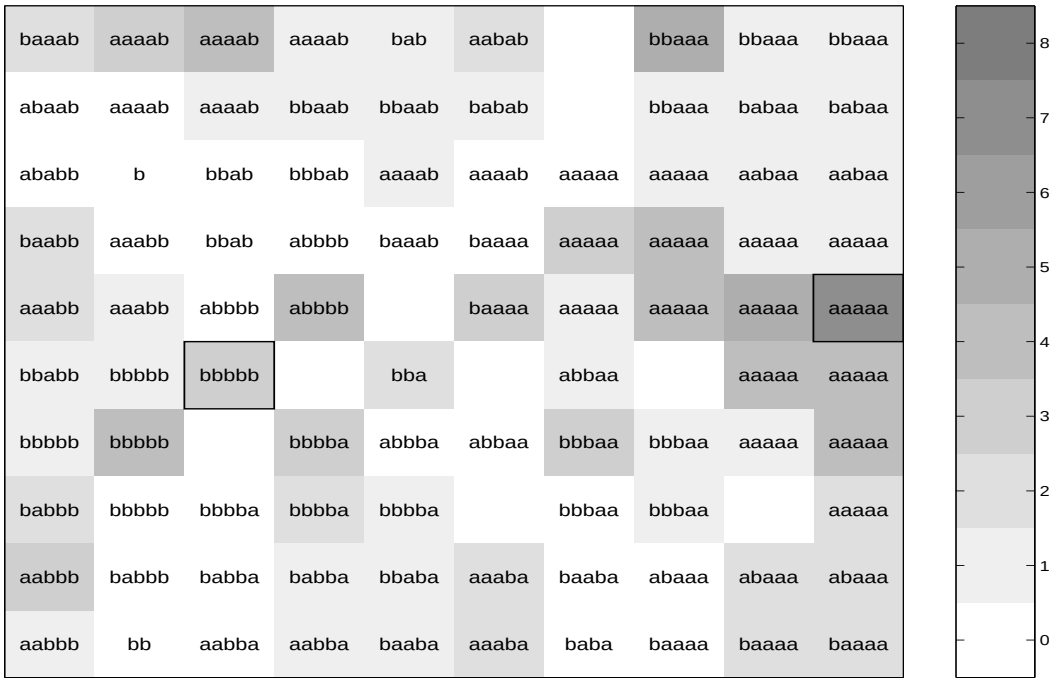

Figure 5: Receptive fields of RecSOM trained on the stochastic two-state automaton. Topographic organization is observed with respect to most recent symbols (only 5 symbols are shown for clarity). Empty slots signify neurons that were not registered as best-matching units when processing the data. Receptive field of each unit i is shaded according to the local topography preservation measure $\ell(\mathcal{R}_i)$. For each input symbol $s \in \{a, b\}$, we mark the position of the fixed point attractor i_s of the induced (fixed input) dynamics on the map by a square around its RF.

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, e.g. there is a unique dimension (map unit) of pronounced maximum activity.

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 RFs in 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:

$$\begin{aligned} i_s(t) &= \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmax}} \; y_i(t) \\ &= \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmax}} \; F_{s,i}(\mathbf{y}(t-1)). \end{aligned} \tag{43}$$

To illustrate the dynamics (43), 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

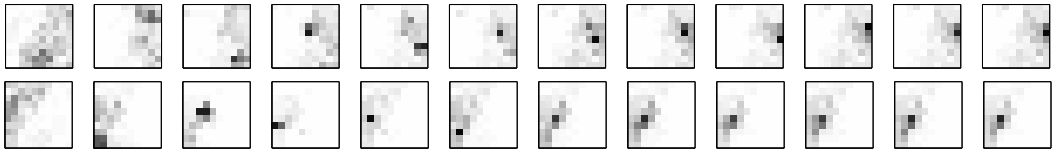

Figure 6: Fixed-input dynamics of RecSOM trained on the stochastic automaton – symbol 'a' (1st row), symbol 'b' (2nd row) . The activations settle to a stable unimodal profile after roughly 10 iterations.

on the map, (43), settles down in neuron i_s, corresponding to the mode of $\mathbf{y}_s$:

$$i_s = \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmax}} \; y_{s,i}. \tag{44}$$

Position of the neuron i_s is marked in figure 5 by a square around its RF. 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. Indeed, for each $s \in \{a, b\}$, the RF of i_s ends with a long block of symbols s and the local topography preservation $\ell(\mathcal{R}_{i_s})$ around i_s is high.

This mechanism for creating suffix-based RF organization is reminiscent of the Markovian fractal subsequence representations used in (Tiňo & Dorffner, 2001) to build Markov models with context dependent length. In the next subsection we compare 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 as[11]: $\Upsilon(a) = 0.0226$ 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) (Barnsley, 1988) can be used to transform temporal structure of symbolic sequences into a spatial structure of points in a metric space (Tiňo & Dorffner, 2001). The points represent subsequences in a Markovian manner: Subsequences sharing a common suffix are

[11] Again, we write the actual input symbols, rather than their vector encodings $\mathbf{s}$.

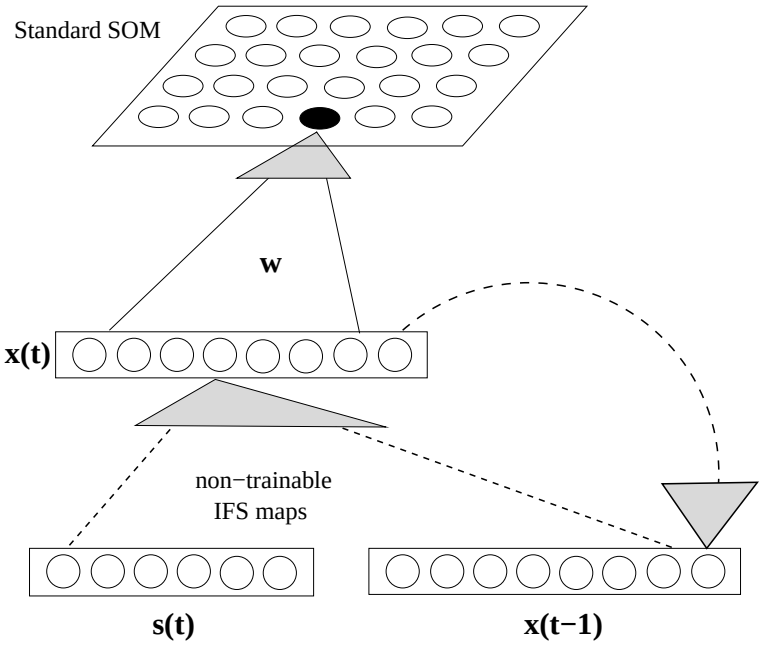

Figure 7: Standard SOM operating on IFS representations of symbolic streams (IFS+SOM model). Solid lines represent trainable feed-forward connections. No learning takes place in the dynamic IFS part responsible for processing temporal contexts in the input stream.

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 on an m-dimensional unit hypercube $[0,1]^m$, where[12] $m = \lceil \log_2 A \rceil$. With each symbol $s \in \mathcal{A}$ we associate an affine contraction on $[0,1]^m$,

$$s(\mathbf{x}) = k\mathbf{x} + (1-k)\mathbf{t}_s, \quad \mathbf{t}_s \in \{0,1\}^m, \quad \mathbf{t}_s \neq \mathbf{t}_{s'} \text{ for } s \neq s', \tag{45}$$

with contraction coefficient $k \in (0, \frac{1}{2}]$. The attractor of the IFS (45) is the unique set $K \subseteq [0,1]^m$, known as the Sierpinski sponge (Kenyon & Peres, 1996), for which $K = \bigcup_{s \in \mathcal{A}} s(K)$ (Barnsley, 1988).

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

$$u(\mathbf{x}) = u_n(u_{n-1}(...(u_2(u_1(\mathbf{x})))...)) = (u_n \circ u_{n-1} \circ ... \circ u_2 \circ u_1)(\mathbf{x}) \tag{46}$$

constitutes a spatial representation of the prefix u under the IFS (45). 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 (Tiňo, 2002). The IFS-based Markovian coding scheme can be used to construct generative probabilistic

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

models on sequences analogous to the variable memory length Markov models (Tiňo & Dorffner, 2001). 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 (Kohonen, 1990) 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 a part of the map itself, but is performed outside the map as a preprocessing step before feeding the standard SOM (see figure 7). As shown in figure 8, the combination of IFS representations[13] and standard SOM[14] leads to a suffix-based organization of RFs on the map, similar to that produced by RecSOM. In both models, the RFs are topographically ordered with respect to the most recent input symbols.

The dynamics of SOM activations $\mathbf{y}$, driven by the IFS dynamics (45) and (46), again induces the dynamics (43) 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 i_s of $\mathbf{t}_s$. Note how fixed points i_s of the induced winning neuron dynamics shape the suffix-based organization of receptive fields in figure 8.

To compare RecSOM and IFS+SOM maps in terms of quantization depth (QD) (41) and topography preservation (TP) (42), we varied RecSOM parameters α and β and ran 40 training sessions for each setting of (α, β). The resulting TP and QD values were compared with those of the IFS+SOM maps constructed in 40 independent training sessions. Other parameters were the same as in the previous simulations, and were identical for both models. We chose a 3×3 design using $\alpha \in \{1, 2, 3\}$ and $\beta \in \{0.2, 0.7, 1\}$ attempting to meaningfully cover the parameter space of RecSOM (see (Voegtlin, 2002)). Each RecSOM model was compared to IFS+SOM using a two-tail t-test. Results are shown in Table 1. The number of stars denotes the significance level of the difference ($p < 0.05, 0.01, 0.001$, in ascending order). Almost all differences are significant. Specifically, QD of RecSOM is significantly higher for all combinations of α and β[15]. The TP for RecSOM is also significantly higher in most cases, except for those with higher β/α ratio. As explained in

[13]IFS coefficient $k = 0.3$

[14]parameters such as learning rate and schedule for neighborhood width σ were taken from RecSOM

[15]Due to the constraints imposed by topographic organization of RFs, the quantizer depths of the maps are smaller than that of the theoretically optimal (unconstrained) quantizer computed by Voegtlin (2002) as QD = 7.08.

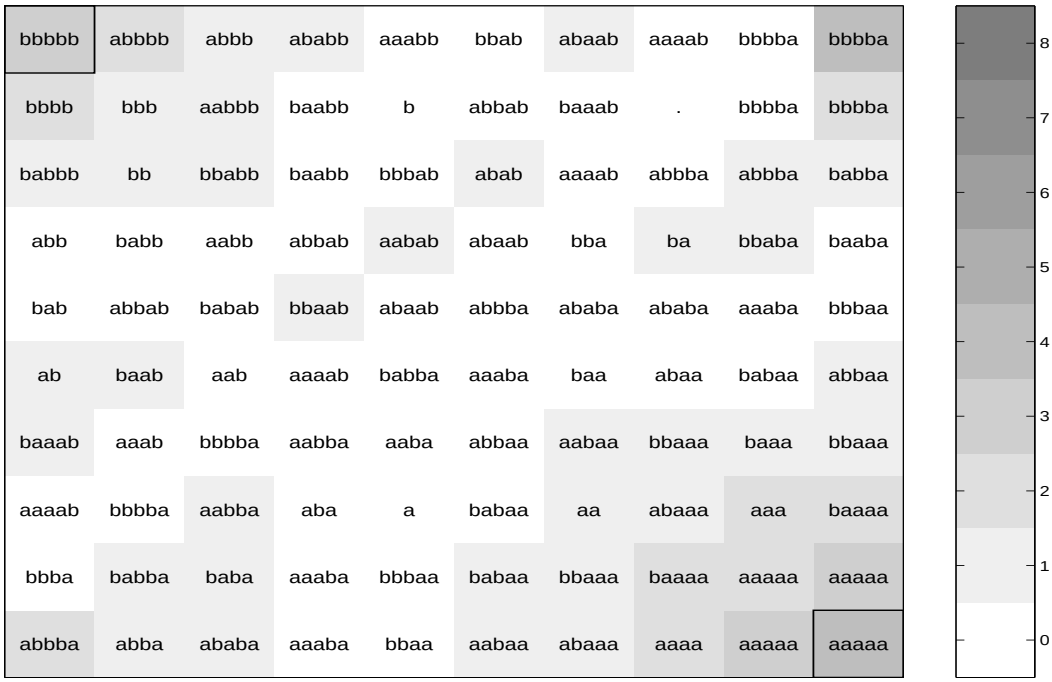

Figure 8: Receptive fields of a standard SOM trained on IFS representations of sequences obtained from the automaton. Suffix-based topographic organization similar to that found in RecSOM is apparent. Receptive field of each unit i is shaded according to the local topography preservation measure $\ell(\mathcal{R}_i)$. For each $s \in \{a, b\}$, we mark the position of the fixed point attractor i_s of the induced dynamics on the map by a square around its RF.

section 4.1, trivial contractive fixed input dynamics dominated by unique attractive fixed points lead to Markovian suffix-based RF organizations. Lower TP values are observed for higher β/α ratios, because in those cases, more complicated fixed-input dynamics can arise, breaking the Markovian RF maps.

$\beta \setminus \alpha$	1.0		2.0		3.0	
0.2	5.84^{***}	2.51^{***}	6.15^{***}	2.77^{***}	6.05^{***}	2.55^{***}
0.7	6.11^{***}	1.27^{***}	6.14^{***}	2.21^{***}	6.12^{***}	2.43^{***}
1.0	5.75^{***}	0.75^{**}	5.87^{***}	2.00	5.92^{***}	2.10^{***}

Table 1: Means of the (QD, TP) measures, averaged over 40 simulations, for RecSOM trained on the stochastic automaton. Corresponding means for IFS+SOM were as follows: QD = 5.55 and TP = 1.96.

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 sym-

bolic stream over 4 symbols (as in (Tiňo & Köteles, 1999; Tiňo & Dorffner, 2001; Tiňo, Čerňanský & Beňušková, 2004)) 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 was qualitatively the same as in the previous experiment. The map of RFs was topographically ordered with respect to most recent symbols. 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 $\mathbf{y}_d$. As before, the dynamics of winning units on the map induced by the fixed-input dynamics $\mathbf{y}(t) = \mathbf{F}_s(\mathbf{y}(t-1))$, $s \in \{a, b, c, d\}$, settled down in the mode position i_s 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 (45). Again, both RecSOM and the combination of IFS with standard SOM[16] lead to suffix-based maps of RFs, i.e. the maps of RFs were topographically ordered with respect to most recent input symbols.

Analogously to the previous experiment, we compared quantization depths and topography preservation measures of RecSOMs and IFS+SOM in a large set of experimental runs with varying RecSOM parameters α and β. Results are shown in Table 2. As before, QD of RecSOM is always significantly higher, whereas the TP measure for RecSOM is higher except for cases of higher β/α ratio.

$\beta \setminus \alpha$	1.0		2.0		3.0	
0.2	5.07^{***}	1.44	5.52^{***}	1.77^{***}	4.92^{***}	1.63^{***}
0.7	5.49^{***}	0.75^{***}	5.92^{***}	1.41	5.81^{***}	1.59^{***}
1.0	5.66^{***}	0.55^{***}	6.75^{***}	1.16^{***}	6.73^{***}	1.42

Table 2: Means of the (QD, TP)measures, averaged over 40 simulations, for RecSOM trained on the laser data. Corresponding means for IFS+SOM were as follows: QD = 4.63 and TP = 1.42.

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

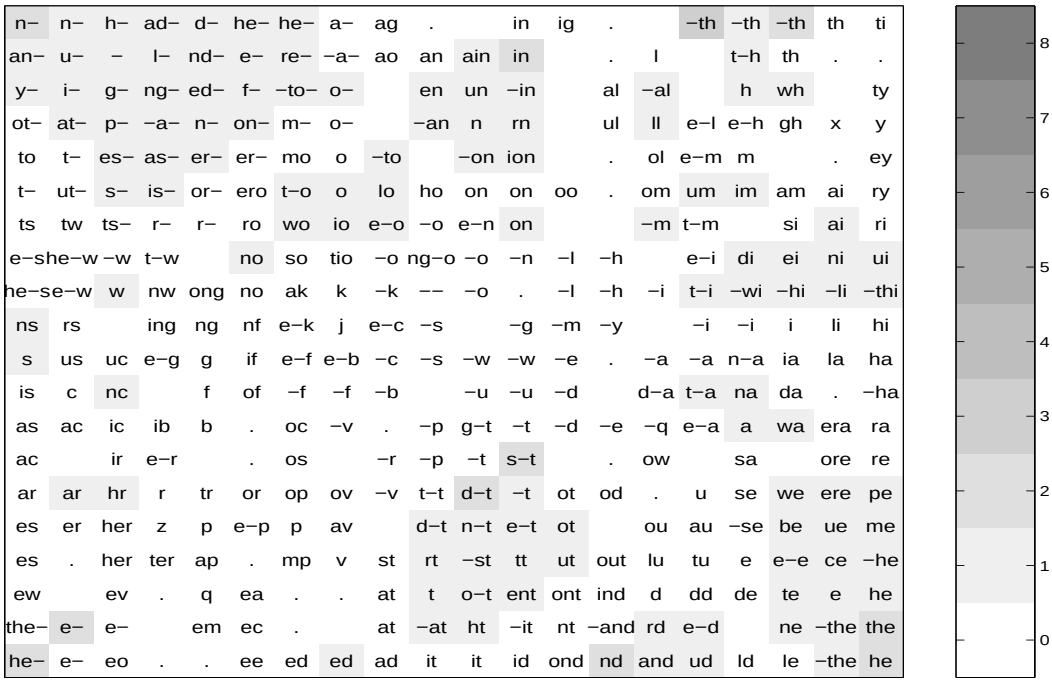

Figure 9: Receptive fields of RecSOM trained on English text. Dots denote units with empty RFs. Receptive field of each unit i is shaded according to the local topography preservation measure $\ell(\mathcal{R}_i)$.

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 (Voegtlin, 2002), 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 9.

Figure 10 illustrates asymptotic regimes of the fixed-input RecSOM dynamics (10) in terms of map activity differences between consecutive time steps[17]. We observed a variety

[17]Because 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 $i = 1, 2, ..., N$, 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

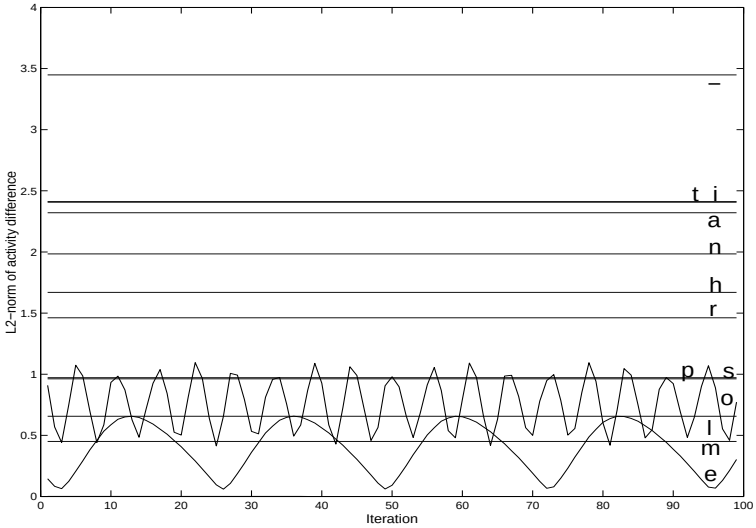

Figure 10: Fixed-input asymptotic dynamics of RecSOM after training on English text. Plotted are L_2 norms of the differences of map activities between the successive iterations. Labels denote the associated input symbols (for clarity, not all labels are shown).

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 aperiodic trajectory[18].

Dynamics of the winner units on the map induced by the fixed-input dynamics of $\mathbf{F}_s$ are shown in figure 11 (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 9 is shaped by the underlying collection of the fixed input dynamics of $\mathbf{F}_s$ (illustrated in figure 11 (left) through the induced dynamics on the map).

Theoretical upper bounds on β (eq. (40)) are shown in figure 12. Whenever for an input symbol s the bound $\Upsilon(s)$ is above $\beta = 0.7$ (dashed horizontal line) used to train RecSOM (symbols 'z', 'j', 'q', 'x'), we can be certain that the fixed input dynamics given by

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.

[18] A detailed investigation revealed that the same holds for the autonomous dynamics under symbol 'e' (even though this is less obvious by scanning figure 10).

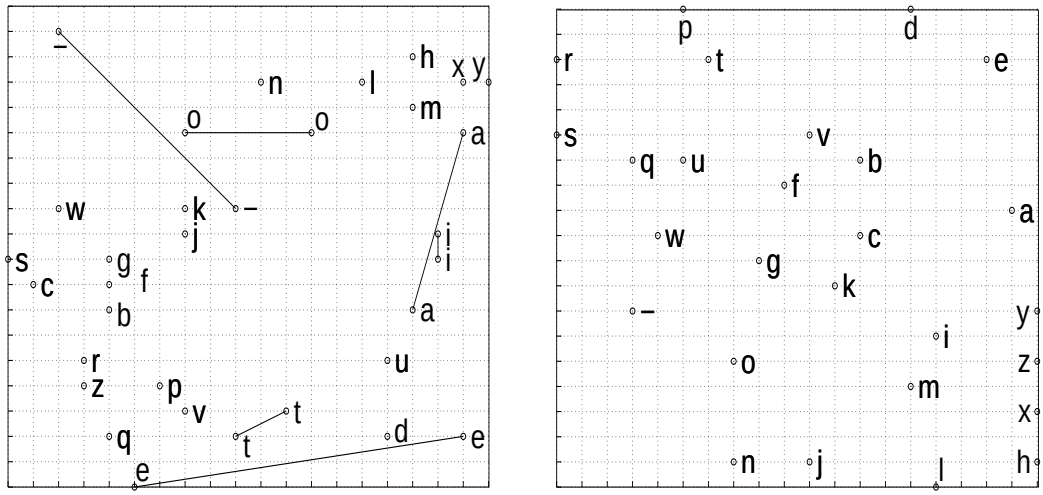

Figure 11: Dynamics of the winning units on the RecSOM (left) and IFS+SOM (right) maps induced by the fixed-input dynamics. The maps were trained on a corpus of written English ("Brave New World" by Aldous Huxley).

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 $\mathbf{F}_s$. We marked β-bounds for all symbols s with asymptotic fixed-input dynamics that goes beyond a single stable sink by an asterisk. Obviously, as seen in the previous experiments, $\Upsilon(s) < \beta$ does not necessarily imply more complicated fixed input dynamics on symbol s. However, in this case, for most symbols s with $\Upsilon(s) < \beta$, the associated fixed-input dynamics was indeed different from the trivial one dominated by a single attractive fixed point.

We also trained a standard SOM with 20×20 neurons on five-dimensional inputs created by the IFS[19] (45). The map is shown in figure 13. The induced dynamics on the map is illustrated in figure 11 (right). The suffix based organization of RFs is shaped by the underlying collection of autonomous attractive IFS dynamics.

Table 3 compares the QD and TP measures of the RecSOMs to IFS+SOM maps. In this case higher β/α ratios quickly lead to rather complicated fixed-input RecSOM dynamics, breaking the Markovian suffix-based RF organization of contractive maps. This has a negative effect on the QD and TP measures.

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

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

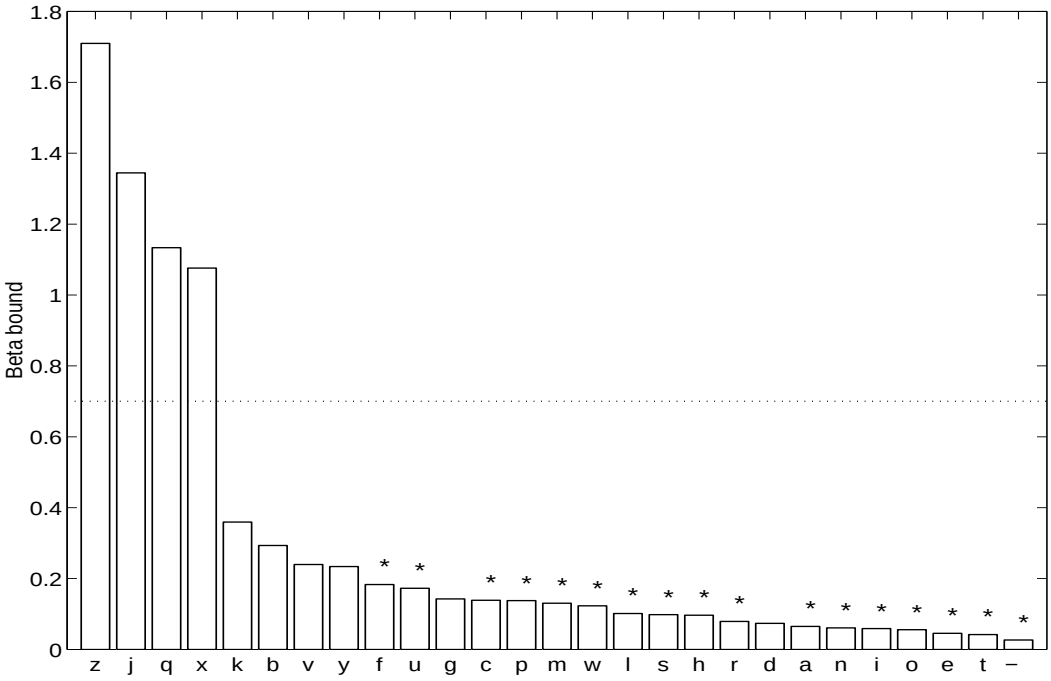

Figure 12: Theoretical bounds on β for RecSOM trained on the English text. β-bounds for all symbols with asymptotic fixed-input dynamics richer than a single stable sink are marked by an asterisk.

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) (Jeffrey, 1990; Oliver et al., 1993; Román-Roldan, Bernaola-Galván & Oliver, 1994; Fiser, Tusnady & Simon, 1994; Hao, 2000; Hao, Lee & Zhang, 2000; Tiňo, 2002). 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 (Tiňo, 2002)).

More importantly, it is the attractive character of the individual fixed-input IFS maps that shapes the Markovian organization of the state space. Imagine we feed the IFS with a long string $s_1...s_{p-2}s_{p-1}s_p...s_{r-2}s_{r-1}s_r...$ 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 $s_{1:p} = s_1...s_{p-2}s_{p-1}s_p$ and $s_{1:r} = s_1...s_{r-2}s_{r-1}s_r$ share a common suffix, the corresponding IFS states (see eqs. (45-46)), $s_{1:p}(\mathbf{x})$ 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

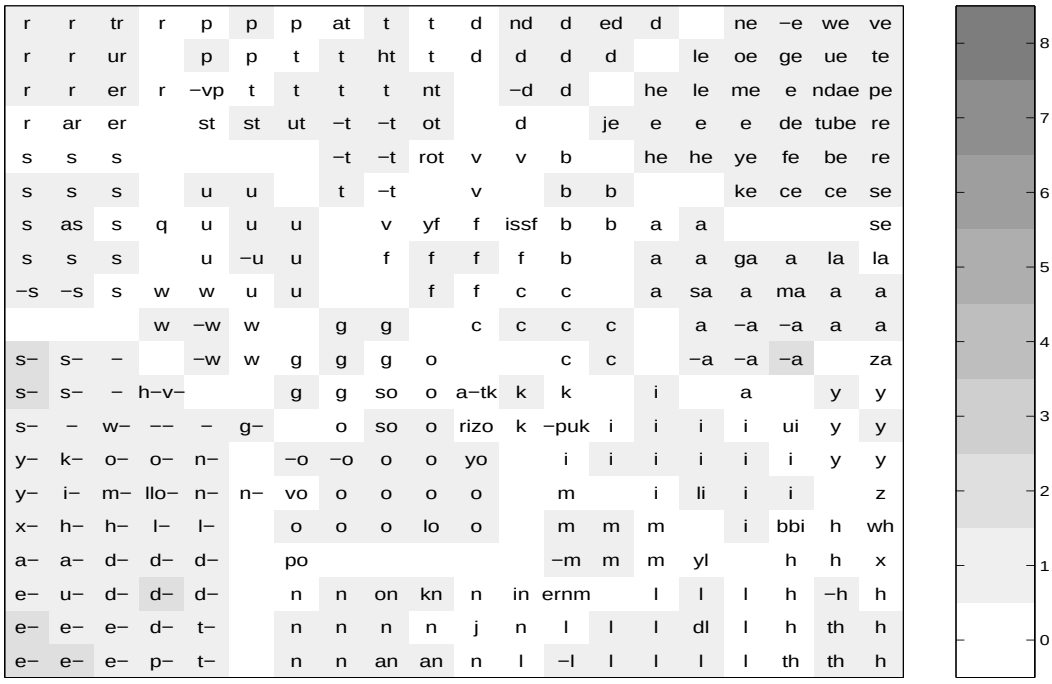

Figure 13: Receptive fields of a standard SOM with 20×20 units trained on IFS outputs, obtained on the English text. Topographic organization is observed with respect to the most recent symbols. Receptive field of each unit i is shaded according to the local topography preservation measure $\ell(\mathcal{R}_i)$.

position $\mathbf{x} \in [0,1]^m$, $m = \lceil \log_2 A \rceil$,

$$\|s_{1:p}(\mathbf{x}) - s_{1:r}(\mathbf{x})\| \leq k^L \sqrt{m}, \tag{47}$$

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

$$\begin{aligned} \mathbf{F}_{s_{1:n}}(\mathbf{y}) &= \mathbf{F}_{s_n}(\mathbf{F}_{s_{n-1}}(...(\mathbf{F}_{s_2}(\mathbf{F}_{s_1}(\mathbf{y})))...)) \\ &= (\mathbf{F}_{s_n} \circ \mathbf{F}_{s_{n-1}} \circ ... \circ \mathbf{F}_{s_2} \circ \mathbf{F}_{s_1})(\mathbf{y}). \end{aligned} \tag{48}$$

$\beta \setminus \alpha$	1.0		2.0		3.0	
0.2	1.68^{*}	0.30^{***}	2.03^{***}	0.58^{***}	2.04^{***}	0.64^{***}
0.7	1.15^{***}	0.10^{***}	1.89^{***}	0.30^{***}	1.93^{***}	0.39^{***}
1.0	0.92^{***}	0.06^{***}	1.66	0.21^{***}	1.81^{***}	0.31^{***}

Table 3: Means of the (QD, TP) measures, averaged over 40 simulations, for RecSOM trained on the language data. Corresponding means for IFS+SOM were as follows: QD = 1.65 and TP = 0.59.

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

$$\|\mathbf{F}_{s_{1:p}}(\mathbf{y}) - \mathbf{F}_{s_{1:r}}(\mathbf{y})\| \leq \rho_{max}^{L}\sqrt{N}, \tag{49}$$

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,[20]

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

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 follows from 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 symbols[21]).

Our experiments suggest that, compared with IFS+SOM maps, RecSOM maps with lower β/α ratio, e.g. RecSOM maps constructed with stronger emphasis on recently observed history of inputs, are capable of developing Markovian organizations of RFs with significantly superior memory depth and topography preservation (quantified by the QD (41) and TP (42) measures, respectively).

[20] or at least mode locations on neighboring grid points of the map

[21] in some rare cases even deeper

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 (Voegtlin, 2002), we always create a suffix based map of RFs. Such RF maps are designed to illustrate the temporal structure learnt 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. Such cases result in lower values of the topography preservation measure TP (42). 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 9. 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. Wiemer, 2003), 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 (2002) is the quantizer depth (QD) (eq. (41)). Another possible measure quantifying topology preservation on maps is the TP measure of equation (42). If coding efficiency of induced RFs and their topography preservation is a desirable property[22], then RecSOM with Markovian maps seem to be superior candidates to IFS+SOM models. In other words, having a trainable dynamic part in self-organizing maps has its merits. Indeed, in our experiments RecSOM

[22]Here we mean coding efficiency of RFs constrained by the two-dimensional map structure. Obviously, unconstrained codebooks will always lead to better coding efficiency.

maps with lower β/α ratio, lead to Markovian RF organizations with significantly superior QD and TP values.

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 appealing to notice in the RF map of figure 9 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 Linking RecSOM parameter β to Markovian RF organizations

RecSOM parameter β weighs the significance of importing information about possibly distant past into processing of sequential data. Intuitively, it is not surprising that when β is sufficiently small, e.g. when information about the very recent inputs dominates processing in RecSOM, the resulting maps will have Markovian flavour. This intuition was given a more rigorous form in section 3. Contractive fixed input mappings are likely to produce Markovian organizations of RFs on the RecSOM map. We have established theoretical bounds on parameter β that guarantee contractiveness of the fixed input maps. Using corollary 3.7, we obtain:

Corollary 5.1 *Provided*

$$\beta < \frac{e}{2N}, \tag{50}$$

irrespective of the input $\mathbf{s}$*, the map* $\mathbf{F_S}$ *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.

We experimentally tested the validity of the β bounds bellow which the fixed input dynamics was proved to be driven exclusively by attractive fixed points. Using 5×5 map grid ($N = 25$), with $\beta = 0.05$ (slightly below the bound (50)), and setting $\alpha = 1$, we ran a batch of RecSOM training sessions for each of the three data sets considered in this paper. The other model parameters were set as described in section 4. We evaluated the autonomous dynamics for each symbol, assessed by the L_2 norm of consecutive differences of the map activity profiles. In all cases, the differences vanished in less than 10 iterations. Because the β parameter was set bellow the theoretical bound (50), the training process could never induce an autonomous dynamics beyond the trivial one dominated by an attractive fixed point.

We constructed a bifurcation diagram for the RecSOM architecture described in section 4.4. After training, we varied β, while keeping other model parameters fixed. For each $0 \leq \beta \leq 3$ with step 0.01, we computed map activity differences between consecutive time steps during 100 iterations (after initial 400 preiterations). The activation differences for input symbol 'o' are shown in figure 14. Three dominant types of autonomous dynamics were observed: fixed point dynamics, period-2 attractors, and aperiodic oscillations (roughly for $0.6 < \beta < 1$). As expected, for small values of β, the dynamics is always governed by a unique fixed point. For higher values of β, the dynamics switches between periodic, fixed-point and aperiodic regimes[23].

We conclude by noting that when the inputs and input weights are taken from a set of diameter ξ, we have for the bound (40),

$$\frac{e}{2N} \leq \Upsilon(\mathbf{s}) \leq \frac{e}{2N} e^{2\alpha\xi^2}.$$

The lower bound follows from Corollary 5.1, the upper bound follows from minimizing $\|\mathbf{G}_\alpha(\mathbf{s})\|$ in (40).

5.4 Related work

It has been recently observed in (Hammer et al., 2004) that Markovian representations of sequence data occur naturally in topographic maps governed by leaky integration, such as Temporal Kohonen Map (Chappell & Taylor, 1993). Moreover, under some imposed circumstances, SOM for structured data (Hagenbuchner, Sperduti & Tsoi, 2003) 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 RecSOM map, potentially capable of very complicated dynamic representations, as non-autonomous dynamical systems governed by a collection of fixed input dynamics. Corollary 5.1 states

[23]The results are shown for one particular initial condition $\mathbf{y}(0)$. Qualitatively similar diagrams were obtained for a variety of initial conditions $\mathbf{y}(0)$.

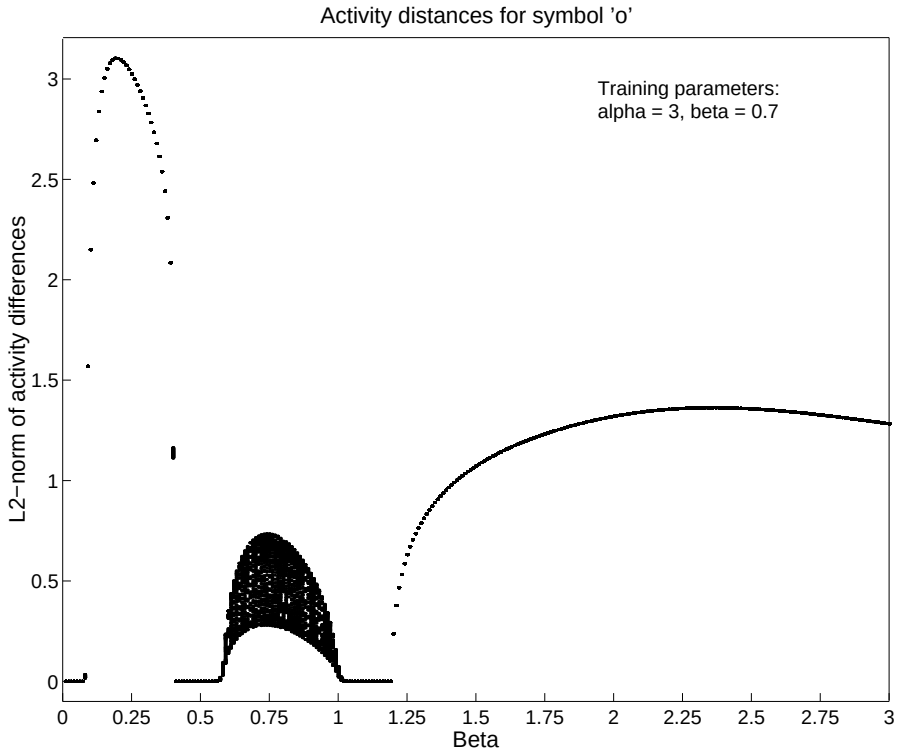

Figure 14: RecSOM map activity differences between consecutive time steps during 100 iterations (after initial 400 preiterations) for $0 \leq \beta \leq 3$, step size 0.01. The input is fixed to symbol 'o'. The map was trained on the language data as described in section 4.4.

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 (Hammer et al., 2004), the context influence has to be small for time series data to avoid instabilities in the model. Indeed, the RecSOM experiments of Hammer et al. (2004) (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 12.

Theoretical results of section 3 and corollary 5.1 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 (Voegtlin, 2002). Voegtlin also shows, using Taylor expansion arguments, that if $\beta < e/(2N)$, small perturbations of the activities will decay (fixed input maps are locally contractive). Our work extents this result to perturbations of arbitrary size[24]. Based on our analysis, we conclude that for each RecSOM model satisfying Voegtlin's stability bound on β, the fixed input dynamics for *any* input will be dominated by a unique attractive

[24]We are thankful to one of the anonymous reviewers who pointed this out.

fixed point. This renders the map Markovian quality and training stability.

Finally, we note that it has been shown that representation capabilities of merge SOM (Strickert & Hammer, 2003) and SOM for structured data (Hagenbuchner, Sperduti & Tsoi, 2003) operating on sequences transcend those of finite memory Markovian models in the sense that finite automata can be simulated (Strickert & Hammer, 2005; Hammer et al., 2004). It was assumed that there is no topological ordering among the units of the map. Also, the proofs are constructive in nature and it is not obvious that deeper memory automata structures can be actually learnt with Hebbian learning (Hammer et al. 2004). It should be emphasised that the type of analysis presented in this paper would not be feasible for the merge SOM and SOM for structured data models, since their fixed-input dynamics are governed by discontinuous mappings (due to discrete winner determination when calculating the context)[25].

5.5 Relation between IFS+SOM and recurrent SOM models

In this section we show that in the test mode (no learning), the IFS+SOM model acts exactly like the recurrent SOM (RSOM) model (Koskela et al., 1998). Given a sequence $s_1 s_2 ...$ 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

$$\mathbf{d}_i(t) = \nu \ (\mathbf{t}_{s_t} - \mathbf{w}_i) + (1 - \nu) \ \mathbf{d}_i(t-1), \tag{51}$$

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 (45) and (46))

$$\mathbf{x}(t) = k \ \mathbf{x}(t-1) + (1 - k) \ \mathbf{t}_{s_t}, \tag{52}$$

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

$$\mathbf{D}_i(t) = \mathbf{x}(t) - \mathbf{w}_i = k \ \mathbf{x}(t-1) + (1 - k) \ \mathbf{t}_{s_t} - \mathbf{w}_i. \tag{53}$$

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

$$\mathbf{D}_i(t) = k \ \mathbf{D}_i(t-1) + (1 - k) \ (\mathbf{t}_{s_t} - \mathbf{w}_i), \tag{54}$$

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

$$\mathbf{D}_i(t) = \nu \ (\mathbf{t}_{s_t} - \mathbf{w}_i) + (1 - \nu) \ \mathbf{D}_i(t-1). \tag{55}$$

[25] This was pointed out by one of the anonymous reviewers.

Provided $\nu = 1 - k$, the equations (51) and (55) 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) (Voegtlin, 2002), 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 Chappell & Taylor, 1993), recurrent SOM (Koskela et al., 1998), feedback SOM (Horio & Yamakawa, 2001), RecSOM (Voegtlin, 2002) and merge SOM (Strickert & Hammer, 2003), 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. However, allowing trainable feedback connections does seem to benefit the map formation, even in the Markovian case: compared with topographic maps fed by the fixed dynamic module, RecSOM maps are capable of developing Markovian organizations of receptive fields with significantly superior memory depth and topography preservation.

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.

Acknowledgements

Igor Farkaš and Peter Tiňo were supported by the Slovak Grant Agency for Science (#1/2045/05). Jort van Mourik was supported in part by the European Community's Human Potential Programme under contract number HPRN-CT-2002-00319. The authors are thankful to the anonymous reviewers for suggestions that helped to improve presentation of the paper.

References

Barnsley, M.F. (1988). *Fractals everywhere*. New York: Academic Press.

Chappell, G., & Taylor, J. (1993). The temporal Kohonen map. *Neural Networks*, 6, 441–445.

Barreto, de A.G., Araújo, A., & Kremer, S. (2003). A taxanomy of spatiotemporal connectionist networks revisited: The unsupervised case. *Neural Computation*, 15(6), 1255–1320.

Fiser, A., Tusnady, G., & Simon, I. (1994). Chaos game representation of protein structures. *Journal of Molecular Graphics*, 12(4), 302–304.

Hagenbuchner, M., Sperduti, A., & Tsoi, A. (2003). Self-organizing map for adaptive processing of structured data. *IEEE Transactions on Neural Networks*, 14(3), 491–505.

Hammer, B., Micheli, A., Sperduti, A., & Strickert, M. (2004). Recursive self-organizing network models. *Neural Networks*, 17(8-9), 1061–1085.

Hammer, B., Micheli, A., Strickert, M., & Sperduti, A. (2004a). A general framework for unsupervised processing of structured data. *Neurocomputing*, 57, 3–35.

Hammer, B., & Tiňo, P. (2003). Neural networks with small weights implement finite memory machines. *Neural Computation*, 15(8), 1897–1926.

Hao, B.-L. (2000). Fractals from genomes – exact solutions of a biology-inspired problem. *Physica A*, 282, 225–246.

Hao, B.-L., Lee, H., & Zhang, S. (2000). Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*, 11, 825–836.

Horio, K., & Yamakawa, T. (2001). Feedback self-organizing map and its application to spatio-temporal pattern classification. *International Journal of Computational Intelligence and Applications*, 1(1), 1–18.

James, D., & Miikkulainen, R. (1995). Sardnet: A self-organizing feature map for sequences. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Proceedings of the Advances in Neural Information Processing Systems*, Volume 7, (pp. 577–584). Morgan Kaufmann.

Jeffrey, J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8), 2163–2170.

Kenyon, R., & Peres, Y. (1996). Measures of full dimension on affine invariant sets. *Ergodic Theory and Dynamical Systems*, 16, 307–323.

Kohonen, T. (1982). Self–organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464–1479.

Koskela, T., Varsta, M., Heikkonen, J., & Kaski, K. (1998). Recurrent SOM with local linear models in time series prediction. In *6th European Symposium on Artificial Neural Networks* (pp. 167–172). D-facto Publications.

Lee, J., & Verleysen, M. (2002). Self-organizing maps with recursive neighborhood adaptation. *Neural Networks*, 15(8–9), 993–1003.

Oliver, J.L., Bernaola-Galván, P., Guerrero-Garcia, J., & Román-Roldan, R. (1993). Entropic profiles of dna sequences through chaos-game-derived images. *Journal of Theor. Biology*, (160), 457–470.

Principe, J., Euliano, N., & Garani, S. (2002). Principles and networks for self-organization in space-time. *Neural Networks*, 15(8–9), 1069–1083.

Román-Roldan, R., Bernaola-Galván, P., & Oliver, J. (1994). Entropic feature for sequence pattern through iteration function systems. *Pattern Recognition Letters*, 15, 567–573.

Schulz, R., & Reggia, J. (2004). Temporally asymmetric learning supports sequence processing in multi-winner self-organizing maps. *Neural Computation*, 16(3), 535–561.

Strickert, M., & Hammer, B. (2003). Neural gas for sequences. In T. Yamakawa (Ed.), *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)* (pp. 53–57). Kyushu Institute of Technology.

Strickert, M., & Hammer, B. (2005). Merge som for temporal data. *Neurocomputing*, 64, 39–72.

Tiňo, P. (2002). Multifractal properties of Hao's geometric representations of DNA sequences. *Physica A: Statistical Mechanics and its Applications*, 304(3–4), 480–494.

Tiňo, P., & Dorffner, G. (2001). Predicting the future of discrete sequences from fractal representations of the past. *Machine Learning*, 45(2), 187–218.

Tiňo, P., & Hammer, B. (2004). Architectural bias in recurrent neural networks: Fractal analysis. *Neural Computation*, 15(8), 1931–1957.

Tiňo, P., & Köteles, M. (1999). Extracting finite state representations from recurrent neural networks trained on chaotic symbolic sequences. *IEEE Transactions on Neural Networks*, 10(2), 284–302.

Tiňo, P., Čerňanský, M., & Beňušková, L. (2004). Markovian architectural bias of recurrent neural networks. *IEEE Transactions on Neural Networks*, 15(1), 6–15.

Voegtlin, T. (2002). Recursive self-organizing maps. *Neural Networks*, 15(8–9), 979–992.

Wiemer, J. (2003). The time-organized map algorithm: Extending the self-organizing map to spatiotemporal signals. *Neural Computation*, 16, 1143–1171.

Yin, H. (2002). ViSOM - a novel method for multivariate data projection and structure visualisation. *IEEE Transactions on Neural Networks*, 13(1), 237–243.