

Architectural Bias in Recurrent Neural Networks - Fractal Analysis

Peter Tiño

School of Computer Science

University of Birmingham

Edgbaston, Birmingham B15 2TT, UK

Barbara Hammer

Department of Mathematics/ComputerScience

University of Osnabrück

D-49069 Osnabrück, Germany

Abstract

We have recently shown that when initialized with “small” weights, recurrent neural networks (RNNs) with standard sigmoid-type activation functions are inherently biased towards Markov models, i.e. even prior to any training, RNN dynamics can be readily used to extract finite memory machines (Hammer & Tiño, 2002; Tiño, Čerňanský & Beňušková, 2002; Tiño, Čerňanský & Beňušková, 2002a). Following Christiansen and Chater (1999), we refer to this phenomenon as the *architectural bias of RNNs*. In this paper we further extend

our work on the architectural bias in RNNs by performing a rigorous fractal analysis of recurrent activation patterns. We assume the network is driven by sequences obtained by traversing an underlying finite-state transition diagram – a scenario that has been frequently considered in the past e.g. when studying RNN-based learning and implementation of regular grammars and finite-state transducers. We obtain lower and upper bounds on various types of fractal dimensions, such as box-counting and Hausdorff dimensions. It turns out that not only can the recurrent activations inside RNNs with small initial weights be explored to build Markovian predictive models, but also the activations form fractal clusters the dimension of which can be bounded by the scaled entropy of the underlying driving source. The scaling factors are fixed and are given by the RNN parameters.

1 Introduction

There is a considerable amount of literature devoted to connectionist processing of sequential symbolic structures (e.g. (Kremer, 2001)). For example, researchers have been interested in formulating models of human performance in processing linguistic patterns of various complexity (e.g. (Christiansen & Chater, 1999)). Recurrent neural networks (RNNs) constitute a well established approach for dealing with linguistic data, and the capability of RNNs of processing finite automata and some context-free and context-sensitive languages is well known (see e.g. (Bodén & Wiles, 2002; Forcada & Carrasco, 1995; Tiño et al., 1998)). The underlying dynamics which emerges through training has been investigated in the work of Blair, Bodén, Pollack and others (Blair & Pollack, 1997; Bodén & Blair, 2002; Rodriguez, Wiles & Elman,

1999). It turns out that even though theoretically RNNs are able to explore a wide range of possible dynamical regimes, trained RNNs often achieve complicated behavior by combining appropriate fixed point dynamics, including attractive fixed points as well as saddle points.

It should be mentioned that RNNs can be canonically generalized to so-called recursive networks for processing tree structures and directed acyclic graphs, hence providing a very attractive connectionist framework for symbolic or structural data (Frasconi, Gori & Sperduti, 1998; Hammer, 2002).

For both, recurrent and recursive networks, the interior connectionist representation of input data given by the hidden neuron's activation profile is of particular interest. The hidden neuron's activation might allow the detection of important information about relevant substructures of the inputs for the respective task as has been demonstrated for example for quantitative structure activity relationship prediction tasks with recursive networks (Micheli et al., 2002). The hidden neuron's activations often show considerable profiles such as clusters and structural differentiation. For RNNs, hidden nodes' activation profile allows to infer parts of the underlying dynamical components which account for the network behavior (Blair & Pollack, 1997).

However, it has been known for some time that when training RNNs to process symbolic sequences, activations of recurrent units display a considerable amount of structural differentiation even *prior to learning* (Christiansen & Chater, 1999; Kolen, 1994; Kolen, 1994a; Manolios & Fanelli, 1994). Following Christiansen and Chater (1999), we refer to this phenomenon as the *architectural bias of RNNs*. Now the question arises: which parts of structural differentiation are due to the learning process and hence are likely to encode information related to the specific learning task, and which parts emerge au-

tomatically, even without training, due to the architectural bias of RNNs.

We have recently shown, both empirically and theoretically, the meaning of the architectural bias for RNNs: when initialized with “small” weights, RNNs with standard sigmoid-type activation functions are inherently biased towards Markov models, i.e. even prior to any training, RNN dynamics can be readily used to extract finite memory machines (Hammer & Tiño, 2002; Tiño, Čerňanský & Beňušková, 2002; Tiño, Čerňanský & Beňušková, 2002a). In this study we further extend our work by rigorously analyzing the “size” of recurrent activation patterns in such RNNs. Since the activation patterns are of fractal nature, their size is expressed through fractal dimensions (e.g. (Falconer, 1990)). The dimensionality of hidden neuron’s activation patterns provides one characteristic for comparing different networks and the effect of learning algorithms on the network’s interior connectionistic representation of data. We concentrate on the case where the RNN is driven by sequences generated from an underlying finite-state automaton – a scenario frequently studied in the past in the context of RNN-based learning and implementation of regular grammars and finite-state transducers (Casey, 1996; Cleeremans, Servan-Schreiber & McClelland, 1989; Elman, 1990; Forcada & Carrasco, 1995; Frasconi et al., 1996; Giles et al., 1992; Manolios & Fanelli, 1994; Tiño & Šajda, 1995; Watrous & Kuhn, 1992). We will derive bounds on the dimensionality of the recurrent activation patterns which reflect complexity of the input source and the bias caused by the recurrent architecture initialized with small weights.

The paper has the following organization: After a brief introduction to fractal dimensions and automata-directed iterated function systems (IFS) in Sections 2 and 3, respectively, we prove in Section 4 dimension estimates for invariant sets of general automata-directed IFSs composed of non-similarities.

In Section 5 we reformulate RNNs driven by sequences over finite alphabets as IFSs. Section 6 is devoted to dimension estimates of recurrent activations inside contractive RNNs. The discussion in Section 7 puts our findings into the context of previous work. We confront the theoretically calculated dimension bounds with empirical fractal dimension estimates of recurrent activations in Section 8. Section 9 concludes the paper by summarizing the key messages of this study.

2 Fractal dimensions

In this section we briefly introduce various notions of “size” for geometrically complicated objects called fractals. For a more detailed information, we refer the interested reader to e.g. (Falconer, 1990).

Consider a metric space X . Let K be a totally bounded subset of X . For each $\delta > 0$, define $N_\delta(K)$ to be the smallest number of sets of diameter $\leq \delta$ that can cover K (δ -fine cover of K). Since K is bounded, $N_\delta(K)$ is finite for each $\delta > 0$. The rate of increase of $N_\delta(K)$ as $\delta \rightarrow 0$ tells us something about the “size” of K .

Definition 2.1 *The upper and lower box-counting dimensions of K are defined as*

$$\dim_B^+ K = \limsup_{\delta \rightarrow 0} \frac{\log N_\delta(K)}{-\log \delta} \quad \text{and} \quad \dim_B^- K = \liminf_{\delta \rightarrow 0} \frac{\log N_\delta(K)}{-\log \delta}, \quad (1)$$

respectively.

Definition 2.2 *Let $s > 0$. For $\delta > 0$, define*

$$\mathcal{H}_\delta^s(K) = \inf_{\Gamma_\delta(K)} \sum_{B \in \Gamma_\delta(K)} (\text{diam } B)^s, \quad (2)$$

where the infimum is taken over the set $\Gamma_\delta(K)$ of all countable δ -fine covers of K . Define

$$\mathcal{H}^s(K) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(K).$$

The Hausdorff dimension of the set K is

$$\dim_H K = \sup\{s \mid \mathcal{H}^s(K) = \infty\} = \inf\{s \mid \mathcal{H}^s(K) = 0\}. \quad (3)$$

The Hausdorff dimension is more “subtle” than the box-counting dimensions in the sense that the former can capture details not detectable by the latter. In a way, box-counting dimensions can be thought of as indicating the efficiency with which a set may be covered by small sets of equal size. In contrast, Hausdorff dimension involves coverings by sets of small but perhaps widely varying size (Falconer, 1990). It is well known that

$$\dim_H K \leq \dim_B^- K \leq \dim_B^+ K. \quad (4)$$

3 Finite automata and automata-directed IFS

We now introduce some terminology related to iterated function systems (IFS) with an emphasis on IFS driven by sequences obtained by traversing finite automata. First, we briefly recall basic notions related to sequences over finite alphabets.

3.1 Symbolic sequences over a finite alphabet

Consider a finite alphabet $\mathcal{A} = \{1, 2, \dots, A\}$. The sets of all finite (non-empty) and infinite sequences over \mathcal{A} are denoted by \mathcal{A}^+ and \mathcal{A}^ω , respectively. Sequences from \mathcal{A}^ω are also referred to as ω -words. Denoting by λ the empty word, we write $\mathcal{A}^* = \mathcal{A}^+ \cup \{\lambda\}$. The set of all sequences consisting of a finite,

or an infinite number of symbols from \mathcal{A} is then $\mathcal{A}^\infty = \mathcal{A}^+ \cup \mathcal{A}^\omega$. The set of all sequences over \mathcal{A} with exactly n symbols (n -blocks) is denoted by \mathcal{A}^n . The number of symbols in a sequence w is denoted by $|w|$.

Let $w = s_1s_2\dots \in \mathcal{A}^\infty$ and $i \leq j$. By w_i^j we denote the string $s_i s_{i+1} \dots s_j$, with $w_i^i = s_i$. For $i > j$, $w_i^j = \lambda$. If $|w| = n \geq 1$, then $w^- = w_1^{n-1}$ is obtained¹ by omitting the last symbol of w .

A partial order \leq may be defined on \mathcal{A}^* as follows: write $w_1 \leq w_2$ if and only if w_1 is a prefix of w_2 , i.e. there is some $w_3 \in \mathcal{A}^*$, such that $w_2 = w_1 w_3$. Two strings w_1, w_2 are incomparable, if neither $w_1 \leq w_2$, nor $w_2 \leq w_1$. For each finite string $c \in \mathcal{A}^*$, the cylinder $[c]$ is the set of all infinite strings $w \in \mathcal{A}^\omega$ that begin with c , i.e. $[c] = \{w \in \mathcal{A}^\omega \mid c \leq w\}$.

3.2 Iterated Function Systems

Let X be a complete metric space. A (contractive) iterated function system (IFS) (Barnsley, 1988) consists of A contractions $f_a : X \rightarrow X$, $a \in \{1, 2, \dots, A\} = \mathcal{A}$, operating on X . In the basic setting, at each time step all the maps f_a are used to transform X in an iterative manner:

$$X_0 = X, \tag{5}$$

$$X_{t+1} = \bigcup_{a \in \mathcal{A}} f_a(X_t), \quad t \geq 0. \tag{6}$$

There is a unique compact invariant set $K \subset X$, $K = \bigcup_{a \in \mathcal{A}} f_a(K)$, called the attractor of the IFS. Actually, it can be shown that the sequence $\{X_t\}_{t=0}^\infty$ converges² to K . For any ω -word $w = s_1s_2\dots \in \mathcal{A}^\omega$, the images of X under

¹if $|w| = 1$, then $w^- = \lambda$.

²under a Hausdorff distance on the power set of X

compositions of IFS maps,

$$(f_{s_1}(f_{s_2}(\dots(f_{s_n}(X))\dots))) = f_{w_1^n}(X)$$

converge to an element of K as $n \rightarrow \infty$.

More complicated IFS systems can be devised e.g. by constraining the sequences of maps that can be applied to X in the process of defining the attractive invariant set K . One possibility is to demand that the words w corresponding to allowed compositions of IFS maps belong to a language over \mathcal{A} (Prusinkiewicz & Hammel, 1992). A specific example of this are recurrent IFS (Barnsley, Elton & Hardin, 1989), where the allowed words w are specified by a topological Markov chain. In this paper we concentrate on IFSs with map compositions restricted to sequences that can be read-out by traversing labeled state-transition graphs (see e.g. (Culik & Dube, 1993)).

3.3 IFS associated with state-transition graphs

Consider a labeled directed multigraph $\mathcal{G} = (V, E, \kappa)$, where the elements $v \in V$ and $e \in E$ are the *vertices* (or nodes) and *edges* of the multigraph \mathcal{G} , respectively. Each edge $e \in E$ is labeled by a symbol $\kappa(e) \in \mathcal{A}$ from the input alphabet $\mathcal{A} = \{1, 2, \dots, A\}$, as prescribed by the map $\kappa : E \rightarrow \mathcal{A}$. The map κ can be naturally extended to operate on subsets E' of E and sequences γ over E :

$$\text{for } E' \subseteq E, \quad \kappa(E') = \bigcup_{e \in E'} \kappa(e), \quad (7)$$

$$\text{for } \gamma = e_1 e_2 \dots, \quad \kappa(\gamma) = \kappa(e_1) \kappa(e_2) \dots \in \mathcal{A}^\infty. \quad (8)$$

The set $E_{u \rightarrow v} \subseteq E$ contains all edges from $u \in V$ to $v \in V$. $E_{u \rightarrow} = \bigcup_{v \in V} E_{u \rightarrow v}$ is the set of all edges starting at the vertex u .

A *path* in the graph is a sequence $\gamma = e_1 e_2 \dots$ of edges, such that the terminal vertex of each edge e_i is the initial vertex of the next edge e_{i+1} . The initial vertex of e_1 is the initial vertex $ini(\gamma)$ of γ . For a finite path $\gamma = e_1 e_2 \dots e_n$ of length n , the terminal vertex of e_n is the terminal vertex $term(\gamma)$ of γ . The sequence of labels read out along the path γ is $\kappa(\gamma)$. As in the case of sequences over \mathcal{A} , for $i \leq j \leq n$, γ_i^j denotes the path $e_i e_{i+1} \dots e_j$, and γ^- denotes the path γ without the last edge e_n , i.e. $\gamma^- = \gamma_1^{n-1}$.

We write $E_{u \rightarrow v}^{(n)}$ for the set of all paths of length n with initial vertex u and terminal vertex v . The set of all paths of length n with initial vertex u is denoted by $E_{u \rightarrow}^{(n)}$. Analogously, $E_{u \rightarrow v}^{(*)}$ denotes the set of all finite paths from u to v ; $E_{u \rightarrow}^{(*)}$ and $E_{u \rightarrow}^{(\omega)}$ denote the sets of all finite and infinite paths, respectively, starting at vertex u .

Definition 3.1 A strictly contracting state-transition graph (SCSTG) is a labeled directed multigraph $\mathcal{G} = (V, E, \kappa)$ together with a number $0 < r_a < 1$ for each symbol a from the label alphabet \mathcal{A} .

We will assume that the state-transition graph \mathcal{G} is both *deterministic*, i.e. for each vertex u there are no two distinct paths initiated at u with the same label read-out, and *strongly connected*, i.e. there is a path from any vertex to any other.

Definition 3.2 An iterated function system (IFS) associated with a SCSTG $(V, E, \kappa, \{r_a\}_{a \in \mathcal{A}})$ consists of a complete metric space $(X, \|\cdot\|)$ ³ and a set of A contractions $f_a : X \rightarrow X$, one for each input symbol $a \in \mathcal{A}$, such that for all $x, y \in X$,

$$\|f_a(x) - f_a(y)\| \leq r_a \|x - y\|, \quad a \in \mathcal{A}. \quad (9)$$

³the metric is expressed through a norm

The allowed compositions of maps f_a are controlled by the sequences of labels associated with the paths in $\mathcal{G} = (V, E, \kappa)$. The IFS image of a point $x \in X$ under a (finite) path $\gamma = e_1 e_2 \dots e_n$ in \mathcal{G} is

$$\gamma(x) = (f_{\kappa(e_1)}(f_{\kappa(e_2)}(\dots(f_{\kappa(e_n)}(x))\dots))) = (f_{\kappa(e_1)} \circ f_{\kappa(e_2)} \circ \dots \circ f_{\kappa(e_n)})(x). \quad (10)$$

The contraction factor $r(\gamma)$ of the path γ is calculated as

$$r(\gamma) = \prod_{i=1}^n r_{\kappa(e_i)}. \quad (11)$$

For the empty path λ , $\lambda(x) = x$ and $r(\lambda) = 1$. For an infinite path $\gamma = e_1 e_2 \dots$, the corresponding image of a point $x \in X$ is

$$\gamma(x) = \lim_{n \rightarrow \infty} (f_{\kappa(e_1)} \circ f_{\kappa(e_2)} \circ \dots \circ f_{\kappa(e_n)})(x). \quad (12)$$

The maps $\gamma(\cdot)$ are extended to subsets X' of X as follows: $\gamma(X') = \{\gamma(x) \mid x \in X'\}$. As the length of paths γ in \mathcal{G} increases, the points $\gamma(x)$ tend to closely approximate the attractive set of the IFS. The attractor has now a more intricate structure than in the case of the basic “unrestricted” IFS. There is a list of invariant compact sets $K_u \subseteq X$, one for each vertex $u \in V$ (Edgar & Golds, 1999), such that

$$K_u = \bigcup_{v \in V} \bigcup_{e \in E_{u \rightarrow v}} f_{\kappa(e)}(K_v). \quad (13)$$

Each set K_u contains (possibly non-linearly) compressed copies of the sets K_v .

We denote the union of the sets K_u by K ,

$$K = \bigcup_{u \in V} K_u. \quad (14)$$

4 Dimension estimates for K_u

In this section we develop a theory of dimension estimates for the invariant sets K_u of IFSs driven by sequences obtained by traversing deterministic finite

automata. In particular, we ask whether upper and lower bounds on Lipschitz coefficients of contractive IFS maps f_a translate into bounds on dimensions of K_u . In answering this question, we extend the results of Falconer (1990), who considered the original IFS setting (IFSs driven by \mathcal{A}^ω), and Edgar & Golds (1999), who studied graph-directed IFSs (GDIFS). In GDIFS one associates each edge in the underlying graph with a distinct IFS map, so the calculations are somewhat more straightforward. However, many of the ideas applied in proofs in (Edgar & Golds, 1999) were transferable to proofs in this section.

4.1 Upper bounds

Definition 4.1 *Let $(V, E, \kappa, \{r_a\}_{a \in \mathcal{A}})$ be a SCSTG. For each $s > 0$, denote by $\mathbf{M}(s)$ a square matrix (rows and columns are indexed by vertices⁴ V) with entries*

$$M_{uv}(s) = \sum_{e \in E_{u \rightarrow v}} (r_{\kappa(e)})^s. \quad (15)$$

The unique⁵ number s_1 such that the spectral radius of $\mathbf{M}(s_1)$ is equal to 1 is called the dimension of the SCSTG $(V, E, \kappa, \{r_a\}_{a \in \mathcal{A}})$.

Theorem 4.2 *Let $\mathcal{G} = (V, E, \kappa)$ be a labeled directed multigraph and $(X, \{f_a\}_{a \in \mathcal{A}})$ an IFS associated with the SCSTG $(\mathcal{G}, \{r_a\}_{a \in \mathcal{A}})$. Let s_1 be the dimension of the SCSTG $(\mathcal{G}, \{r_a\}_{a \in \mathcal{A}})$. Then for all $u \in V$, $\dim_B^+ K_u \leq s_1$, and also $\dim_B^+ K \leq s_1$.*

⁴we slightly abuse mathematical notation by identifying the vertices of \mathcal{G} with integers $1, 2, \dots, |V|$.

⁵Uniqueness follows from the Perron-Frobenius theory (see e.g. (Minc, 1988)).

Proof: By the Perron-Frobenius theory of nonnegative irreducible⁶ matrices (Minc, 1988), the spectral radius (equal to 1) of the matrix $\mathbf{M}(s_1)$ is actually an eigenvalue of $\mathbf{M}(s_1)$ with the associated positive eigenvector $\{\lambda_u\}_{u \in V}$ satisfying

$$\lambda_u = \sum_{v \in V} \lambda_v \sum_{e \in E_{u \rightarrow v}} (r_{\kappa(e)})^{s_1}, \quad \sum_{u \in V} \lambda_u = 1, \quad \lambda_u > 0. \quad (16)$$

Given a vertex $u \in V$, it is possible to define a measure μ on the set of cylinders

$$C_u = \{[w] \mid w \in \kappa(E_{u \rightarrow}^{(*)})\} \quad (17)$$

by postulating:

$$\text{for } w = \kappa(\gamma), \gamma \in E_{u \rightarrow v}^{(*)}; \quad \mu([w]) = \lambda_v \cdot r(\gamma)^{s_1}. \quad (18)$$

Indeed, for $\gamma \in E_{u \rightarrow v}^{(*)}$, $w = \kappa(\gamma)$, we have

$$\mu([w]) = \sum_{e \in E_{v \rightarrow}} \mu([w\kappa(e)]).$$

The measure μ extends to Borel measures on $E_{u \rightarrow}^{(\omega)}$ (Edgar & Golds, 1999).

Consider a vertex $u \in V$. Fix a positive number δ . We define a (cross-cut) set

$$T_\delta = \{\gamma \mid \gamma \in E_{u \rightarrow}^{(*)}, r(\gamma) < \delta \leq r(\gamma^-)\}. \quad (19)$$

Note that T_δ is a finite set such that for every infinite path $\gamma' \in E_{u \rightarrow}^{(\omega)}$ there is exactly one prefix length n such that $(\gamma')_1^n \in T_\delta$. The sets T_δ and $\kappa(T_\delta)$ partition the sets $E_{u \rightarrow}^{(\omega)}$ and $\kappa(E_{u \rightarrow}^{(\omega)})$, respectively. The collection

$$\mathcal{C}_u = \{\gamma(K_{term(\gamma)}) \mid \gamma \in T_\delta\} \quad (20)$$

⁶Matrix $\mathbf{M}(s_1)$ is irreducible because the underlying graph (V, E, κ) is strongly connected.

covers the set K_u with sets of diameter less than $\eta = \delta \cdot \text{diam}(X)$.

In order to estimate $N_\eta(K_u)$, which is upper bounded by the cardinality $\text{card}(T_\delta)$ of T_δ , we write

$$\lambda_u = \mu(\kappa(E_{u \rightarrow}^{(\omega)})) = \sum_{\gamma \in T_\delta} \lambda_{\text{term}(\gamma)} \cdot (r(\gamma))^{s_1}. \quad (21)$$

By (19), we have for each $\gamma = e_1 e_2 \dots e_n \in T_\delta$,

$$r(\gamma) = r(\gamma^-) \cdot r(e_n) \geq \delta \cdot r_{\min}$$

and so from (21) we obtain

$$\lambda_{\max} \geq \lambda_u \geq \text{card}(T_\delta) \cdot \lambda_{\min} \cdot (\delta r_{\min})^{s_1}, \quad (22)$$

where $\text{card}(T_\delta)$ is the cardinality of T_δ and

$$\lambda_{\max} = \max_{v \in \mathcal{V}} \lambda_v, \quad \lambda_{\min} = \min_{v \in \mathcal{V}} \lambda_v \quad \text{and} \quad r_{\min} = \min_{a \in \mathcal{A}} r_a. \quad (23)$$

Hence,

$$N_\eta(K_u) \leq \text{card}(T_\delta) \leq \frac{\lambda_{\max}}{\lambda_{\min}} (\delta r_{\min})^{-s_1} = \frac{\lambda_{\max}}{\lambda_{\min}} \left(\frac{r_{\min}}{\text{diam}(X)} \right)^{-s_1} \eta^{-s_1}.$$

It follows that

$$\log N_\eta(K_u) \leq -s_1 \log \eta + \text{Const}, \quad (24)$$

where Const is a constant term with respect to δ . Dividing (24) by $-\log \eta > 0$, and letting the diameter η of the covering sets diminish, we obtain

$$\dim_B^+ K_u = \limsup_{\eta \rightarrow 0} \frac{\log N_\eta(K_u)}{-\log \eta} \leq s_1. \quad (25)$$

If we denote the number of vertices in V by $\text{card}(V)$, then the number of sets with diameter less than η needed to cover the set K can be upper bounded by

$$\begin{aligned} N_\eta(K) &\leq \text{card}(V) \cdot \max_{u \in V} N_\eta(K_u) \\ &\leq \text{card}(V) \frac{\lambda_{\max}}{\lambda_{\min}} \left(\frac{r_{\min}}{\text{diam}(X)} \right)^{-s_1} \eta^{-s_1} \end{aligned}$$

and we again obtain

$$\log N_\eta(K) \leq -s_1 \log \eta + \text{Const}' \quad (26)$$

where Const' is a constant term with respect to δ . Repeating the above argument leads to $\dim_B^+ K \leq s_1$. \square

We can derive a less tight bound, that is closely related to the theories of symbolic dynamics and formal languages over the alphabet \mathcal{A} .

The *adjacency matrix* \mathbf{G} of a labeled directed multigraph $\mathcal{G} = (V, E, \kappa)$ is defined element-wise as follows⁷:

$$G_{uv} = \text{card}(E_{u \rightarrow v}). \quad (27)$$

Theorem 4.3 *Let $\mathcal{G} = (V, E, \kappa)$ be a labeled directed multigraph and $(X, \{f_a\}_{a \in \mathcal{A}})$ an IFS associated with the SCSTG $(\mathcal{G}, \{r_a\}_{a \in \mathcal{A}})$. Let $r_{\max} = \max_{a \in \mathcal{A}} r_a$. Then for all $u \in V$,*

$$\dim_B^+ K_u \leq s_1 \leq \frac{\log \rho(\mathbf{G})}{-\log r_{\max}},$$

where $\rho(\mathbf{G})$ is the maximum eigenvalue of \mathbf{G} .

Proof: Consider a SCSTG $(\mathcal{G}, r_a = r_{\max}, a \in \mathcal{A})$ with a fixed contraction rate r_{\max} for all symbols in \mathcal{A} . The matrix (15) for this SCSTG has the form $\mathbf{N}(s) = r_{\max}^s \mathbf{G}$. The spectral radius of $\mathbf{N}(s)$, which coincides with the maximum eigenvalue of $\mathbf{N}(s)$, is $\rho(\mathbf{N}(s)) = r_{\max}^s \cdot \rho(\mathbf{G})$, where $\rho(\mathbf{G})$ is the spectral radius (maximum eigenvalue) of \mathbf{G} . The dimension s_{\max} associated with such a SCSTG is the solution of $\rho(\mathbf{N}(s)) = 1$. In particular, $r_{\max}^s \cdot \rho(\mathbf{G}) =$

⁷Although we do not allow $s = 0$ in the definition of $\mathbf{M}(s)$, \mathbf{G} can be formally thought of as a matrix $\mathbf{M}(s)$ computed with s set to 0.

1, and so

$$s_{max} = \frac{\log \rho(\mathbf{G})}{-\log r_{max}}. \quad (28)$$

Next we show that $s_{max} \geq s_1$.

Let $\mathbf{M}(s_1)$ be the matrix (15) of spectral radius 1 for the original SCSTG $(\mathcal{G}, \{r_a\}_{a \in \mathcal{A}})$.

Define $\Delta = s_{max} - s_1$. Note that $\mathbf{N}(s_{max}) = r_{max}^\Delta \cdot \mathbf{N}(s_1)$.

Since for all $a \in \mathcal{A}$, $r_{max} \geq r_a$, we have $0 < \mathbf{M}(s_1) \leq \mathbf{N}(s_1)$, where $\mathbf{M} \leq \mathbf{N}$ for two positive matrices \mathbf{M}, \mathbf{N} of the same size means that \leq applies element-wise, i.e. $0 \leq M_{uv} \leq N_{uv}$. Hence, $\rho(\mathbf{M}(s_1)) \leq \rho(\mathbf{N}(s_1))$.

Observe that

$$\rho(\mathbf{N}(s_{max})) = 1 = \rho(r_{max}^\Delta \cdot \mathbf{N}(s_1)) = r_{max}^\Delta \cdot \rho(\mathbf{N}(s_1)).$$

But $\rho(\mathbf{N}(s_1)) \geq \rho(\mathbf{M}(s_1)) = 1$, so $r_{max}^\Delta \leq 1$ must hold. Since $r_{max} < 1$, we have that $\Delta \geq 0$, which means $s_{max} \geq s_1$. \square

4.2 Lower bounds

Imagine that besides the Lipschitz bounds (9) for the IFS maps $\{f_a\}_{a \in \mathcal{A}}$, we also have lower bounds: for all $x, y \in X$,

$$\|f_a(x) - f_a(y)\| \geq r'_a \|x - y\|, \quad r'_a > 0, \quad a \in \mathcal{A}. \quad (29)$$

We will show, that under certain conditions, the bounds (29) induce lower bounds on the Hausdorff dimension of the sets K_u . Roughly speaking, to estimate lower bounds on the dimension of K_u 's, we need to make sure that our lower bounds on covering set size are not invalidated by trivial overlaps between the IFS basis sets.

Definition 4.4 Let $\mathcal{G} = (V, E, \kappa)$ be a labeled directed multigraph. We say that an IFS $(X, \{f_a\}_{a \in \mathcal{A}})$ associated with the SCSTG $(\mathcal{G}, \{r_a\}_{a \in \mathcal{A}})$ falls into the disjoint case if, for any $u, v, v' \in V$, $e \in E_{u \rightarrow v}$, $e' \in E_{u \rightarrow v'}$, $e \neq e'$, we have⁸ $f_{\kappa(e)}(K_v) \cap f_{\kappa(e')}(K_{v'}) = \emptyset$.

Theorem 4.5 Consider a labeled directed multigraph $\mathcal{G} = (V, E, \kappa)$ and an IFS $(X, \{f_a\}_{a \in \mathcal{A}})$ associated with the SCSTG $(\mathcal{G}, \{r_a\}_{a \in \mathcal{A}})$. Suppose the IFS falls into the disjoint case. Let s_2 be the dimension of the SCSTG $(\mathcal{G}, \{r'_a\}_{a \in \mathcal{A}})$. Then for each vertex $u \in V$, $\dim_H K_u \geq s_2$.

Proof: First, define a metric on the power set of X as

$$d(B_1, B_2) = \inf_{x \in B_1} \inf_{y \in B_2} \|x - y\|, \quad B_1, B_2 \subseteq X.$$

Because the IFS falls into the disjoint case and the sets K_u , $u \in V$, are compact, there is some $\zeta > 0$ such that for all $u, v, v' \in V$, $e \in E_{u \rightarrow v}$, $e' \in E_{u \rightarrow v'}$, $e \neq e'$, we have

$$d(f_{\kappa(e)}(K_v), f_{\kappa(e')}(K_{v'})) > \zeta. \quad (30)$$

Fix a vertex $u \in V$ and consider two paths γ_1, γ_2 from $E_{u \rightarrow}^{(*)}$ such that the corresponding label sequences $\kappa(\gamma_1)$ and $\kappa(\gamma_2)$ are incomparable. Let w be the longest common prefix of $\kappa(\gamma_1)$, $\kappa(\gamma_2)$ and let γ be the initial subpath of γ_1 that supports the label sequence w , i.e. $\gamma = (\gamma_1)_1^{|w|}$. We have $|w| < |\kappa(\gamma_1)|$, and so $w \leq \kappa(\gamma_1)^-$, which means $r'(\gamma) \geq r'(\gamma_1^-)$. The paths γ_1 and γ_2 can be written as $\gamma_1 = \gamma e \gamma'_1$ and $\gamma_2 = \gamma e' \gamma'_2$, where the edges e and e' are distinct and $\kappa(e) \neq \kappa(e')$ (\mathcal{G} is deterministic). The paths γ'_1 , γ'_2 may happen to be empty. In any case,

$$\gamma'_1(K_{\text{term}(\gamma_1)}) \subseteq K_{\text{term}(e)}, \quad \gamma'_2(K_{\text{term}(\gamma_2)}) \subseteq K_{\text{term}(e')}$$

⁸By determinicity of \mathcal{G} , $\kappa(e) \neq \kappa(e')$.

and we have

$$d(e\gamma'_1(K_{term(\gamma_1)}), e'\gamma'_2(K_{term(\gamma_2)})) \geq d(e(K_{term(e)}), e'(K_{term(e')})) > \zeta.$$

This leads to

$$d(\gamma_1(K_{term(\gamma_1)}), \gamma_2(K_{term(\gamma_2)})) = \tag{31}$$

$$d(we\gamma'_1(K_{term(\gamma_1)}), we'\gamma'_2(K_{term(\gamma_2)})) \geq \tag{32}$$

$$r'(\gamma) \cdot d(e\gamma'_1(K_{term(\gamma_1)}), e'\gamma'_2(K_{term(\gamma_2)})) > \tag{33}$$

$$r'(\gamma) \cdot \zeta \geq r'(\gamma_1^-) \cdot \zeta \tag{34}$$

Now, as in the upper bound case (Section 4.1), we construct a matrix $\mathbf{M}'(s)$,

$$M'_{uv}(s) = \sum_{e \in E_{u \rightarrow v}} (r'_{\kappa(e)})^s, \tag{35}$$

Let s_2 be the unique number such that the spectral radius of $\mathbf{M}'(s_2)$ is equal to 1. Again, 1 is an eigenvalue of $\mathbf{M}'(s_2)$ with the associated eigenvector $\{\lambda'_u\}_{u \in V}$ satisfying

$$\lambda'_u = \sum_{v \in V} \lambda'_v \sum_{e \in E_{u \rightarrow v}} (r'_{\kappa(e)})^{s_2}, \quad \sum_{u \in V} \lambda'_u = 1, \quad \lambda'_u > 0. \tag{36}$$

As before, we define a measure μ' on the set of cylinders C_u (see (17)) as follows:

$$\text{for } w = \kappa(\gamma), \gamma \in E_{u \rightarrow v}^{(*)}; \quad \mu'([w]) = \lambda'_v \cdot r'(\gamma)^{s_2}. \tag{37}$$

The measure μ' extends to Borel measures on $E_{u \rightarrow}^{(\omega)}$.

Fix a (Borel) set $B \subset K_u$ and define

$$\delta = \frac{\text{diam}(B)}{\zeta}. \tag{38}$$

The corresponding cross-cut set is

$$T'_\delta = \{\gamma \mid \gamma \in E_{u \rightarrow}^{(*)}, r'(\gamma) < \delta \leq r'(\gamma^-)\}. \tag{39}$$

From (38) and the definition of T'_δ , we have

$$\text{for each } \gamma \in T'_\delta; \quad \text{diam}(B) = \zeta \cdot \delta \leq \zeta \cdot r'(\gamma^-). \quad (40)$$

The label strings $\kappa(\gamma)$ associated with the paths γ in T'_δ are incomparable, implying that there is at most one path $\gamma \in T'_\delta$ such that $\gamma(K_{\text{term}(\gamma)}) \cap B \neq \emptyset$. To see this, note that by (31)–(34) and (40), for any two paths $\gamma_1, \gamma_2 \in T'_\delta$, we have

$$d(\gamma_1(K_{\text{term}(\gamma_1)}), \gamma_2(K_{\text{term}(\gamma_2)})) > r'(\gamma_1^-) \cdot \zeta \geq \text{diam}(B).$$

Denote by \tilde{B} the set of all ω -sequences associated with infinite paths starting in the vertex u that map X into a point in B , i.e.

$$\tilde{B} = \{\kappa(\gamma) \mid \gamma \in E_u^{(\omega)}, \gamma(X) \in B\}. \quad (41)$$

Since the cross-cut set T'_δ partitions the set $E_u^{(\omega)}$, there exists some $\gamma_B \in T'_\delta$ such that $\gamma_B(K_{\text{term}(\gamma_B)})$ meets $B \subseteq K_u$. In fact, as argued above, it is the only such path from T'_δ , and so $\tilde{B} \subseteq [\kappa(\gamma_B)]$. In terms of measure (see Eq. (37)),

$$\mu'(\tilde{B}) \leq \mu'([\kappa(\gamma_B)]) \leq \lambda'_{\max} \cdot r'(\gamma_B)^{s_2}, \quad (42)$$

where $\lambda'_{\max} = \max_{v \in \mathcal{V}} \lambda'_v$. From the definition of T'_δ we have

$$\text{diam}(B) = \zeta \cdot \delta > \zeta \cdot r'(\gamma_B).$$

Using this inequality in (42) we obtain

$$\mu'(\tilde{B}) \leq \lambda'_{\max} \left(\frac{\text{diam}(B)}{\zeta} \right)^{s_2}. \quad (43)$$

Now, for *any* countable cover \mathcal{C}_u of K_u by Borel sets we have

$$\sum_{B \in \mathcal{C}_u} (\text{diam}(B))^{s_2} \geq \frac{\zeta^{s_2}}{\lambda'_{\max}} \sum_{B \in \mathcal{C}_u} \mu'(\tilde{B}) = \frac{\zeta^{s_2}}{\lambda'_{\max}} \sum_{B \in \mathcal{C}_u} \nu(B) \geq \frac{\zeta^{s_2}}{\lambda'_{\max}} \nu(K_u), \quad (44)$$

where ν is the pushed forward measure on K_u induced by the measure μ' . Note that this is a correct construction, since the IFS falls into the disjoint case.

Consulting the Definition 2.2 and noting that it is always the case that

$$\mathcal{H}^{s_2}(K_u) \geq \frac{\zeta^{s_2}}{\lambda_{max}^{s_2}} \nu(K_u) > 0,$$

we conclude that $\dim_H K_u \geq s_2$. \square

As in the case of upper bounds, we transform the results of the previous theorem into a weaker, but more accessible lower bound.

Theorem 4.6 *Under the assumptions of Theorem 4.5, for each vertex $u \in V$,*

$$\dim_H K_u \geq s_2 \geq \frac{\log \rho(\mathbf{G})}{-\log r_{min}},$$

where $r_{min} = \min_{a \in \mathcal{A}} r_a$.

Proof: As in the proof of Theorem 4.3, consider a SCSTG $(\mathcal{G}, r_a = r_{min}, a \in \mathcal{A})$. The matrix (15) for this SCSTG is $\mathbf{N}(s) = r_{min}^s \mathbf{G}$ with $\rho(\mathbf{N}(s)) = r_{min}^s \cdot \rho(\mathbf{G})$. The dimension s_{min} associated with such a SCSTG is

$$s_{min} = \frac{\log \rho(\mathbf{G})}{-\log r_{min}}. \quad (45)$$

Define $\Delta = s_{min} - s_2$. We have $\mathbf{N}(s_{min}) = r_{min}^\Delta \cdot \mathbf{N}(s_2)$. Since $r_{min} \leq r_a$, for all $a \in \mathcal{A}$, it holds $0 < \mathbf{N}(s_2) \leq \mathbf{M}(s_2)$, and so $\rho(\mathbf{N}(s_2)) \leq \rho(\mathbf{M}(s_2))$.

Note that

$$\rho(\mathbf{N}(s_{min})) = 1 = r_{min}^\Delta \cdot \rho(\mathbf{N}(s_2))$$

and $\rho(\mathbf{N}(s_2)) \leq \rho(\mathbf{M}(s_2)) = 1$. It follows that $r_{min}^\Delta \geq 1$ must hold. Since $r_{min} < 1$, we have $\Delta \leq 0$, which means that $s_{min} \leq s_2$. \square

4.3 Discussion of the quantities involved in the bounds

The numbers s_1 and s_2 are well-known quantities in the literature on fractal sets generated by constructions employing maps with different contraction rates (see e.g. (Fernau & Staiger, 1994; Fernau & Staiger, 2001; Edgar & Golds, 1999)). Since there is no closed-form analytical solution for s_1 as introduced in Definition 4.1, the dimensions s_1, s_2 are defined only implicitly. Loosely speaking, spectral radius of $\mathbf{M}(s)$ is related to the variation of the “cover function” $\mathcal{H}^s(K)$ (see Definition 2.2) and the only way of getting a finite change-over point s in Eq. (3) is to set the spectral radius to 1.

The situation changes when we confine ourselves to a constant contraction rate across all IFS maps. As seen in the proof of Theorem 4.3, in this case there is a closed-form solution for s . The solution is equal to the scaled spectral radius of the adjacency matrix, which determines the growth rate of allowed label strings as the sequence length increases. The major difficulty with having different contraction rates in the IFS maps is that the areas on the fractal support that need to be covered by sets of decreasing diameters cannot be simply estimated by counting the number of different label sequences of certain length that can be obtained by traversing the underlying multigraph.

5 Recurrent networks as IFS

In this paper we concentrate on first-order (discrete-time) recurrent neural networks RNNs of N recurrent neurons driven with dynamics

$$x_n(t) = g \left(\sum_{j=1}^D v_{nj} i_j(t) + \sum_{j=1}^N w_{nj} x_j(t-1) + d_n \right), \quad (46)$$

where $x_j(t)$ and $i_j(t)$ are elements, at time t , of the state and input vectors, $\mathbf{x}(t) = (x_1(t), \dots, x_N(t))^T$ and $\mathbf{i}(t) = (i_1(t), \dots, i_D(t))^T$, respectively, w_{nj} and v_{nj} are recurrent and input connection weights, respectively, d_n 's constitute the bias vector $\mathbf{d} = (d_1, \dots, d_N)^T$, and $g(\cdot)$ is an injective non-constant differentiable activation function of bounded derivative from \mathbb{R} to a bounded interval $\Omega \subset \mathbb{R}$ of length $|\Omega|$. Denoting by \mathbf{V} and \mathbf{W} the $N \times D$ and $N \times N$ weight matrices (v_{nj}) and (w_{nj}) , respectively, we rewrite (46) in matrix form

$$\mathbf{x}(t) = G(\mathbf{V}\mathbf{i}(t) + \mathbf{W}\mathbf{x}(t-1) + \mathbf{d}), \quad (47)$$

where $G : \mathbb{R}^N \rightarrow \Omega^N$ is the element-wise application of g .

Assume RNN is processing strings over the finite alphabet $\mathcal{A} = \{1, 2, \dots, A\}$. Symbols $a \in \mathcal{A}$ are presented at the network input as unique D -dimensional codes $\mathbf{c}_a \in \mathbb{R}^D$, one for each symbol $a \in \mathcal{A}$. RNN can be viewed as a non-linear IFS consisting of a collection of A maps $\{f_a\}_{a=1}^A$ acting on Ω^N ,

$$f_a(\mathbf{x}) = (G \circ T_a)(\mathbf{x}), \quad (48)$$

where

$$T_a(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{d}_a \quad (49)$$

and

$$\mathbf{d}_a = \mathbf{V}\mathbf{c}_a + \mathbf{d}. \quad (50)$$

For a set $B \subseteq \mathbb{R}^N$, we denote by $[B]_i$ the slice of B defined as

$$[B]_i = \{x_i \mid \mathbf{x} = (x_1, \dots, x_N)^T \in B\}, \quad i = 1, 2, \dots, N. \quad (51)$$

When symbol $a \in \mathcal{A}$ is at the network input, the range of possible net-in activations on recurrent units is the set

$$R_a = \bigcup_{1 \leq i \leq N} [T_a(\Omega^N)]_i. \quad (52)$$

Recall that singular values $\alpha_1, \dots, \alpha_N$ of the matrix \mathbf{W} are positive square roots of the eigenvalues of $\mathbf{W}\mathbf{W}^T$. The singular values are lengths of the (mutually perpendicular) principal semi-axes of the image of the unit ball under the linear map defined by the matrix \mathbf{W} . We assume that \mathbf{W} is non-singular and adopt the convention that $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_N > 0$. Denote by $\alpha_{max}(\mathbf{W})$ and $\alpha_{min}(\mathbf{W})$ the largest and the smallest singular values of \mathbf{W} , respectively, i.e. $\alpha_{max}(\mathbf{W}) = \alpha_1$ and $\alpha_{min}(\mathbf{W}) = \alpha_N$. The next lemma gives us conditions under which the maps f_a are contractive.

Lemma 5.1 *If*

$$r_a^{max} = \alpha_{max}(\mathbf{W}) \cdot \sup_{z \in R_a} |g'(z)| < 1, \quad (53)$$

then the map f_a , $a \in \mathcal{A}$ (Eq. (48)), is contractive.

Proof: The result follows from two facts:

1. A map f from a metric space $(U, \|\cdot\|_U)$ to a metric space $(V, \|\cdot\|_V)$ is contractive, if it is Lipschitz continuous, i.e. for all $\mathbf{x}, \mathbf{y} \in U$, $\|f(\mathbf{x}) - f(\mathbf{y})\|_V \leq r_f \|\mathbf{x} - \mathbf{y}\|_U$, and the Lipschitz constant r_f is smaller than 1.
2. The Lipschitz constant r of a composition $(f_1 \circ f_2)$ of two Lipschitz continuous maps f_1 and f_2 with Lipschitz constants r_1 and r_2 is equal to $r = r_1 \cdot r_2$.

Note that $\alpha_{max}(\mathbf{W})$ is a Lipschitz constant of the affine maps T_a and that $\sup_{z \in R_a} |g'(z)|$ is a Lipschitz constant of the map G with domain $T_a(\Omega^N)$. \square

By arguments similar to those in the proof of Lemma 5.1, we get

Lemma 5.2 *For*

$$r_a^{min} = \alpha_{min}(\mathbf{W}) \cdot \inf_{z \in R_a} |g'(z)|, \quad (54)$$

it holds: $\|f_a(\mathbf{x}) - f_a(\mathbf{y})\| \geq r_a^{min} \|\mathbf{x} - \mathbf{y}\|$, for all $\mathbf{x}, \mathbf{y} \in \Omega^N$.

6 RNNs driven by finite automata

Major research effort has been devoted to the study of learning and implementation of regular grammars and finite-state transducers using RNNs (Casey, 1996; Cleeremans, Servan-Schreiber & McClelland, 1989; Elman, 1990; Forcada & Carrasco, 1995; Frasconi et al., 1996; Giles et al., 1992; Manolios & Fanelli, 1994; Tiño & Šajda, 1995; Watrous & Kuhn, 1992). In most cases, the input sequences presented at the input of RNNs were obtained by traversing the underlying finite state automaton/machine. Of particular interest to us are approaches that did not reset the RNN state after presentation of each input string, but learned the appropriate resetting behavior during the training process e.g. (Forcada & Carrasco, 1995; Tiño & Šajda, 1995). In such cases RNNs are fed by a concatenation of input strings over some input alphabet \mathcal{A}' , separated by a special end-of-string symbol $\# \notin \mathcal{A}'$.

We can think of such sequences as finite substrings of infinite sequences generated by traversing the underlying state-transition graph extended by adding edges labeled by $\#$ that terminate in the initial state. The added edges start at all the states – in case of transducers, or at the final states – in case of acceptors.

Let $\mathcal{G} = (V, E, \kappa)$ represent such an extended state-transition graph. Let $\mathcal{A} = \mathcal{A}' \cup \{\#\}$. If for all $a \in \mathcal{A}$, $r_a^{max} < 1$ (see Lemma 5.1), the recurrent network can then be viewed as an IFS associated with a SCSTG $(\mathcal{G}, \{r_a^{max}\}_{a \in \mathcal{A}})$ (Definition 3.2) operating on the RNN state space Ω^N .

When driving RNNs with symbolic sequences generated by traversing a

strongly connected state-transition graph \mathcal{G} , by action of the IFS $\{f_a\}_{a \in \mathcal{A}}$ we get a set of recurrent activations $\mathbf{x}(t)$ that tend to group in well-separated clusters (Kolen, 1994; Kolen, 1994a; Manolios & Fanelli, 1994). In case of contractive RNNs (that are capable of emulating finite-memory machines, see (Hammer & Tiño, 2002; Tiño, Čerňanský & Beňušková, 2002; Tiño, Čerňanský & Beňušková, 2002a)), the activations $\mathbf{x}(t)$ approximate the invariant attractor sets $K_u \subseteq \Omega^N$, $u \in V$. In this situation, we can get size estimates for the activation clusters.

Define (see Eq. (52))

$$\ell = \sup_{z \in \bigcup_a R_a} |g'(z)| = \max_{a \in \mathcal{A}} \sup_{z \in R_a} |g'(z)| \quad (55)$$

$$q = \min_{a, a' \in \mathcal{A}, a \neq a'} \|\mathbf{V}(\mathbf{c}_a - \mathbf{c}_{a'})\|. \quad (56)$$

Theorem 6.1 *Let RNN (47) be driven with sequences generated by traversing a strongly connected state-transition graph $\mathcal{G} = (V, E, \kappa)$ with adjacency matrix \mathbf{G} . Suppose $\alpha_{max}(\mathbf{W}) < \ell^{-1}$. Let s_1 be the dimension of the SCSTG $(\mathcal{G}, \{r_a^{max}\}_{a \in \mathcal{A}})$. Then the (upper box-counting) fractal dimensions of attractors K_u , $u \in V$, approximated by the RNN activation vectors $\mathbf{x}(t)$, are upper-bounded by*

$$\forall u \in V; \quad \dim_B^+ K_u \leq s_1 \leq \frac{\log \rho(\mathbf{G})}{-\log r_{max}} \quad (57)$$

and also

$$\dim_B^+ K \leq s_1 \leq \frac{\log \rho(\mathbf{G})}{-\log r_{max}},$$

where $r_{max} = \max_{a \in \mathcal{A}} r_a^{max}$.

Proof: Since $\alpha_{max}(\mathbf{W}) < \ell^{-1}$, by (55) and Lemma 5.1, each IFS map f_a , $a \in \mathcal{A}$, is a contraction with contraction coefficient r_a^{max} .

One is now tempted to invoke Theorems 4.2 and 4.3. Note however that in the theory developed in Sections 3 and 4, for allowed paths γ in \mathcal{G} , the compositions of IFS maps $\gamma(\cdot)$ are applied in the *reversed manner*⁹ (see Eq. (10)), i.e. if $\kappa(\gamma) = w = s_1 s_2 \dots s_n \in \mathcal{A}^+$, $\gamma(\mathbf{x}) = (f_{s_1} \circ f_{s_2} \circ \dots \circ f_{s_n})(\mathbf{x})$. Actually, this does not pose any difficulty, since we can work with a new state-transition graph $\mathcal{G}^R = (V, E^R, \kappa^R)$ associated with $\mathcal{G} = (V, E, \kappa)$. \mathcal{G}^R is completely the same as \mathcal{G} up to orientation of the edges. For every edge $e \in E_{u \rightarrow v}$ with label $\kappa(e)$ there is an associated edge $e^R \in E_{v \rightarrow u}^R$ labeled by $\kappa^R(e^R) = \kappa(e)$.

The matrices $\mathbf{M}(s)$ (15) corresponding to IFSs based on \mathcal{G} are just transposed versions of the matrices $\mathbf{M}^R(s)$ corresponding to IFSs based on \mathcal{G}^R . However, spectral radius is invariant with respect to the transpose operation, so the results of Section 4 can be directly employed. \square

We now turn to lower bounds.

Theorem 6.2 *Consider a RNN (47) driven with sequences generated by traversing a strongly connected state-transition graph $\mathcal{G} = (V, E, \kappa)$ with adjacency matrix \mathbf{G} . Suppose*

$$\alpha_{max}(\mathbf{W}) < \min \left\{ \frac{1}{\ell}, \frac{q}{\sqrt{N} |\Omega|} \right\}. \quad (58)$$

Let s_2 be the dimension of the SCSTG $(\mathcal{G}, \{r_a^{min}\}_{a \in \mathcal{A}})$. Then the (Hausdorff) dimensions of attractors K_u , $u \in V$, approximated by the RNN activation vectors $\mathbf{x}(t)$, are lower-bounded by

$$\forall u \in V; \quad \dim_H K_u \geq s_2 \geq \frac{\log \rho(\mathbf{G})}{-\log r_{min}}, \quad (59)$$

⁹This approach, usual in the IFS literature, is convenient because of the convergence properties of $\gamma(X)$, $\gamma \in \mathcal{A}^\omega$.

where $r_{min} = \min_{a \in \mathcal{A}} r_a^{min}$.

Proof: Note that $diam(\Omega^N) = \sqrt{N} \cdot |\Omega|$. Furthermore, by (58), we have

$$q > \alpha_{max}(\mathbf{W}) \cdot \sqrt{N} \cdot |\Omega| = \alpha_{max}(\mathbf{W}) \cdot diam(\Omega^N).$$

It follows that

$$q = \min_{a \neq a'} \|\mathbf{V}(\mathbf{c}_a - \mathbf{c}_{a'})\| = \min_{a \neq a'} \|\mathbf{d}_a - \mathbf{d}_{a'}\| > \alpha_{max}(\mathbf{W}) \cdot diam(\Omega^N)$$

and so

$$\text{for all } a, a' \in \mathcal{A}, a \neq a'; \quad T_a(\Omega^N) \cap T_{a'}(\Omega^N) = \emptyset.$$

Since G is injective, we also have

$$\forall a, a' \in \mathcal{A}, a \neq a'; \quad (G \circ T_a)(\Omega^N) \cap (G \circ T_{a'})(\Omega^N) = \emptyset,$$

which, by (48), implies $f_a(\Omega^N) \cap f_{a'}(\Omega^N) = \emptyset$, for all $a, a' \in \mathcal{A}, a \neq a'$. Hence, the IFS $\{f_a\}_{a \in \mathcal{A}}$ falls into the disjoint case. We can now invoke Theorems 4.5 and 4.6. \square

Using (4), we summarize the bounds in the following corollary.

Corollary 6.3 *Under the assumptions of Theorems 6.1 and 6.2, for all $u \in V$,*

$$\frac{\log \rho(\mathbf{G})}{-\log r_{min}} \leq s_2 \leq \dim_H K_u \leq \dim_B^- K_u \leq \dim_B^+ K_u \leq s_1 \leq \frac{\log \rho(\mathbf{G})}{-\log r_{max}}. \quad (60)$$

7 Discussion

Recently, we have extended the work of Kolen and others (e.g. (Christiansen & Chater, 1999; Kolen, 1994; Kolen, 1994a; Manolios & Fanelli, 1994)) by pointing out that in dynamical tasks of symbolic nature, when initialized with

“small” weights, recurrent neural networks (RNN), with e.g. sigmoid activation functions, form contractive IFS. In particular, the dynamical systems (48) are driven by a single attractive fixed point, one for each $a \in \mathcal{A}$ (Tiño, Horne & Giles, 2001; Tiño et al., 1998). Actually, this type of simple dynamics is a reason for starting to train recurrent networks from small weights. Unless one has a strong prior knowledge about the network dynamics (Giles & Omilin, 1993), the sequence of bifurcations leading to a desired network behavior may be hard to achieve when starting from an arbitrarily complicated network dynamics (Doya, 1992).

The tendency of RNNs to form clusters in the state space *before* training was first reported by Servan-Schreiber, Cleeremans and McClelland (1989), later by Kolen (1994; 1994a) and Christiansen & Chater (1999). We have shown that RNNs randomly initiated with small weights are inherently biased towards Markov models, i.e. even prior to any training, RNN dynamics can be readily used to extract finite memory machines (Hammer & Tiño, 2002; Tiño, Čerňanský & Beňušková, 2002; Tiño, Čerňanský & Beňušková, 2002a). In other words, even prior to training, the recurrent activation clusters are perfectly reasonable and are biased towards finite-memory computations.

This paper further extends our work by showing that in such cases, a rigorous analysis of fractal encodings in the RNN state space can be performed. We have derived (lower and upper) bounds on several types of fractal dimensions, such as Hausdorff and (upper and lower) box-counting dimensions of activation patterns occurring in the RNN state space when the network is driven by an underlying finite-state automaton, as has been the case in many studies reported in the literature (Casey, 1996; Cleeremans, Servan-Schreiber & McClelland, 1989; Elman, 1990; Forcada & Carrasco, 1995; Frasconi et al., 1996;

Giles et al., 1992; Manolios & Fanelli, 1994; Tiño & Šajda, 1995; Watrous & Kuhn, 1992).

Our results have a nice information theoretic interpretation. The entropy of a language $L \subseteq \mathcal{A}^*$ is defined as the rate of exponential increase in the number of distinct words belonging to L , as the word length increases. With respect to the scenario we have in mind, i.e. RNN driven by traversing a finite-state automaton, a more appropriate context is that of regular ω -languages. Roughly speaking, a regular ω -language is a set $L \subseteq \mathcal{A}^\omega$ of infinite strings over \mathcal{A} that, when parsing a (traditional) finite-state automaton, end up visiting the set of final states infinitely often¹⁰ (e.g. (Thomas, 1990)). The entropy of a regular ω -language L is equal to the entropy of the set of all finite prefixes of the words belonging to L and this is equal to $\log \rho(\mathbf{G})$, where \mathbf{G} is the adjacency matrix of the underlying finite state automaton (Fernau & Staiger, 2001). Hence, the architectural bias phenomenon has an additional flavor: not only can the recurrent activations inside RNNs with small weights be explored to build Markovian predictive models, but also the activations form fractal clusters the dimension of which can be upper- (and under some conditions lower-) bounded by the scaled entropy of the underlying driving source. The scaling factors are fixed and are given by the RNN parameters.

Our work is related to valuations of languages, e.g. (Fernau & Staiger, 1994). Briefly, a valuation of a symbol $a \in \mathcal{A}$ can be thought of as a “contraction ratio” r_a of the associated IFS map f_a . Valuations of finite words over \mathcal{A} are then obtained by postulating that in general the valuation β is a monoid morphism mapping $(\mathcal{A}^*, +, \lambda)$ to $((0, \infty), \cdot, 1)$, where $+$ and \cdot denote the oper-

¹⁰This is by no means the only acceptance criterion that has been investigated. For a survey see (Thomas, 1990).

ations of word concatenation and real number multiplication, respectively (see Eq. (11)). It should be stressed that, strictly speaking, not all the maps f_a need to be contractions. There can be words with valuation ≥ 1 . It is important, however, to have a control over recursive structures. For example, in our setting of sequences drawn by traversing finite-state diagrams, it is desirable that valuations of cycles be < 1 . Actually, in this paper we could have softened in this manner our demand on contractiveness of each IFS map. However, this would unnecessarily complicate the presentation. In addition, we are dealing with the architectural bias in RNNs where all the maps f_a are contractive anyway. Our Hausdorff dimension estimates are related to β -entropy of languages. However, valuations in principle assume that maps f_a are similarities. In this paper we deal with a more general case of non-similarities, since the RNN maps f_a are non-linear. In addition, the theory of valuations operates with contraction ratios only, but in order to obtain lower bounds on the Hausdorff dimension of the IFS attractor one has to consider other details of the IFS maps.

Results in this paper extend the work of Blair, Bodén, Pollack and others (e.g. (Blair & Pollack, 1997; Bodén & Blair, 2002; Rodriguez, Wiles & Elman, 1999)) analyzing fixed point representations of context-free languages in trained RNNs. They observed that the networks often developed an intricate combination of attractive and saddle-type fixed points. Compared with Cantor-set-like representations emerging from attractive-point-only dynamics (the case studied here), the inclusion of saddle points can lead to better generalization on longer strings. Our theory describes the complexity of recurrent patterns in the early stages of RNN training, before bifurcating into dynamics enriched with saddle points.

Finally, our fractal dimension estimates can prove helpful in neural approaches to the (highly non-trivial) inverse problem of fractal image compression through IFSs (e.g. (Melnik & Pollack; 1998; Stucki & Pollack, 1992)): Given an image \mathcal{I} , identify an IFS that generates \mathcal{I} . The inverse fractal problem is an area of active research with many alternative approaches proposed in the literature. However, a general algorithm for solving the inverse fractal problem is still missing (Melnik & Pollack; 1998). As the neural implementations of IFS maps are contractions and the feeding input label sequences are generated by Bernoulli sources, the fractal dimension estimates derived in this paper can be used e.g. to assess in an evolutionary setting the fitness of neural candidates for generating the target image with a known pre-computed fractal dimension. Such assessment will be most appropriate in the early stages of population evolution and is potentially much cheaper (see Section 8) than having to produce at each step and for each candidate iteratively generated images and then compute their Hausdorff distances to the target image.

8 Experiments

Even though our main result is of theoretical nature – the size of recurrent activations in the early stages of RNN training can be upper- (and sometimes lower-) bounded by the scaled information theoretic complexity of the underlying input driving source and the scaling factors are determined purely by the network parameters – one can ask a more practical question about tightness of the derived bounds¹¹. In general, wider differences between contraction ratios of the IFS maps will result in weaker bounds.

¹¹thanks to an anonymous reviewer for raising this question

We ran an experiment with randomly initialized RNNs having $N = 5$ recurrent neurons and standard logistic sigmoid activation functions $g(u) = 1/(1 + \exp(-u))$. The weights $\mathbf{W} = (w_{nj})$ were sampled from the uniform distribution over a symmetric interval $[-w_{max}, w_{max}]$, with $w_{max} \in [0.05, 0.35]$. We assumed that the networks were input-driven by sequences from a fair Bernoulli source over $A = 4$ symbols. In other words, the underlying automaton has a single state with A symbol loops. The topological entropy of such a source is $\log A$. For each weight range $w_{max} \in \{0.05, 0.06, 0.07, \dots, 0.35\}$ we generated 1000 networks and approximated the upper and lower bounds on fractal dimensions presented in Corollary 6.3. In particular two types of approximations were made: **(1)** we assumed that the principal semi-axes of $\mathbf{W}B_0$ (B_0 is an N -dimensional ball centered at the origin) are aligned with the neuron-wise coordinate system in \mathbb{R}^N and **(2)** we did not check whether the IFSs fall into the disjoint case. Both simplifications, while potentially over-estimating the bounds, considerably speed-up the computations. The results are shown in Figure 1(a). The means of the bounds across 1000 networks are plotted as plain solid lines; the dashed lines indicate the spread of one standard deviation.

To verify the theoretical bounds, we generated from the Bernoulli source a long sequence ω of 10000 symbols. Next we constructed for each weight range $w_{max} \in \{0.1, 0.2, 0.3\}$ 10 randomly initialized RNNs and drove each network with ω . We collected the 10000 recurrent activations from every RNN and estimated their box-counting fractal dimension using the FD3 system (version 4) (Sarraille & Myers, 1994). The mean dimension estimates are shown in Figure 1(a) as squares connected by a solid line; the error bars correspond to the spread of \pm one standard deviation. For comparison, we

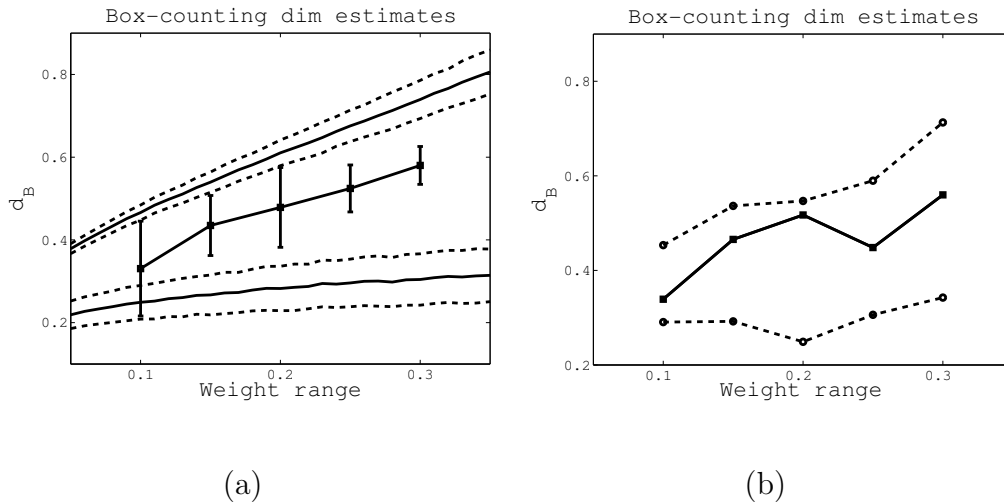


Figure 1: (a) Mean lower and upper bounds on fractal dimensions of RNNs (plain solid lines) estimated for weight ranges in $[0.05, 0.35]$ across 1000 RNN realizations. The dashed lines indicate the spread of one standard deviation. Solid line with squares shows the mean empirical dimension estimates from 10000 recurrent activations obtained by driving each network with an external input sequence. (b) Theoretical bounds (circles) and empirical dimension estimates (squares) for a single RNN randomly initialized with weight range $w_{max} \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.

show in figure 1(b) the theoretical bounds (circles) and empirical dimension estimates (squares) for a single RNN randomly initialized with weight range $w_{max} \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.

The empirical box-counting estimates based on recurrent activation patterns inside RNNs fall between the theoretical bounds computed based on networks' parameters. Tightness of the computed bounds depends on the range of contraction ratios of the individual IFS maps within a RNN and on the degree to which the simplifying assumption of axes-aligned principal semi-axes of \mathbf{WB}_0 holds. It turns that the danger of overestimating the theoretical

lower bounds because of not checking the disjoint case condition never materialized. The lower bounds are based on the minimal contraction rate across the IFS maps in a RNN and so the potential overestimation is balanced by higher contraction rates in the rest of the IFS maps. Also, even though the disjoint case condition may not be fully satisfied, deterioration of the bounds follows a “graceful degradation” pattern as the level of overlap between the IFS images $f_a(\Omega^N)$ increases.

The framework presented in this paper is general and can be readily applied to any RNN with dynamics (46) and injective non-constant differentiable activation functions of bounded derivative, mapping \mathbb{R} into a bounded interval. Many extensions and/or refinements are possible. For example, the theory can be easily extended to second-order RNNs and can be made more specific for the case of unary (one-of- A) encodings \mathbf{c}_a of input symbols $a \in \mathcal{A}$.

9 Conclusion

It has been reported that recurrent neural networks (RNNs) form reasonably looking activation clusters even *prior to any training* (Christiansen & Chater, 1999; Kolen, 1994; Kolen, 1994a; Servan-Schreiber, Cleeremans & McClelland, 1989). We have recently shown that RNNs randomly initialized with small weights are inherently biased towards Markov models, i.e. even without any training, RNN dynamics can be readily used to extract finite memory machines (Hammer & Tiño, 2002; Tiño, Čerňanský & Beňušková, 2002; Tiño, Čerňanský & Beňušková, 2002a).

In this paper we have shown that in such cases a rigorous analysis of fractal encodings in the RNN state space can be performed. We have derived (lower

and upper) bounds on the Hausdorff and box-counting dimensions of activation patterns occurring in the RNN state space when the network is driven by an underlying finite-state automaton, as has been the case in many studies reported in the literature (Casey, 1996; Cleeremans, Servan-Schreiber & McClelland, 1989; Elman, 1990; Forcada & Carrasco, 1995; Frasconi et al., 1996; Giles et al., 1992; Manolios & Fanelli, 1994; Tiño & Šajda, 1995; Watrous & Kuhn, 1992). It turns out that the recurrent activations form fractal clusters the dimension of which can be bounded by the scaled entropy of the underlying driving source. The scaling factors are fixed and are given by the RNN parameters.

Acknowledgments

This work was supported by the VEGA MS SR and SAV 1/9046/02.

References

- Barnsley, M.F. (1988). *Fractals everywhere*. New York: Academic Press.
- Barnsley, M.F., Elton, J.H., & Hardin, D.P. (1989). Recurrent iterated function system. *Constructive Approximation*, 5, 3–31.
- Blair, A.D., & Pollack, J.B. (1997). Analysis of Dynamical Recognizers. *Neural Computation*, 9(5), 1127–1142.
- Bodén, M., & Blair, A.D. (2002). Learning the dynamics of embedded clauses. *Applied Intelligence*, to appear.

- Bodén, M., & Wiles, J. (2002). On learning context free and context sensitive languages. *IEEE Transactions on Neural Networks*, 13(2), 491–493.
- Casey, M.P. (1996). The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction. *Neural Computation*, 8(6), 1135–1178.
- Christiansen, M.H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23, 417–437.
- Cleeremans, A., Servan-Schreiber, D., & McClelland, J.L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, 1(3), 372–381.
- Culik(II), K., & Dube, S. (1993). Affine automata and related techniques for generation of complex images. *Theoretical Computer Science*, 116(2), 373–398.
- Doya, K. (1992). Bifurcations in the learning of recurrent neural networks. In *Proc. of 1992 IEEE Int. Symposium on Circuits and Systems*, (pp. 2777–2780).
- Edgar, G.A., & Golds, J. (1999). A fractal dimension estimate for a graph-directed iterated function system of non-similarities. *Indiana University Mathematics Journal*, 48, 429–447.
- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.

- Falconer, K.J. (1990). *Fractal Geometry: Mathematical Foundations and Applications*. New York: John Wiley and Sons.
- Fernau, H., & Staiger, L. (1994). Valuations and unambiguity of languages, with applications to fractal geometry. In S. Abiteboul, & E. Shamir (Eds.), *Automata, Languages and Programming, 21st International Colloquium, ICALP 94* (pp. 11–22).
- Fernau, H., & Staiger, L. (2001). Iterated function systems and control languages. *Information and Computation*, 168(2), 125–143.
- Forcada, M.L., & Carrasco, R.C. (1995). Learning the initial state of a second-order recurrent neural network during regular-language inference. *Neural Computation*, 7(5), 923–930.
- Frasconi, P., Gori, M., & Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, 9(5), 768–786.
- Frasconi, P., Gori, M., Maggini, M., & Soda, G. (1996). Insertion of finite state automata in recurrent radial basis function networks. *Machine Learning*, 23, 5–32.
- Giles, C.L., Miller, C.B., Chen, D., Chen, H.H., Sun, G.Z., & Lee, Y.C. (1992). Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation*, 4(3), 393–405.
- Giles, C.L., & Omlin, C.W. (1993). Insertion and refinement of production rules in recurrent neural networks. *Connection Science*, 5(3).

- Hammer, B. (2002). Recurrent networks for structured data - a unifying approach and its properties. *Cognitive Systems Research*, to appear.
- Hammer, B., & Tiño, P. (2002). *Neural networks with small weights implement finite memory machines* (Tech. Rep. P-241). Osnabrück, Germany: Dept. of Math./Comp.Science, University of Osnabrück.
- Kolen, J.F. (1994). The origin of clusters in recurrent neural network state space. In *Proceedings from the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 508–513). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kolen, J.F. (1994a). Recurrent networks: state machines or iterated function systems? In M.C. Mozer, P. Smolensky, D.S. Touretzky, J.L. Elman, & A.S. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School* (pp. 203–210). Hillsdale, NJ: Erlbaum Associates.
- Kremer, S.C. (2001). Spatio-temporal connectionist networks: A taxonomy and review. *Neural Computation*, 13(2), 249–306.
- Manolios, P., & Fanelli, R. (1994). First order recurrent neural networks and deterministic finite state automata. *Neural Computation*, 6(6), 1155–1173.
- Melnik, O., & Pollack, J.B. (1998). A gradient descent method for a neural fractal memory. In *Proceedings of WCCI 98. International Joint Conference on Neural Networks*. IEEE Press.
- Micheli, A., Sperduti, A., Starita, A., & Bianucci, A.M. (2002). QSPR/QSAR Based on Neural Networks for Structures, In L.M. Sztandera, H. Cartwright

(eds.), *Soft Computing Approaches in Chemistry*. Heidelberg: Physica Verlag, to appear.

Minc, H. (1988). *Nonnegative Matrices*. Wiley.

Prusinkiewicz, P., & Hammel, M. (1992). Escape-time visualization method for language-restricted iterated function system. In *Proceedings of Graphics Interface '92* (pp. 213–223).

Rodriguez, P., Wiles, J., & Elman, J.L. (1991). A recurrent neural network that learns to count. *Connection Science*, 11, 5–40.

Sarraille, J.J., & Myers, L.S. (1994). FD3: A Program for Measuring Fractal Dimensions. *Educational and Psychological Measurement*, 54(1), 94–97.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. (1989). Encoding sequential structure in simple recurrent networks. In D.S. Touretzky (Eds.), *Advances in Neural Information Processing Systems 1*. San Mateo, CA: Morgan Kaufmann.

Stucki, D.J., & Pollack, J.B. (1992). Fractal (reconstructive analogue) memory. In *Proceedings of 14th Annual Cognitive Science Conference* (pp. 118–123).

Tiño, P., Horne, B.G., & Giles, C.L. (2001). Attractive periodic sets in discrete time recurrent networks (with emphasis on fixed point stability and bifurcations in two–neuron networks). *Neural Computation*, 13(6), 1379–1414.

Tiño, P., Horne, B.G., Giles, C.L., & Collingwood, P.C. (1998). Finite state machines and recurrent neural networks – automata and dynamical

- systems approaches. In J.E. Dayhoff, & O. Omidvar (Eds.), *Neural Networks and Pattern Recognition* (pp. 171–220). Academic Press.
- Tiňo, P., Čerňanský, M., & Beňušková, L. (2002). Markovian architectural bias of recurrent neural networks. In P. Sinčák, J. Vaščák, V. Kvasnička, & J. Pospichal (Eds.), *Intelligent Technologies - Theory and Applications* (pp. 17–23). Amsterdam: IOS Press.
- Tiňo, P., Čerňanský, M., & Beňušková, L. (2002a). *Markovian architectural bias of recurrent neural networks*. (Tech. Rep. NCRG/2002/008). Birmingham, UK: NCRG, Aston University.
- Tiňo, P., & Šajda, J. (1995). Learning and extracting initial mealy machines with a modular neural network model. *Neural Computation*, 7(4), 822–844.
- Thomas, W. (1990). Automata on infinite objects. In J.van Leeuwen (Eds.), *Handbook of Theoretical Computer Science*, volume B (pp. 133–191). Amsterdam: Elsevier.
- Watrous, R.L., & Kuhn, G.M. (1992). Induction of finite-state languages using second-order recurrent networks. *Neural Computation*, 4(3), 406–414.