Architectural Bias in Recurrent Neural Networks - Fractal Analysis

Peter Tiňo[1] and Barbara Hammer[2]

[1] Neural Computing Research Group, Aston University, Birmingham B4 7ET, UK
[2] University of Osnabrück, D-49069 Osnabrück, Germany
tinop@aston.ac.uk, hammer@informatik.uni-osnabrueck.de

Abstract. We have recently shown that when initiated 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 [6, 8]. Following [2], 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 obtain both lower and upper bounds on various types of fractal dimensions, such as box-counting and Hausdorff dimensions.

1 Introduction

There is a considerable amount of literature devoted to connectionist processing of sequential symbolic structures. For example, researchers have been interested in formulating models of human performance in processing linguistic patterns of various complexity (e.g. [2]).

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* [2, 4, 5, 7]. Following [2], we refer to this phenomenon as the *architectural bias of RNNs.*

We have recently shown, both empirically and theoretically, the meaning of the architectural bias: when initiated 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 [6, 8]. In this study we further extend our work by rigorously analyzing the "size" of recurrent activation patterns in such RNNs. Since the activations are of fractal nature, the size is expressed through fractal dimensions.

2 Recurrent networks as IFS

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

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

where $x_j(t)$ and $i_j(t)$ are elements, at time t, of the state and input vectors, $\mathbf{x}(t) = (x_1(t), ..., x_N(t))^T$ and $\mathbf{i}(t) = (i_1(t), ..., i_D(t))^T$, respectively, w_{nj} and v_{nj} are recurrent and input connection weights, respectively, d_n's constitute the bias vector $\mathbf{d} = (d_1, ..., d_N)^T$, and $g(\cdot)$ is an injective non-constant differentiable activation function of bounded derivative from $\Re$ to a bounded interval Ω 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 (1) in matrix form

$$\mathbf{x}(t) = G(\mathbf{V}\mathbf{i}(t) + \mathbf{W}\mathbf{x}(t-1) + \mathbf{d}), \tag{2}$$

where $G : \Re^N \to \Omega^N$ is the element-wise application of g.

Assume RNN is processing strings over a finite alphabet $\mathcal{A} = \{1, 2, ..., A\}$ of A symbols. Symbols $a \in \mathcal{A}$ are presented at the network input as unique D-dimensional codes $\mathbf{c}_a \in \Re^D$, one for each symbol $a \in \mathcal{A}$. RNN can be viewed as a non-linear iterative function system (IFS) [1], i.e. as a collection of A maps

$$\mathbf{x}(t) = F_a(\mathbf{x}(t-1)) = (G \circ T_a)(\mathbf{x}(t-1)) = G(\mathbf{W}\mathbf{x}(t-1) + \mathbf{d}_a), \quad a \in \mathcal{A}, \tag{3}$$

acting on Ω^N, where $\mathbf{d}_a = \mathbf{V}\mathbf{c}_a + \mathbf{d}$.

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

$$[B]_i = \{x_i |\ \mathbf{x} = (x_1, ..., x_N)^T \in B\},\ i = 1, 2, ..., N. \tag{4}$$

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

$$C_a = \bigcup_{1 \le i \le N} [T_a(\Omega^N)]_i. \tag{5}$$

Recall that singular values $\alpha_1, ..., \alpha_N$ 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 semiaxes of the image of the unit ball under the linear map defined by the matrix $\mathbf{W}$. We assume $\mathbf{W}$ is non-singular and adopt the convention that $\alpha_1 \ge \alpha_2 \ge ... \ge \alpha_N > 0$. Denote by $\alpha_+(\mathbf{W})$ and $\alpha_-(\mathbf{W})$ the largest and the smallest singular values of $\mathbf{W}$, respectively, i.e. $\alpha_+(\mathbf{W}) = \alpha_1$ and $\alpha_+(\mathbf{W}) = \alpha_N$. An easy lemma gives us conditions under which the maps F_a are contractive.

Lemma 1: *If*

$$k_a^{max} = \alpha_+(\mathbf{W}) \cdot \sup_{v \in C_a} |g'(v)| < 1, \tag{6}$$

then the map F_a, $a \in \mathcal{A}$, is contractive.
Proof: The result follows from two facts:

(1) a map f from a metric space[1] $(U, \|.\|_U)$ to a metric space $(V, \|.\|_V)$ is contractive, if it is Lipschitz continuous, i.e. for all $\mathbf{x}, \mathbf{y} \in U$, $\|f(\mathbf{x}) - f(\mathbf{y})\|_V \leq k_f \|\mathbf{x} - \mathbf{y}\|_U$, and the Lipschitz constant k_f is smaller than 1.
(2) The Lipschitz constant k of a composition $(f_1 \circ f_2)$ of two Lipschitz continuous maps f_1 and f_2 with Lipschitz constants k_1 and k_2 is equal to $k = k_1 \cdot k_2$.

Note that $\alpha_+(\mathbf{W})$ is a Lipschitz constant of the affine maps T_a and that $\sup_{v \in C_a} |g'(v)|$ 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 1, we get

Lemma 2: *For*

$$k_a^{min} = \alpha_-(\mathbf{W}) \cdot \inf_{v \in C_a} |g'(v)|, \tag{7}$$

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

3 Fractal dimensions

Let $K \subseteq \Omega^N$. 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 the RNNs state space Ω^N is bounded, $N_\delta(K)$ is finite for each $\delta > 0$. The rate of increase of $N_\delta(K)$ as $\delta \to 0$ tells us something about the "size" of K.

Definition 1: The *upper* and *lower box-counting dimensions* of K are defined as

$$\dim_B^+ K = \limsup_{\delta \to 0} \frac{\log N_\delta(K)}{-\log \delta} \quad \text{and} \quad \dim_B^- K = \liminf_{\delta \to 0} \frac{\log N_\delta(K)}{-\log \delta}, \tag{8}$$

respectively.

Definition 2: Let $s > 0$. For $\delta > 0$, define

$$\mathcal{H}_\delta^s(K) = \inf_{\Gamma_\delta(K)} \sum_{B \in \Gamma_\delta(K)} (diam\ B)^s, \tag{9}$$

where the infimum is taken over the set $\Gamma_\delta(K)$ of all countable δ-fine covers of K. Define $\mathcal{H}^s(K) = \lim_{\delta \to 0} \mathcal{H}_\delta^s(K)$. The *Hausdorff dimension of the set* K is

$$\dim_H K = \inf\{s|\ \mathcal{H}^s(K) = 0\}. \tag{10}$$

To find out more about fractal dimensions, we refer the interested reader to [3]. It is well known that

$$\dim_H K \leq \dim_B^- K \leq \dim_B^+ K. \tag{11}$$

[1] the metric is expressed through a norm

4 Fractal dimension estimates for RNNs driven by Bernoulli source

Consider a Bernoulli source $\mathcal{S}$ over the alphabet $\mathcal{A}$. Without loss of generality assume that all symbol probabilities are nonzero. When we drive a RNN with symbolic sequences generated by $\mathcal{S}$, by action of the IFS $\{F_a\}_{a\in\mathcal{A}}$, eq. (3), we get a set of recurrent activations $\mathbf{x}(t)$ that tend to group in well-separated clusters [4, 5, 7]. In case of contractive RNNs (that are capable of emulating finite-memory machines, see [6, 8]), the activations $\mathbf{x}(t)$ approximate the unique attractor $K \subset \Omega^N$ of the IFS $\{F_a\}_{a\in\mathcal{A}}$ via the so-called chaos game algorithm [1]. In this situation, we can get size estimates for the activation clusters.

Define

$$\ell = \sup_{v\in\bigcup_a C_a} |g'(v)| = \max_{a\in\mathcal{A}} \sup_{v\in C_a} |g'(v)|, \qquad q = \min_{a,b\in\mathcal{A},a\neq b} \|\mathbf{V}(\mathbf{c}_a - \mathbf{c}_b)\|. \tag{12}$$

Theorem 1:
(i) *Suppose* $\alpha_+(\mathbf{W}) < \ell^{-1}$. *Let* s_{max} *be the (unique) solution of (see (6))*

$$\sum_{a\in\mathcal{A}} (k_a^{max})^s = 1. \tag{13}$$

Then, $\dim_B^+ K \leq s_{max}$.
(ii) *Suppose*

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

and let s_{min} *be the (unique) solution of (see (7))*

$$\sum_{a\in\mathcal{A}} (k_a^{min})^s = 1. \tag{14}$$

Then, $\dim_H K \geq s_{min}$.
Proof: Since $\alpha_+(\mathbf{W}) < \ell^{-1}$, by (12) and lemma 1, the IFS $\{F_a\}_{a\in\mathcal{A}}$ is contractive. Furthermore, the contractions F_a have Lipschitz constants k_a^{max}. Statement **(i)** follows from [3], proposition 9.6.

For the lower bound **(ii)**, we need the open set condition (OSC)[2]:
(j) for all $a \in \mathcal{A}$, $F_a(\Omega^N) \subseteq \Omega^N$ and
(jj) $F_a(\Omega^N) \cap F_b(\Omega^N) = \emptyset$, for all $a, b \in \mathcal{A}$, $a \neq b$.

(j) is automatically satisfied, since all F_a's are contractions. To see that **(jj)** holds, note that $\alpha_+(\mathbf{W}) < \frac{q}{\sqrt{N}\ |\Omega|}$, which implies[3]

$$q = \min_{a\neq b} \|\mathbf{V}(\mathbf{c}_a - \mathbf{c}_b)\| = \min_{a\neq b} \|\mathbf{d}_a - \mathbf{d}_b\| > 2\ \alpha_+(\mathbf{W})\ \frac{diam\ \Omega^N}{2},$$

[2] It is not necessary to apply the OSC to the whole state space, as done here. An open subset of Ω^N, such that **(j)** and **(jj)** hold, would suffice.
[3] radius of the sphere encircling Ω^N is $(diam\ \Omega^N)/2 = \sqrt{N}\ |\Omega|/2$

and so $T_a(\Omega^N) \cap T_b(\Omega^N) = \emptyset$, for all $a, b \in \mathcal{A}$, $a \neq b$. But since G is injective, $(G \circ T_a)(\Omega^N) \cap (G \circ T_b)(\Omega^N) = \emptyset$, which, by (3), is equivalent to **(jj)**. The statement **(ii)** then follows from [3], proposition 9.7. □

Using (11), we get

Corollary 1: *Under the assumptions of Theorem 1* **(ii)***,* $s_{min} \leq \dim K \leq s_{max}$*, where* s_{max} *and* s_{min} *are determined by (13) and (14), respectively, and '*$\dim$*' is any of the fractal dimensions*[4] $\dim_H$*,* $\dim_B^-$ *and* $\dim_B^+$*.*

4.1 Closed-form, less tight bounds

Theorem 2: *Let*

$$k^{max} = \max_{a\in\mathcal{A}} k_a^{max}, \quad k^{min} = \min_{a\in\mathcal{A}} k_a^{min}, \quad g'_{max} = \sup_{v\in\Re} |g'(v)|, \quad w_{max} = \max_{1\leq i,j\leq N} |w_{ij}|.$$

(i) *Suppose* $\alpha_+(\mathbf{W}) < \ell^{-1}$*. Then*

$$\dim_B^+ K \leq s_{max} \leq -\frac{\log A}{\log k^{max}} \leq -\frac{\log A}{\log N + \log w_{max} + \log g'_{max}}.$$

(ii) *Suppose*

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

Then

$$\dim_H K \geq s_{min} \geq -\frac{\log A}{\log k^{min}}.$$

Proof: Fractal dimension s_+ corresponding to the IFS with A *similarities* of contraction ratio k^{max} is the unique solution of

$$\sum_{a\in\mathcal{A}} (k^{max})^s = 1, \tag{15}$$

and so is given by $s_+ = -\frac{\log A}{\log k^{max}}$. Subtracting (13) from (15) and denoting $\Delta = s_{max} - s_+$, we obtain

$$\sum_{a\in\mathcal{A}} (k^{max})^{s_+} - (k_a^{max})^{s_+}(k_a^{max})^{\Delta} = 0.$$

Since for all $a \in \mathcal{A}$, $k^{max} \geq k_a^{max}$, it must be that $(k_a^{max})^{\Delta} \geq 1$. Because the IFS $\{F_a\}_{a\in\mathcal{A}}$ is contractive, Δ must be ≤ 0, and so $s_{max} \leq s_+$. **(i)** follows from Theorem 1(i), and from realizing that $N \cdot w_{max} \cdot g'_{max} \geq k^{max}$.

To prove **(ii)**, note that the fractal dimension of IFS with A *similarities* of contractive ratio k^{min} is $s_- = -\frac{\log A}{\log k^{min}}$. Similarly to the case **(i)**, we write

$$\sum_{a\in\mathcal{A}} (k^{min})^{s_-} - (k_a^{min})^{s_-}(k_a^{min})^{\Delta} = 0, \quad \Delta = s_{min} - s_-.$$

Since $k^{min} \leq k_a^{min}$, and $k_a^{min} < 1$, for all $a \in \mathcal{A}$, it must be that $(k_a^{min})^{\Delta} \leq 1$, and so $\Delta \geq 0$, which means $s_{min} \geq s_-$. □

[4] and in fact many more, e.g. packing dimension, see [3]

5 Discussion and conclusion

Recently, we have extended the work of Kolen and others (e.g. [2, 4, 5, 7]) by pointing out that in dynamical tasks of symbolic nature, when initiated with "small" weights, recurrent neural networks (RNN), with e.g. sigmoid activation functions, form contractive IFS. Such networks are inherently biased towards Markov models, i.e. even prior to any training, RNN dynamics can be readily used to extract finite memory machines [6, 8]. In other words, even without any 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.

The theory presented in this paper is general, but can be readily applied to any RNN with dynamics (1) and injective non-constant differentiable activation functions of bounded derivative, mapping $\Re$ into a bounded interval. For example, when the standard logistic sigmoid $g(u) = 1/(1+\exp(-u))$ is used, initiating recurrent weights of N recurrent neurons from the interval $(-r/N, r/N)$, $0 < r < 4$, will insure contractive symbol maps F_a. By theorem 2 (i), the upper box-counting dimension of recurrent activations is upper-bounded by $\log A/\log(4/r)$, where A is the number of symbols in the input alphabet $\mathcal{A}$.

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 $a \in \mathcal{A}$.

References

1. M.F. Barnsley: Fractals everywhere. Academic Press, New York (1988)
2. M.H. Christiansen, N. Chater: Toward a connectionist model of recursion in human linguistic performance. Cognitive Science, **23** (1999) 157–205
3. K. Falconer: Fractal Geometry- Mathematical Foundations and Applications. John Wiley & Sons, Chichester, UK (1990)
4. J.F. Kolen: Recurrent networks: state machines or iterated function systems? In: Proceedings of the 1993 Connectionist Models Summer School. Lawrence Erlbaum Associates, Hillsdale, NJ (1994) 203–210
5. J.F. Kolen: The origin of clusters in recurrent neural network state space. In: Proceedings from the Sixteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Hillsdale, NJ (1994) 508–513
6. B. Hammer, P. Tiňo: Neural networks with small weights implement finite memory machines. Technical Report P-241, Dept. of Math./Comp.Science, University of Osnabrück, Germany (2002)
7. P. Manolios and R. Fanelli: First order recurrent neural networks and deterministic finite state automata. Neural Computation **6** (1994) 1155–1173
8. P. Tiňo, M. Čerňanský, L. Beňušková: Markovian Architectural Bias of Recurrent Neural Networks. In: 2nd Euro-International Symposium on Computational Intelligence. Springer-Verlag Studies in CI (2002) to appear