Neural Networks 46 (2013) 116–123

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

The learning problem of multi-layer neural networks

Jung-Chao Ban [a], Chih-Hung Chang [b,*]

[a] Department of Applied Mathematics, National Dong Hwa University, Hualien 970003, Taiwan, ROC
[b] Department of Applied Mathematics, Feng Chia University, Taichung 40724, Taiwan, ROC

ARTICLE INFO

Article history:
Received 26 December 2012
Received in revised form 6 March 2013
Accepted 7 May 2013

Keywords:
Multi-layer neural networks
Topological entropy
Sofic shift
Learning problem
Linear separation

ABSTRACT

This manuscript considers the learning problem of multi-layer neural networks (MNNs) with an activation function which comes from cellular neural networks. A systematic investigation of the partition of the parameter space is provided. Furthermore, the recursive formula of the transition matrix of an MNN is obtained. By implementing the well-developed tools in the symbolic dynamical systems, the topological entropy of an MNN can be computed explicitly. A novel phenomenon, the asymmetry of a topological diagram that was seen in Ban, Chang, Lin, and Lin (2009) [J. Differential Equations 246, pp. 552–580, 2009], is revealed.

1. Introduction

In the past few decades, multi-layer neural networks (MNNs, Hornik, Stinchcombe, & White, 1989, Widrow & Lehr, 1990) have received considerable attention and have been successfully applied to many areas such as combinatorial optimization (Hopfield & Tank, 1985; Peterson & Söderberg, 1989), signal processing, pattern recognition (Alsultanny & Aqul, 2003; Widrow, 1962) and artificial intelligence (AI) (Bengio, 2009).

One important reason for coupling NNs is the simulation of the visual systems of mammals (Fukushima, 2013a, 2013b, each layer symbolizes a single cortex in the visual system) and it is proved that the mammal brain is organized in deep architectures[1] (Serre et al., 2007), i.e., the number of layers in MNNs is large. Due to the architecture depth of the mammal brain, scientists have been interested in learning and training deep architectures (Bengio & LeCun, 2007; Utgoff & Stracuzzi, 2002) since 2002. Hinton et al. obtained great successful results on deep architectures in 2006 by using Deep Belief Networks (DBNs) (Hinton, Osindero, & Teh, 2006) and Restricted Boltzmann Machine (RBM) (Freund & Haussler, 1994) methods, and have applied to many fields since then, e.g., classification tasks, regression, dimensionality reduction, modeling textures, modeling motion, natural language processing, object segmentation, information retrieval, robotics and collaborative filtering. A general reference is referred to Bengio (2009), and reader could find the complete bibliography therein. Related topics are that how to distinguish the different hidden layers and how two hidden layers make a difference (Kurková & Sanguineti, 2013). In Ban, Chang, and Lin (2012), Ban et al. established the mathematical foundation on the structures of hidden layers. More precisely, some checkable conditions are provided to ensure whether two hidden layers are conjugate, shift equivalent or finite shift equivalent (Lind & Marcus, 1995). This gives a connection between deep and shallow architectures.

Due to the learning algorithm and training processing, the investigation of mosaic solutions is most essential in MNN models, and such models indeed produce abundant output patterns and make the learning algorithm more efficient. In neural networks, many types of activation function, e.g., linear, McCulloch–Pitts, signum, Sigmoid, Ramp functions, are chosen for many specific purposes. The activation function indicates that under what conditions the synapses will be activated. Different activation functions make different output function spaces and produce different dynamical systems. In this paper, we consider a different activation function which comes from cellular neural networks (CNNs). Namely,

$$f(x) = \frac{1}{2}(|x+1| - |x-1|).$$

CNNs, introduced by Chua and Yang in 1988 (Chua & Yang, 1988), have many applications in the area of image processing (Chua, 1998). The topics of pattern formation and spatial chaos for mosaic

* Corresponding author.
E-mail addresses: jcban@mail.ndhu.edu.tw (J.-C. Ban), chihhung@mail.fcu.edu.tw (C.-H. Chang).

[1] Except for the simulation of the visual systems of mammals, deep architectures are often used to learn some complicated functions expressing high-level abstractions, e.g., language and AI-level tasks.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.05.006

solutions have been discussed in CNN (Juang & Lin, 2000) and multi-layer CNN (MCNN, Ban et al., 2009) models. However, it seems that there are a few studies on MNN models. The aim of this paper is to set up the mathematical foundation for the MNN model with the above activation function. The method we have provided herein is more general, an easy extension leading us to consider the classical McCulloch–Pitts model and signum activation function. More or less, this elucidation has provided a connection of learning algorithm between MNN and MCNN, or CNN and NN.

This paper is organized as follows. In Section 2 we consider the two-layer NN. The partition of parameter space, stable local patterns and the generation of global mosaic patterns are discussed. We also prove that the global mosaic solution space forms a sofic space in classical symbolic dynamical systems. Thus, the complexity (topological entropy) can be computed by using the knowledge of sofic space from symbolic dynamical systems. In Section 3 we consider the general case of MNN and some results parallel to a two-layer case are provided. Finally, the strange phenomena of the asymmetry of topological entropy are presented in Section 4.

2. Two-layer neural networks

A one-dimensional multi-layer neural network (MNN) is realized as

$$\begin{cases} \dfrac{d}{dt}x_i^{(k)}(t) = -x_i^{(k)}(t) + z^{(k)} + a^{(k)} f(x_i^{(k)}(t)) \\ \qquad + \displaystyle\sum_{\ell\in\mathcal{N}} b_\ell^{(k)} f(x_{i+\ell}^{(k-1)}(t)), \\ \dfrac{d}{dt}x_i^{(1)}(t) = -x_i^{(1)}(t) + z^{(1)} + a^{(1)} f(x_i^{(1)}(t)) \\ \qquad + \displaystyle\sum_{\ell\in\mathcal{N}} a_\ell^{(1)} f(x_{i+\ell}^{(1)}(t)), \end{cases} \tag{1}$$

for some $N \in \mathbb{N}$, $k = 2, \ldots, N$ and $i \in \mathbb{Z}$. We called the finite subset $\mathcal{N} \subset \mathbb{Z}$ the *neighborhood*, and the piecewise linear map $f(x) = \frac{1}{2}(|x+1| - |x-1|)$ is called the *output function*. The *template* $\mathbb{T} = [\mathbf{A}, \mathbf{B}, \mathbf{z}]$ is composed of a *feedback template* $\mathbf{A} = (A_1, A_2)$ with $A_1 = (a^{(1)}, \ldots, a^{(N)})$, $A_2 = (a_\ell^{(1)})_{\ell\in\mathcal{N}}$, a *controlling template* $\mathbf{B} = (B_2, \ldots, B_N)$, and the *threshold* $\mathbf{z} = (z^{(1)}, \ldots, z^{(N)})$, where $B_k = (b_\ell^{(k)})_{\ell\in\mathcal{N}}$ for $k \geq 2$. A stationary solution $\mathbf{x} = (x_i^{(1)}, \ldots, x_i^{(N)})_{i\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}_\infty \times N}$ of (1) is called *mosaic* if $|x_i^{(k)}| > 1$ for $1 \leq k \leq N$, $i \in \mathbb{Z}$. The output $\mathbf{y} = (y_i^{(1)} \ldots y_i^{(N)})_{i\in\mathbb{Z}} \in \{-1, 1\}^{\mathbb{Z}_\infty \times N}$ of a mosaic solution is called a *mosaic pattern*, where $y_i^{(k)} = f(x_i^{(k)})$. The *solution space* $\mathbf{Y}$ of (1) stores the mosaic patterns $\mathbf{y}$, and the *output space* $\mathbf{Y}^{(N)}$ of (1) is the collection of the output patterns in $\mathbf{Y}$, more precisely,

$$\mathbf{Y}^{(N)} = \{(y_i^{(N)})_{i\in\mathbb{Z}} : (y_i^{(1)} \ldots y_i^{(N)})_{i\in\mathbb{Z}} \in \mathbf{Y}\}.$$

A neighborhood $\mathcal{N}$ is called the nearest neighborhood if $\mathcal{N} = \{-1, 1\}$. The investigation of the output space of (1) is essential for elucidating the complexity of MNNs. The framework is clarified by presenting our methodology of the two-layer neural networks with the nearest neighborhood. The formalized general results are postponed to the next section (see Fig. 1).

2.1. Partition of parameters

To investigate the complexity of the behavior of (1), the prescription of parameters is essential. Generally there is an infinite choice of templates. Since, for MNNs, the neighborhood $\mathcal{N}$ is finite and the template is invariant for each i, the solution space is determined by the so-called *basic set of admissible local patterns*. This subsection demonstrates that the parameter space can be divided into finitely equivalent regions so that the two templates

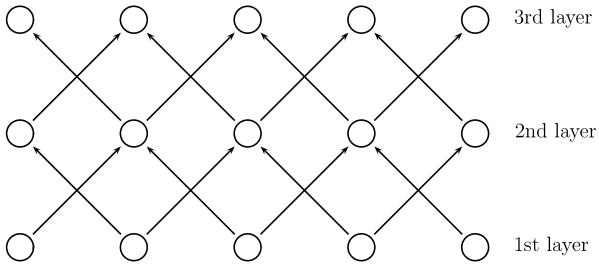


Fig. 1. Three-layer neural networks with nearest neighborhood.

$\mathbb{T}_1, \mathbb{T}_2$ of (1) assert the same solution space if and only if $\mathbb{T}_1$ and $\mathbb{T}_2$ belong to the same region. In other words, there are only finitely different behavior being observed if the neighborhood $\mathcal{N}$ is given.

The basic set of admissible local patterns of the first layer is a subset of

$$\{---, --+, -+-, -++, +--, +-+, ++-, +++\};$$

the basic set of admissible local patterns of the second layer is a subset of the ordered set $\{p_1, \ldots, p_8\}$, where $p_1, \ldots, p_8$ are

$$\begin{matrix} - \\ - \end{matrix}, \begin{matrix} - \\ + \end{matrix}, \begin{matrix} - \\ - \end{matrix}, \begin{matrix} - \\ + \end{matrix}, \begin{matrix} + \\ - \end{matrix}, \begin{matrix} + \\ + \end{matrix}, \begin{matrix} + \\ - \end{matrix}, \begin{matrix} + \\ + \end{matrix}, \tag{2}$$

respectively. (Here we refer to the patterns 1 and -1 as $+$ and $-$, respectively.) First we consider the local patterns of the second layer. For simplicity, we denote the local pattern $\begin{matrix} \alpha \\ \alpha_1 \ \alpha_2 \end{matrix}$ by $\alpha \diamond \alpha_1\alpha_2$. Suppose $\mathbf{y}$ is a mosaic pattern, for each $i \in \mathbb{Z}$; the necessary and sufficient condition for $y_i^{(2)} = 1$ is

$$a^{(2)} - 1 + z_2 > -(b_{-1}^{(2)} y_{i-1}^{(1)} + b_1^{(2)} y_{i+1}^{(1)}), \tag{3}$$

and the necessary and sufficient condition for $y_i^{(2)} = -1$ is

$$a^{(2)} - 1 - z_2 > b_{-1}^{(2)} y_{i-1}^{(1)} + b_1^{(2)} y_{i+1}^{(1)}. \tag{4}$$

It is seen that (3) and (4) divide the $a^{(2)}$–z_2 plane into 25 subregions; each subregion is encoded as $[m, n]$, where $0 \leq m, n \leq 4$, and $[m, n]$ indicates that every pair of parameters $(a^{(2)}, z^{(2)})$ in that region satisfies m and n inequalities in (3) and (4), respectively. Set

$$\begin{aligned} \mathcal{B}^{(2)}(+) &= \left\{ y^{(2)} \diamond y_1^{(1)} y_2^{(1)} : y_1^{(1)}, y_2^{(1)} \in \{-1, 1\} \right. \\ &\qquad \left. \text{satisfy (3)}, y^{(2)} = 1 \right\}, \\ \mathcal{B}^{(2)}(-) &= \left\{ y^{(2)} \diamond y_1^{(1)} y_2^{(1)} : y_1^{(1)}, y_2^{(1)} \in \{-1, 1\} \right. \\ &\qquad \left. \text{satisfy (4)}, y^{(2)} = -1 \right\}. \end{aligned}$$

The basic set of admissible local patterns of the second layer is denoted by $\mathcal{B}^{(2)} = (\mathcal{B}^{(2)}(+), \mathcal{B}^{(2)}(-))$. Similarly, for each $i \in \mathbb{Z}$, the necessary and sufficient conditions for $y_i^{(1)} = 1$ and $y_i^{(1)} = -1$ are

$$a^{(1)} - 1 + z_1 > -(a_{-1}^{(1)} y_{i-1}^{(1)} + a_1^{(1)} y_{i+1}^{(1)}), \tag{5}$$

and

$$a^{(1)} - 1 - z_1 > a_{-1}^{(1)} y_{i-1}^{(1)} + a_1^{(1)} y_{i+1}^{(1)}, \tag{6}$$

respectively. Set

$$\begin{aligned} \mathcal{B}^{(1)}(+) &= \left\{ y_1^{(1)} y^{(1)} y_2^{(1)} : y_1^{(1)}, y_2^{(1)} \in \{-1, 1\} \right. \\ &\qquad \left. \text{satisfy (5)}, y^{(1)} = 1 \right\}, \\ \mathcal{B}^{(1)}(-) &= \left\{ y_1^{(1)} y^{(1)} y_2^{(1)} : y_1^{(1)}, y_2^{(1)} \in \{-1, 1\} \right. \\ &\qquad \left. \text{satisfy (6)}, y^{(1)} = -1 \right\}. \end{aligned}$$

The basic set of admissible local patterns of the first layer is denoted by $\mathcal{B}^{(1)} = (\mathcal{B}^{(1)}(+), \mathcal{B}^{(1)}(-))$. The solution space $\mathbf{Y}$ of (1) is then described as

$$\mathbf{Y} = \{\mathbf{y} = (y_i^{(1)} y_i^{(2)})_{i\in\mathbb{Z}} : y_i^{(2)} \diamond y_{i-1}^{(1)} y_{i+1}^{(1)} \in \mathcal{B}^{(2)}, y_{i-1}^{(1)} y_i^{(1)} y_{i+1}^{(1)} \in \mathcal{B}^{(1)}\}.$$

Since the inequalities (3), (4), (5), and (6) are all linear, each component of the basic set of admissible local patterns, that is, $\mathcal{B}^{(1)}(+)$, $\mathcal{B}^{(1)}(-)$, $\mathcal{B}^{(2)}(+)$, and $\mathcal{B}^{(2)}(-)$, satisfies the so-called *linear separation property* (cf. Ban et al., 2009, Hsu, Juang, Lin, & Lin, 2000, Lay, 1992). More precisely, let $V = \{(-1,-1), (-1,1), (1,-1), (1,1)\}$ be a subset of $\mathbb{Z}^2$ and let B^* be the embedding of patterns in B without a centered pattern in $\mathbb{Z}^2$, where $B = \mathcal{B}^{(1)}(+), \mathcal{B}^{(1)}(-), \mathcal{B}^{(2)}(+), \mathcal{B}^{(2)}(-)$. For instance, if

$$B = \{\begin{smallmatrix} & - \\ - & - \end{smallmatrix}, \begin{smallmatrix} & - \\ - & + \end{smallmatrix}, \begin{smallmatrix} & - \\ + & - \end{smallmatrix}\},$$

then $B^* = \{(-1,-1), (-1,1), (1,-1)\}$. The linear separation property indicates that B is a subset of the basic set of admissible local patterns of (1) if and only if there is a line separating B^* and $V \setminus \mathcal{B}$. Fig. 2 asserts that there are only 6 different geometrical structures satisfying the separation property. Namely, there are 12 packs of admissible local patterns $B \in \{\mathcal{B}^{(1)}(+), \mathcal{B}^{(1)}(-), \mathcal{B}^{(2)}(+), \mathcal{B}^{(2)}(-)\}$.

Let $\mathcal{P}_8 = \{(a^{(2)}, b_{-1}^{(2)}, b_1^{(2)}, z^{(2)}, a^{(1)}, a_{-1}^{(1)}, a_1^{(1)}, z^{(1)})\}$ denote the parameter space of (1). Theorem 2.1 asserts that $\mathcal{P}_8$ can be partitioned into finitely many subregions so that two templates in the same partition exhibit the same basic set of admissible local patterns. Namely, $\mathbb{T}_1, \mathbb{T}_2$ are located in the same subregion if and only if $\mathbf{Y}_1 = \mathbf{Y}_2$, where $\mathbf{Y}_i$ is the solution space of (1) corresponding to the template $\mathbb{T}_i$ for $i = 1, 2$. To see this, define $\xi_1, \xi_2 : \{-1, 1\}^{\mathbb{Z}_{2\times 1}} \to \mathbb{R}$ by

$$\xi_1(w_1, w_2) = a_{-1}^{(1)} w_1 + a_1^{(1)} w_2,$$
$$\xi_2(w_1, w_2) = b_{-1}^{(2)} w_1 + b_1^{(2)} w_2.$$

Since (3) and (4) partition the $a^{(2)}$–$z^{(2)}$ plane into 25 regions, the "order" (i.e., the relative position) of lines $a^{(2)} - 1 + (-1)^\ell z_2 = (-1)^\ell \xi_2(y_{i-1}^{(1)}, y_{i+1}^{(1)})$, $\ell = 1, 2$, can be uniquely determined according to the following procedures:

(1) The signs of $b_{-1}^{(2)}, b_1^{(2)}$ (i.e., the parameters are positive or negative).
(2) The magnitude of $b_{-1}^{(2)}, b_1^{(2)}$ (i.e., $|b_{-1}^{(2)}| > |b_1^{(2)}|$ or $|b_{-1}^{(2)}| < |b_1^{(2)}|$).

This partitions the $a^{(2)}$–$b_{-1}^{(2)}$–$b_1^{(2)}$–$z^{(2)}$ plane into $8 \times 25 = 200$ subregions. (Recall that $a^{(2)} - 1 + (-1)^\ell z_2 = (-1)^\ell \xi_2(y_{i-1}^{(1)}, y_{i+1}^{(1)})$, $\ell = 1, 2, y_{i-1}^{(1)}, y_{i+1}^{(1)} \in \{-1, 1\}$.) Similarly, the $a^{(1)}$–$a_{-1}^{(1)}$–$a_1^{(1)}$–$z^{(1)}$ plane is partitioned into 200 subregions. Hence the parameter space $\mathcal{P}_8$ is partitioned into less than 40,000 equivalent subregions.

Theorem 2.1. *There is a positive integer K and a unique set of open subregions $\{P_k\}_{k=1}^K$ satisfying*

(i) $\mathcal{P}_8 = \bigcup_{k=1}^K \overline{P}_k$.
(ii) $P_i \bigcap P_j = \varnothing$ *if* $i \neq j$.
(iii) *Templates* $\mathbb{T}, \mathbb{T}' \in P_k$ *for some* k *if and only if* $\mathcal{B}(\mathbb{T}) = \mathcal{B}(\mathbb{T}')$.

2.2. Spatial complexity of two-layer neural networks

This subsection is devoted to the discussion of the behavior exhibited by MNNs. First the structure of the solution space and output space are investigated. It is shown that the solution space $\mathbf{Y}$ is a topological Markov chain while the output space $\mathbf{Y}^{(2)}$ is a so-called *sofic shift space* in symbolic dynamical systems. After elucidating the topological structure, it was followed by the study of the spatial complexity of the solution space and output space. Herein by spatial complexity we mean the topological entropy, a quantity that measures the growth rate of the number of patterns when enlarging the size of the lattice.

In order to investigate the structure of the solution space, we assign the local patterns an ordering since the solution space $\mathbf{Y}$ is determined by the basic set of admissible local patterns, and then define the *ordering matrix* to clarify the global patterns in the solution space. The ordering matrix of the two-layer NNs which is defined by $\mathbb{X}_2$ is given in Box I. It is seen that $\mathbb{X}_2(p, q)$ consists of two local patterns in $\mathcal{B}^{(2)}$, and $\mathbb{X}_2$ is self-similar; more specifically, if we write $\mathbb{X}_2 = \begin{pmatrix} \mathbb{X}_{2;11} & \mathbb{X}_{2;12} \\ \mathbb{X}_{2;21} & \mathbb{X}_{2;22} \end{pmatrix}$, where $\mathbb{X}_{2;ij}$ is a 4×4 matrix for all i, j, then the bottom patterns of $\mathbb{X}_{2;ij}(p, q)$ and $\mathbb{X}_{2;i'j'}(p, q)$ are identical for all i, i', j, j', p, q. Let

$$a_{00} = --, \quad a_{01} = -+, \quad a_{10} = +-, \quad a_{11} = ++, \tag{7}$$

define

$$a_{i_1 i_2} a_{i'_2 i_3} = \varnothing \Leftrightarrow i_2 \neq i'_2. \tag{8}$$

If $a_{i_1 i_2} a_{i'_2 i_3} \neq \varnothing$, then it is a pattern with size 3×1 and denoted by $a_{i_1 i_2 i_3}$. The ordering matrix of the first layer is defined by

$$\mathbb{X}_1 = \begin{array}{c|cccc} & -- & -+ & +- & ++ \\ \hline -- & \boxed{---} & \boxed{--+} & \varnothing & \varnothing \\ -+ & \varnothing & \varnothing & \boxed{-+-} & \boxed{-++} \\ +- & \boxed{+--} & \boxed{+-+} & \varnothing & \varnothing \\ ++ & \varnothing & \varnothing & \boxed{++-} & \boxed{+++} \end{array}.$$

Moreover, applying the matrix product to $\mathbb{X}_1$ shows that

$$\mathbb{X}_1^2 = \begin{array}{c|cccc} & -- & -+ & +- & ++ \\ \hline -- & \boxed{----} & \boxed{---+} & \boxed{--+-} & \boxed{--++} \\ -+ & \boxed{-+--} & \boxed{-+-+} & \boxed{-++-} & \boxed{-+++} \\ +- & \boxed{+---} & \boxed{+--+} & \boxed{+-+-} & \boxed{+-++} \\ ++ & \boxed{++--} & \boxed{++-+} & \boxed{+++-} & \boxed{++++} \end{array}$$

stores all patterns with size 4×1.

To investigate the complexity of MNNs, we introduce the *transition matrix* first. As the ordering matrix records the behavior of the global patterns, the transition matrix relates to the number of global patterns. Suppose $\mathcal{B}(\mathbb{T}) = (\mathcal{B}^{(1)}, \mathcal{B}^{(2)})$ is the basic set of admissible local patterns of (1) with respect to the template $\mathbb{T}$. The transition matrix $\mathbf{T}$ is defined by

$$\mathbf{T}(i, j) = \begin{cases} 1, & p_i, p_j \in \mathcal{B}^{(2)} \text{ and } \alpha_{i-1}\alpha_{j-1}\alpha_{i+1}, \\ & \alpha_{j-1}\alpha_{i+1}\alpha_{j+1} \in \mathcal{B}^{(1)}; \\ 0, & \text{otherwise;} \end{cases} \tag{9}$$

herein p_k is defined in (2) and is presented as $\alpha_k \diamond \alpha_{k-1}\alpha_{k+1}$ for $k = 1, \ldots, 8$. Furthermore, the transition matrix of the second layer $T_2 \in \mathcal{M}_{8\times 8}(\{0, 1\})$ is defined by

$$T_2(i, j) = 1 \quad \text{if and only if } p_i, p_j \in \mathcal{B}^{(2)} \tag{10}$$

while the transition matrix of the first layer $T_1 \in \mathcal{M}_{4\times 4}(\{0, 1\})$ is defined by

$$T_1(i, j) = 1 \quad \text{if and only if } \mathbb{X}_1(i, j) \in \mathcal{B}^{(1)}. \tag{11}$$

The discussion above demonstrates that T_1^2 stores all admissible patterns of length 4; denote $T_1^2 = (T_{i,j})_{i,j=1}^2$ as four smaller 2×2 matrices. Define $\overline{T}_1$ by

$$\overline{T}_1(p, q) = T_{i,j}(k, l), \quad \text{where } p = 2i + j - 2, q = 2k + l - 2. \tag{12}$$

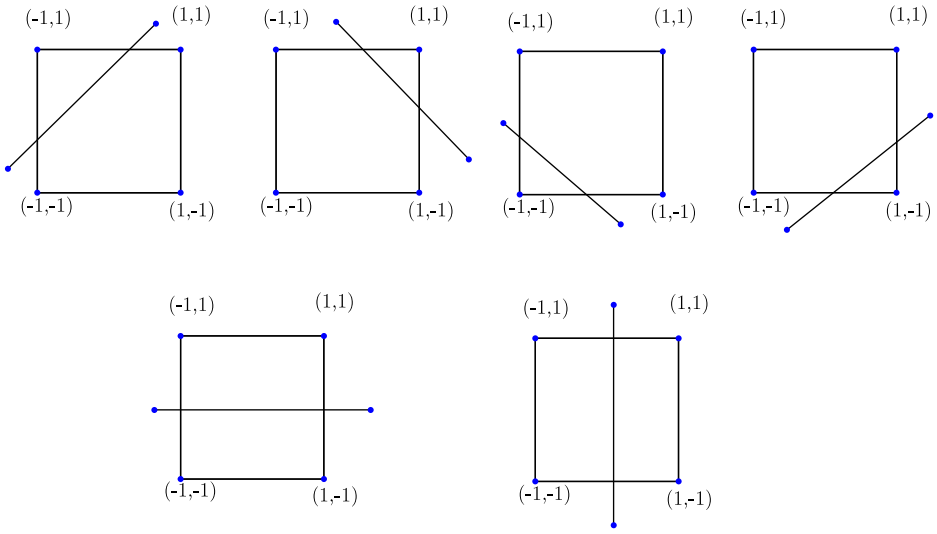
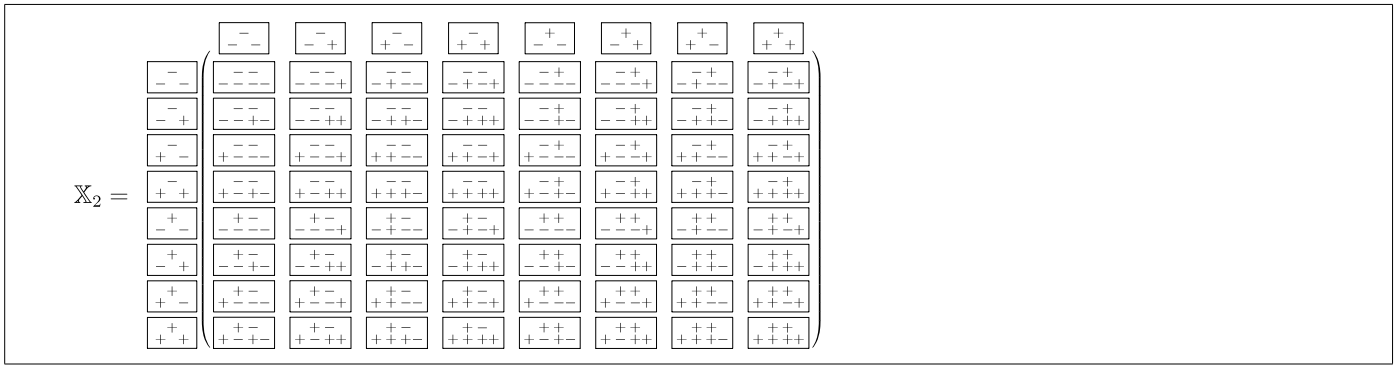


Fig. 2. The basic set of admissible local patterns is constrained from the linear separation property. In other words, there are only 6 combinations for the choice of local patterns without centered entry.



Box I.

A straightforward examination from the structure of the ordering matrices $\mathbb{X}_1$ and $\mathbb{X}_2$ asserts Theorem 2.3, which decomposes $\mathbf{T}$ into the product of T_1 and T_2. Before presenting the theorem, two kinds of products of matrices are defined as follows.

Definition 2.2. Suppose $M \in \mathcal{M}_{k_1\times k_2}(\mathbb{R})$ and $N \in \mathcal{M}_{\ell_1\times\ell_2}(\mathbb{R})$. The *Kronecker product* (or *tensor product*) $M \otimes N$ of M and N is defined by

$$M \otimes N = (M(i,j)N) \in \mathcal{M}_{k_1\ell_1\times k_2\ell_2}(\mathbb{R}). \tag{13}$$

Suppose $P, Q \in \mathcal{M}_{r\times r}(\mathbb{R})$. The *Hadamard product* (or *inner product*) $P \circ Q$ of P and Q is defined by

$$(P \circ Q)(i,j) = P(i,j)Q(i,j). \tag{14}$$

Theorem 2.3. *Suppose* $\mathbf{T}$ *is the transition matrix of* (1)*, and* T_1 *and* T_2 *are the transition matrices of* (1) *with respect to the first and second layer, respectively. Let* $\overline{T}_1$ *be defined as in* (12)*. Then*

$$\mathbf{T} = T_2 \circ (E_2 \otimes \overline{T}_1), \tag{15}$$

where E_k *is a* $k \times k$ *matrix with all entries being* 1*'s.*

As has been demonstrated in Ban et al. (2012, 2009), Juang and Lin (2000), the solution space of a multi-layer cellular neural network is a so-called *shift of finite type* (SFT, also known as a topological Markov shift) in the symbolic dynamical systems, and the output space is a *sofic shift* (sofic), which is the image of an SFT under a surjective map. We give a brief elucidation about the fact that the output space $\mathbf{Y}^{(2)}$ is a sofic shift to make the present manuscript self-contained. A detailed instruction for the symbolic dynamical systems is referred to Lind and Marcus (1995).

A *labeled graph* $\mathcal{G} = (G, \mathcal{L})$ consists of an underlying graph $G = (\mathcal{V}, \mathcal{E})$ and the *labeling* $\mathcal{L} : \mathcal{E} \to \mathcal{A}$ which assigns to each edge a label from the finite alphabet $\mathcal{A}$, where $\mathcal{V}$ and $\mathcal{E}$ refer to the sets of vertices and edges, respectively. A sofic shift $\mathbf{X}$ is defined by

$$\mathbf{X} = \{(\omega_i)_{i\in\mathbb{Z}} : \omega_i = \mathcal{L}(e_i), e_i \in \mathcal{E}, \mathrm{ter}(e_i) = \mathrm{init}(e_{i+1})\}$$

for some labeled graph $\mathcal{G}$, where $\mathrm{ter}(e)$ and $\mathrm{init}(e)$ mean the terminal and initial vertices of the edge $e \in \mathcal{E}$, respectively. Without loss of generality, we may assume that there is at most one edge connecting two vertices. The transition matrix $\mathbf{T}$ of the labeled graph $\mathcal{G}$ is indexed by the vertices $\mathcal{V}$ and $\mathbf{T}(p,q) = 1$ if and only if there is an edge from p to q. Set the alphabet $\mathcal{A} = \{a_{00}, a_{01}, a_{10}, a_{11}\}$, where a_{ij} is defined in (7).

Notably, the transition matrix $\mathbf{T}$ is not capable of recording the exact number of those paths carrying different labels in general. In other words, $\mathbf{T}$ cannot reflect the spatial complexity of the output space $\mathbf{Y}^{(2)}$ properly. The main difficulty is that a labeled path may be recorded several times in the transition matrix. To overcome this, we introduce the *symbolic transition matrix*.

Define the symbolic transition matrix as

$$\mathbf{S} = \left(\begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \otimes E_4\right) \circ \mathbf{T}, \qquad \mathbf{S}(i,j) = \varnothing \quad \text{if } \mathbf{T}(i,j) = 0. \tag{16}$$

Let $\mathcal{V} = \{p_1, \ldots, p_8\}$, and $e_{ij} \in \mathcal{E}$ if $\text{init}(e_{ij}) = p_i$, $\text{ter}(e_{ij}) = p_j$ and $\mathbf{T}(i, j) = 1$. Define $\mathcal{L} : \mathcal{E} \to \mathcal{A}$ by

$$\mathcal{L}(e_{ij}) = a_{\bar{i}\bar{j}}, \quad \text{where } \bar{k} = \left\lfloor \frac{k-1}{4} \right\rfloor,$$

where $\lfloor \cdot \rfloor$ is the Gauss function. Let $\mathcal{G} = (G, \mathcal{L})$ be the labeled graph with an underlying graph $G = (\mathcal{V}, \mathcal{E})$ and labeling $\mathcal{L}$. It is seen that the output space $\mathbf{Y}^{(2)} = \mathbf{X}$ is the sofic shift defined by $\mathcal{G}$. This demonstrates Theorem 2.4.

Theorem 2.4. *The output space $\mathbf{Y}^{(2)}$ is a sofic shift.*

One of the most frequently used quantums for the measure of the spatial complexity is the *topological entropy*, which measures the growth rate of the number of global patterns with respect to the size of lattices. Let X be a symbolic space and let $\Gamma_n(X)$ denote the number of patterns in X of length n. The topological entropy of X is defined by

$$h(X) = \lim_{n\to\infty} \frac{\log \Gamma_n(X)}{n}, \quad \text{provided the limit exists.}$$

The space X is called *pattern formation* if $h(X) = 0$, and *spatial chaos* otherwise. Similar to Ban et al. (2009), it can be verified that the topological entropy of the output space of (1) is $h(\mathbf{Y}^{(2)}) = \log \rho_{\mathbf{T}}$ if the labeled graph constructed from $\mathbf{T}$ is *right-resolving*, where $\rho_{\mathbf{T}}$ is the spectral radius of $\mathbf{T}$. Here a labeled graph $\mathcal{G} = (G, \mathcal{L})$ is called right-resolving if the restriction of $\mathcal{L}$ to $\mathcal{E}_I$ is one-to-one, where $\mathcal{E}_I$ consists of those edges starting from I.

If $\mathcal{G}$ is not right-solving, there exists a labeled graph $\mathcal{H}$, derived by applying the *subset construction method* (SCM) to $\mathcal{G}$, such that the sofic shift defined by $\mathcal{H}$ is identical to the original space. The new labeled graph $\mathcal{H} = (H, \mathcal{L}')$ is constructed as follows.

The vertices I of H are the nonempty subsets of the vertex set $\mathcal{V}$ of G. If $I \in \mathcal{V}'$ and $a \in \mathcal{A}$, let J denote the set of terminal vertices of edges in G starting at some vertices in I and labeled a, i.e., J is the set of vertices reachable from I using the edges labeled a. There are two cases.

(1) If $J = \varnothing$, do nothing.
(2) If $J \neq \varnothing$, $J \in \mathcal{V}'$ and draw an edge in H from I to J labeled a.

Carrying this out for each $I \in \mathcal{V}'$ and each $a \in \mathcal{A}$ produces the labeled graph $\mathcal{H}$. Then, each vertex I in H has at most one edge with a given label starting at I. This implies that $\mathcal{H}$ is right-resolving.

Theorem 2.5 (*Lind & Marcus, 1995*)**.** *Suppose a labeled graph $\mathcal{G} = (G, \mathcal{L})$ is not right-resolving, and $\mathcal{H} = (H, \mathcal{L}')$ is a right-resolving labeled graph constructed via the SCM. Then the sofic shifts defined by $\mathcal{G}$ and $\mathcal{H}$ are identical.*

The above illustration shows that the topological entropy of the output space is related to the transition matrix.

Theorem 2.6. *Let $\mathcal{G}$ be the labeled graph obtained from the transition matrix $\mathbf{T}$ of (1). The topological entropy of the output space $\mathbf{Y}^{(2)}$ is*

$$h(\mathbf{Y}^{(2)}) = \begin{cases} \log \rho_{\mathbf{T}}, & \text{if } \mathcal{G} \text{ is right-resolving;} \\ \log \rho_{\mathbf{H}}, & \text{otherwise;} \end{cases} \tag{17}$$

where $\mathbf{H}$ is the transition matrix of the labeled graph $\mathcal{H}$ which is obtained by applying the SCM to $\mathcal{G}$.

We conclude this subsection by considering the following example.

Example 2.7. Suppose $0 < -a_1^{(1)} < -a_{-1}^{(1)}$ and $0 < -b_1^{(2)} < -b_{-1}^{(2)}$. Pick $[m, n] = [2, 3]$ in the $a^{(1)}$–$z^{(1)}$ plane and $[m, n] = [2, 2]$ in the $a^{(2)}$–$z^{(2)}$ plane. (For instance, $\mathbb{T} = (\mathbf{A}, \mathbf{B}, \mathbf{z})$ with $A_1 = (2.2, 1.7), A_2 = (-4, -2), \mathbf{B} = (-2.6, -1.4)$, and $\mathbf{z} = (-1.2, 0.3)$.) The basic sets of admissible local patterns for the first and second layers are

$$\mathcal{B}^{(1)} = \{-+-, -++, +-+, +--, --+\}$$

and

$$\mathcal{B}^{(2)} = \left\{ \begin{matrix} + \\ - \end{matrix} \; \begin{matrix} - \\ - \end{matrix}, \; \begin{matrix} + \\ + \end{matrix} \; \begin{matrix} + \\ - \end{matrix}, \; \begin{matrix} - \\ + \end{matrix} \; \begin{matrix} + \\ + \end{matrix}, \; \begin{matrix} - \\ + \end{matrix} \; \begin{matrix} - \\ - \end{matrix} \right\},$$

respectively. The transition matrices for the first and second layer are

$$T_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and}$$

$$T_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

respectively. Observe that

$$T_1^2 = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \overline{T}_1 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Therefore, the transition matrix and the symbolic transition matrix of the MNN are

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$\mathbf{S} = \begin{pmatrix} \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \\ \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \\ \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & a_{01} & \varnothing & \varnothing \\ \varnothing & \varnothing & \varnothing & \varnothing & a_{01} & a_{01} & \varnothing & \varnothing \\ \varnothing & \varnothing & a_{10} & \varnothing & \varnothing & a_{11} & \varnothing & \varnothing \\ \varnothing & \varnothing & a_{10} & \varnothing & \varnothing & a_{11} & \varnothing & \varnothing \\ \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \\ \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \end{pmatrix}$$

respectively. Since the labeled graph $\mathcal{G}$, which is obtained from $\mathbf{T}$, is not right-resolving, applying the subset construction method to $\mathcal{G}$ derives a right-resolving labeled graph $\mathcal{H}$ (cf. Fig. 3). The transition matrix of $\mathcal{H}$, indexed by p_3, p_4, p_5, p_6, $\{p_5, p_6\}$, is

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

Theorem 2.6 indicates that the topological entropy of the output space $\mathbf{Y}^{(2)}$ is $h(\mathbf{Y}^{(2)}) = \log \rho_{\mathbf{H}} = \log g$, where $g = \frac{1+\sqrt{5}}{2}$ is the golden mean.

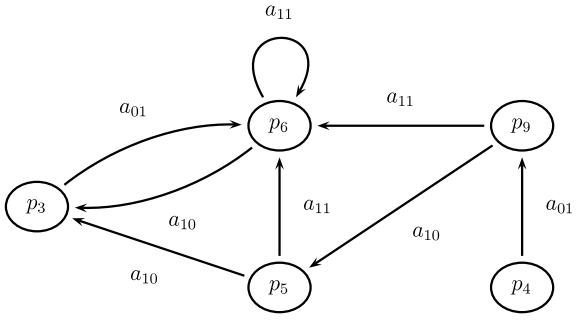


Fig. 3. The labeled graph $\mathcal{H}$ obtained by applying the SCM to $\mathcal{G}$. An extra vertex $p_9 = \{p_5, p_6\}$ is created so that $\mathcal{H}$ is right-resolving.

3. Multi-layer neural networks

This section extends the results that are obtained in the previous section to N-layer neural networks for $N \geq 2$.

Suppose $\mathbb{T} = (\mathbf{A}, \mathbf{B}, \mathbf{z})$ is a template of (1) with respect to a neighborhood $\mathcal{N} = \{-d, \ldots, -1, 1, \ldots, d\}$. The parameter space $\mathcal{P}_n$, where $n = (2d+2)N$, consists of N subspaces $\mathcal{P}^{(k)} = \{(a^{(k)}, z^{(k)}, b^{(k)}_{-d}, \ldots, b^{(k)}_{d})\} = \mathbb{R}^{2d+2}$ for $2 \leq k \leq N$ and $\mathcal{P}^{(1)} = \{(a^{(1)}, z^{(1)}, a^{(1)}_{-d}, \ldots, a^{(1)}_{d})\} = \mathbb{R}^{2d+2}$. For each k there exists $M_k \in \mathbb{N}$ and a unique collection of open subsets $\{P^{(k)}_i\}_{i=1}^{M_k}$ of $\mathcal{P}^{(k)}$ such that

(i) $\mathcal{P}^{(k)} = \bigcup_{i=1}^{M_k} \overline{P^{(k)}}_i$.
(ii) $P^{(k)}_i \bigcap P^{(k)}_j = \varnothing$ if $i \neq j$.
(iii) Templates $\mathbb{T}, \mathbb{T}' \in P^{(k)}_i$ for some i if and only if $\mathcal{B}(\mathbb{T}) = \mathcal{B}(\mathbb{T}')$.

Let $K = K_1 \cdot K_2 \cdots K_N$, define

$$P_i = (P^{(1)}_{i_1}, \ldots, P^{(N)}_{i_N}), \quad i = i_N + \sum_{j=1}^{N-1} \left((i_j - 1) \prod_{\ell=j+1}^{N} K_\ell \right),$$

for $1 \leq i_j \leq K_j$, $1 \leq j \leq N$. Theorem 3.1 asserts that the parameter space $\mathcal{P}_n$ has a unique partition.

Theorem 3.1. *Let $\mathcal{P}_n$ be the parameter space of (1), where $n = (2d+2)N$. There is a positive integer K and unique set of open subregions $\{P_k\}_{k=1}^{K}$ satisfying*

(i) $\mathcal{P}_n = \bigcup_{k=1}^{K} \overline{P}_k$.
(ii) $P_i \bigcap P_j = \varnothing$ *if* $i \neq j$.
(iii) *Templates $\mathbb{T}, \mathbb{T}' \in P_k$ for some k if and only if $\mathcal{B}(\mathbb{T}) = \mathcal{B}(\mathbb{T}')$.*

To clarify the formalism of the ordering matrix $\mathbb{X}_N$ of N-layer NNs, we consider the MNNs with respect to the nearest neighborhood (i.e., $d = 1$) and start from reconstructing the ordering matrix $\mathbb{X}_2$ as a 16×16 matrix by enlarging the size of local patterns into a rectangle (see Fig. 4). The case where $d \geq 2$ can be elucidated analogously. It comes immediately that $\mathbb{X}_2$ still associates with self-similarity. Enlarge the local patterns of (1) so that they are the patterns in the newly constructed ordering matrix. For example, if $- \diamond --$ is an admissible local pattern in $\mathcal{B}^{(2)}$, then the following 8 local patterns are selected: $\mathbb{X}_2(1,1)$, $\mathbb{X}_2(1,5)$, $\mathbb{X}_2(2,3)$, $\mathbb{X}_2(2,7)$, $\mathbb{X}_2(9,1)$, $\mathbb{X}_2(9,5)$, $\mathbb{X}_2(10,3)$, and $\mathbb{X}_2(10,7)$.

Write

$$\mathbb{X}_2 = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{pmatrix},$$
$$X_{ij} = \begin{pmatrix} x_{ij;11} & x_{ij;12} & x_{ij;13} & x_{ij;14} \\ x_{ij;21} & x_{ij;22} & x_{ij;23} & x_{ij;24} \\ x_{ij;31} & x_{ij;32} & x_{ij;33} & x_{ij;34} \\ x_{ij;41} & x_{ij;42} & x_{ij;43} & x_{ij;44} \end{pmatrix} \tag{18}$$

for $1 \leq i, j \leq 4$ as Fig. 4. $x_{ij;kl}$ means the pattern $\begin{matrix} a_{r_1 r_2} a_{r'_2 r_3} \\ a_{s_1 s_2} a_{s'_2 s_3} \end{matrix}$, where

$$\begin{aligned} &r_1 = \left\lfloor \frac{i-1}{2} \right\rfloor, \quad r_2 = i - 1 - 2r_1, \quad r'_2 = \left\lfloor \frac{j-1}{2} \right\rfloor, \\ &r_3 = j - 1 - 2r'_2, \\ &s_1 = \left\lfloor \frac{k-1}{2} \right\rfloor, \quad s_2 = k - 1 - 2s_1, \quad s'_2 = \left\lfloor \frac{l-1}{2} \right\rfloor, \\ &s_3 = l - 1 - 2s'_2. \end{aligned} \tag{19}$$

If $a_{r_1 r_2} a_{r'_2 r_3} = \varnothing$ or $a_{s_1 s_2} a_{s'_2 s_3} = \varnothing$, then $x_{ij;kl} = \varnothing$. Furthermore, if $x_{ij;kl} \neq \varnothing$, then it is denoted by the pattern $\begin{matrix} a_{r_1 a_{r_2} a_{r_3}} \\ a_{s_1 a_{s_2} a_{s_3}} \end{matrix}$ in $\{+, -\}^{\mathbb{Z}_{3\times 2}}$.

By implementing the redefined ordering matrix, we can formulate the explicit expression of the ordering matrix $\mathbb{X}_N$ of N-layer NNs. The ordering matrix $\mathbb{X}_N$ of all possible local patterns in $\{+, -\}^{\mathbb{Z}_{3\times N}}$ is defined recursively as

$$\mathbb{X}_N = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \varnothing & \varnothing \\ \varnothing & \varnothing & \mathbf{X}_{23} & \mathbf{X}_{24} \\ \mathbf{X}_{31} & \mathbf{X}_{32} & \varnothing & \varnothing \\ \varnothing & \varnothing & \mathbf{X}_{43} & \mathbf{X}_{44} \end{pmatrix}, \tag{20}$$

where

$$\mathbf{x}_{i_1 j_1} = \begin{pmatrix} X_{i_1 j_1;11} & X_{i_1 j_1;12} & \varnothing & \varnothing \\ \varnothing & \varnothing & X_{i_1 j_1;23} & X_{i_1 j_1;24} \\ X_{i_1 j_1;31} & X_{i_1 j_1;32} & \varnothing & \varnothing \\ \varnothing & \varnothing & X_{i_1 j_1;43} & X_{i_1 j_1;44} \end{pmatrix}, \tag{21}$$

$$X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k} = \begin{pmatrix} X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;11} & X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;12} & \varnothing & \varnothing \\ \varnothing & \varnothing & X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;23} & X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;24} \\ X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;31} & X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;32} & \varnothing & \varnothing \\ \varnothing & \varnothing & X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;43} & X_{i_1 j_1; i_2 j_2; \cdots; i_k j_k;44} \end{pmatrix}, \tag{22}$$

for $1 \leq k \leq N-2$, and

$$X_{i_1 j_1; i_2 j_2; \cdots; i_{N-1} j_{N-1}} = \begin{pmatrix} x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};11} & x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};12} & \varnothing & \varnothing \\ \varnothing & \varnothing & x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};23} & x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};24} \\ x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};31} & x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};32} & \varnothing & \varnothing \\ \varnothing & \varnothing & x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};43} & x_{i_1 j_1; \cdots; i_{N-1} j_{N-1};44} \end{pmatrix}, \tag{23}$$

where $1 \leq i_k, j_k \leq 4$, and $1 \leq k \leq N$. The construction contains a self-similarity property in $\mathbb{X}_N$. As discussed in the previous section, $x_{i_1 j_1; i_2 j_2; \cdots; i_{N-1} j_{N-1}; i_N j_N}$ means the pattern

$$(a_{r_{11} r_{12}} a_{r'_{12} r_{13}}) \diamond (a_{r_{21} r_{22}} a_{r'_{22} r_{23}}) \diamond \cdots \diamond (a_{r_{N1} r_{N2}} a_{r'_{N2} r_{N3}})$$

in $\{+, -\}^{\mathbb{Z}_{3\times N}}$, where $a_{r_{k1} r_{k2}} a_{r'_{k2} r_{k3}}$ is defined in (8), and

$$\begin{aligned} &r_{k1} = \left\lfloor \frac{i_k - 1}{2} \right\rfloor, \quad r_{k2} = i_k - 1 - 2r_{k1}, \\ &r'_{k2} = \left\lfloor \frac{j_k - 1}{2} \right\rfloor, \quad r_{k3} = j_k - 1 - 2r'_{k2}. \end{aligned}$$

The pattern is $\varnothing$ if $a_{r_{k1} r_{k2}} a_{r'_{k2} r_{k3}} = \varnothing$ for some $1 \leq k \leq N$. Otherwise, it is denoted by the pattern

$$(a_{r_{11}} a_{r_{12}} a_{r_{13}}) \diamond (a_{r_{21}} a_{r_{22}} a_{r_{23}}) \diamond \cdots \diamond (a_{r_{N1}} a_{r_{N2}} a_{r_{N3}})$$

in $\{+, -\}^{\mathbb{Z}_{3\times N}}$.

The newly defined ordering matrix indicates that its corresponding transition matrix is of larger dimension. Notably, Theorem 3.2 asserts that enlarging the local patterns to be rectangles helps for the determination of the transition matrix $\mathbf{T}$ of the solution space. The proof is similar to the elucidation in the previous section and Ban et al. (2009), and thus is omitted.

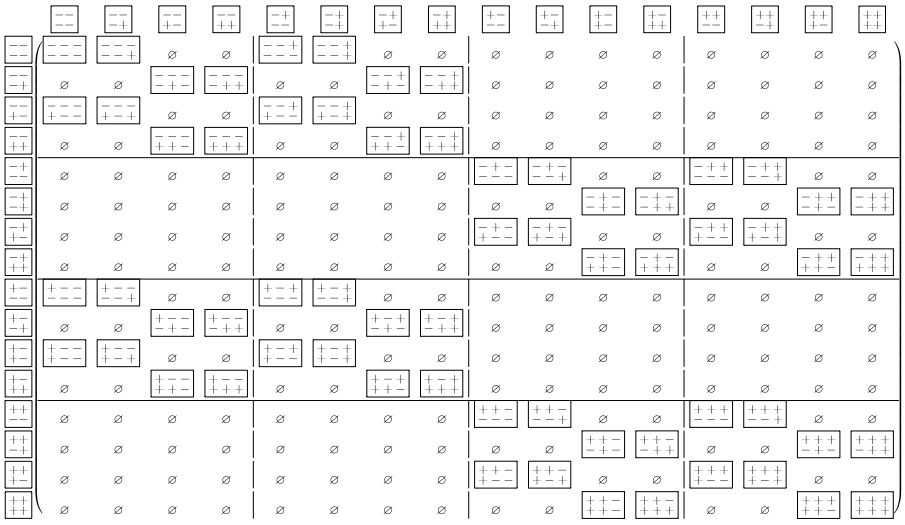


Fig. 4. The enlarged ordering matrix of two-layer neural networks.

Theorem 3.2. *Suppose* $\mathbf{T}$ *is the transition matrix of the solution space of* (1)*, and* T_k *is the transition matrix of the kth layer. Then*

$$\mathbf{T} = (T_N \otimes E_{4^{N-1}}) \circ (E_4 \otimes \mathbf{T}_{N-1}) \in \mathcal{M}_{4^{n+1}\times 4^{n+1}}(\mathbb{R}), \tag{24}$$

where

$$\mathbf{T}_k = (T_k \otimes E_{4^{k-1}}) \circ (E_4 \otimes \mathbf{T}_{k-1}) \in \mathcal{M}_{4^{k+1}\times 4^{k+1}}(\mathbb{R}),$$
$$\text{for } 3 \le k \le N-1, \tag{25}$$

and

$$\mathbf{T}_2 = T_2 \circ (E_4 \otimes T_1) \in \mathcal{M}_{16\times 16}(\mathbb{R}). \tag{26}$$

In particular, if $N = 2$*, then*

$$\mathbf{T} = T_2 \circ (E_4 \otimes T_1). \tag{27}$$

The topological entropy of the output space comes immediately from Theorem 3.2.

Theorem 3.3. *Let* $\mathcal{G}$ *be the labeled graph obtained from the transition matrix* $\mathbf{T}$ *of* (1)*. The topological entropy of the output space* $\mathbf{Y}^{(N)}$ *is*

$$h(\mathbf{Y}^{(N)}) = \begin{cases} \log \rho_{\mathbf{T}}, & \text{if } \mathcal{G} \text{ is right-resolving;} \\ \log \rho_{\mathbf{H}}, & \text{otherwise;} \end{cases} \tag{28}$$

where $\mathbf{H}$ *is the transition matrix of the labeled graph* $\mathcal{H}$ *which is obtained by applying the SCM to* $\mathcal{G}$*.*

4. The asymmetry of a topological entropy diagram

This section reveals a phenomenon that indicates the influence of the controlling template to the topological behavior of the output space, an analogous occurrence that was observed in Ban et al. (2009).

Consider a two-layer neural network with a controlling template $\mathbf{B} = (b_{-1}^{(2)}, b_1^{(2)})$ satisfying $b_1^{(2)} > b_{-1}^{(2)} > 0$. Suppose the basic set of admissible local patterns of the first layer consists of all patterns with size 3×1. In other words, there is no constraint for the input of the second layer. It is seen that, for any region $[m, n]$ in the $a^{(2)}$–$z^{(2)}$ plane such that $m, n \ge 1$, the topological entropy of the output space $\mathbf{Y}^{(2)}$ is $\log 2$. In this case, the topological entropy diagram is symmetric, that is, $h([m, n]) = h([n, m])$ for $m, n \ge 1$. Nevertheless, the symmetry is broken up by a feedback template. Notably, the topological entropy diagram of single layer neural networks is symmetric (cf. Ban & Chang, submitted for publication).

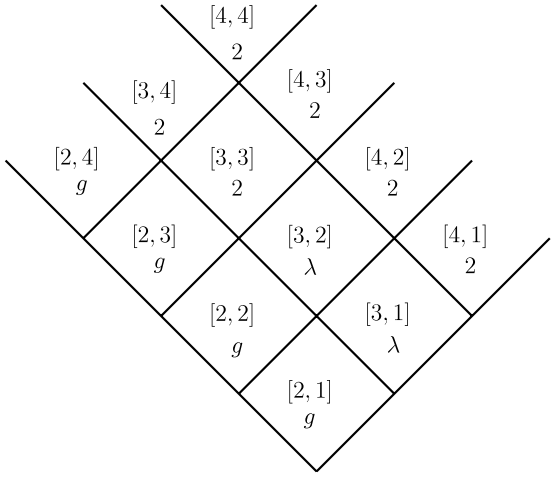


Fig. 5. The topological entropy diagram of the output space might by broken for MNNs. Consider a two-layer NN with parameter $(a^{(1)}, z^{(1)}, a_{-1}^{(1)}, a_1^{(1)}) = (1.5, -2, 2, -4)$ and $(b_{-1}^{(2)}, b_1^{(2)}) = (1, 3)$. The topological entropy for each spatially chaotic region in the $a^{(2)}$–$z^{(2)}$ plane is $h([m, n]) = \log t_{[m,n]}$, where $t_{[3,1]} = t_{[3,2]} = \lambda \approx 1.8019$ is the maximal root of $x^3 - x^2 - 2x + 1$.

Let $(a_{-1}^{(1)}, a_1^{(1)})$ be a pair that satisfies, for example, $-a_1^{(1)} > a_{-1}^{(1)} > 0$. Take $[m, n] = [2, 3]$ in the $a^{(1)}$–$z^{(1)}$ plane. The basic set of admissible local patterns of the first layer is

$$\{---, --+, +-+, -+-, ++-\},$$

and its corresponding transition matrix is

$$T_1 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Observe that the topological entropy of the space derived by the first layer is $\log g$. The topological entropy diagram of the output space, as seen in Fig. 5, is asymmetric. Such a phenomenon illustrates that the influence of the controlling template cannot be depreciated.

5. Conclusion

In this paper, we investigate the topological structure of the solution space $\mathbf{Y}$ and output space $\mathbf{Y}^{(N)}$ of multi-layer neural networks (1). It is demonstrated that $\mathbf{Y}$ is a topological Markov chain and $\mathbf{Y}^{(N)}$ is a sofic shift space in symbolic dynamical systems. More precisely,

the solution space is presented by a directed graph G, while the output space is presented by a labeled graph $\mathcal{G} = (G, \mathcal{L})$ for some labeling $\mathcal{L}$. Applying the theory of symbolic dynamics we indicate that the topological entropy of $\mathbf{Y}$ relates to the transition matrix $\mathbf{T}$, which is obtained from G; meanwhile, the topological entropy of the output space corresponds to the symbolic transition matrix $\mathbf{S}$, which is obtained from $\mathcal{G}$.

The ordering matrix of the solution space $\mathbf{Y}$ exhibits the self-similarity. Following the structure of the self-similarity is the recurrence formula of the transition matrix $\mathbf{T}$, this goes to an algorithm for the computation of the topological entropies of the solution space and output space for arbitrary number of layers we are interested.

We remark that the elucidation of the output space $\mathbf{Y}^{(N)}$ can be applied to the investigation of the nth hidden space $\mathbf{Y}^{(n)}$ for $1 \leq n \leq N-1$. Herein the nth hidden space $\mathbf{Y}^{(n)}$ is indicated by

$$\mathbf{Y}^{(n)} = \{(y_i^{(n)})_{i\in\mathbb{Z}} : (y_i^{(1)} \cdots y_i^{(N)})_{i\in\mathbb{Z}} \in \mathbf{Y}\}.$$

Hence we can investigate the inner structure of MNNs and the relation between any two layers. Further discussion is under preparation.

Acknowledgments

The authors extend their thanks to the anonymous referees for valuable opinions. The suggestions have improved this paper and motivated some further works.

Ban is partially supported by the National Science Council, ROC (Contract No NSC 100-2115-M-259-009-MY2). Chang is grateful for the partial support of the National Science Council, ROC (Contract No NSC 101-2115-M-035-002-).

References

Alsultanny, Y. A., & Aqul, M. M. (2003). Pattern recognition using multilayer neural-genetic algorithm. *Neurocomputing*, *51*, 237–247.

Ban, J.-C., & Chang, C.-H. (2011). Diamond in multi-layer cellular neural networks (submitted for publication).

Ban, J.-C., Chang, C.-H., & Lin, S.-S. (2012). The structure of multi-layer cellular neural networks. *Journal of Differential Equations*, *252*, 4563–4597.

Ban, J.-C., Chang, C.-H., Lin, S.-S., & Lin, Y.-H. (2009). Spatial complexity in multi-layer cellular neural networks. *Journal of Differential Equations*, *246*, 552–580.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*, 1–127.

Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, & J. Weston (Eds.), *Large scale kernel machines*. MIT Press.

Chua, L. O. (1998). *World scientific series on nonlinear science, series A: Vol. 31. Cnn: a paradigm for complexity*. Singapore: World Scietific.

Chua, L. O., & Yang, L. (1988). Cellular neural networks: theory. *IEEE Transactions on Circuits and Systems*, *35*, 1257–1272.

Freund, Y., & Haussler, D. (1994). *Unsupervised learning of distributions on binary vectors using two layer networks. Technical report UCSC-CRL-94-25*. Santa Cruz: University of California.

Fukushima, K. (2013a). Artificial vision by multi-layered neural networks: neocognitron and its advances. *Neural Networks*, *37*, 107–119.

Fukushima, K. (2013b). Training multi-layered neural network neocognitron. *Neural Networks*, *40*, 18–31.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Hopfield, J. J., & Tank, D. W. (1985). Neural computation of decisions in optimization problems. *Biological Cybernetics*, *52*, 141–152.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Hsu, C.-H., Juang, J., Lin, S.-S., & Lin, W.-W. (2000). Cellular neural networks: local patterns for general template. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, *10*, 1645–1659.

Juang, J., & Lin, S.-S. (2000). Cellular neural networks: mosaic pattern and spatial chaos. *SIAM Journal on Applied Mathematics*, *60*, 891–915.

Kurková, V., & Sanguineti, M. (2013). Can two hidden layers make a difference? In *Adaptive and natural computing algorithms* (pp. 30–39). Springer Berlin Heidelberg.

Lay, R. (1992). *Convex sets and their applications*. NY: Wiley.

Lind, D., & Marcus, B. (1995). *An introduction to symbolic dynamics and coding*. Cambridge: Cambridge University Press.

Peterson, C., & Söderberg, B. (1989). A new method for mapping optimization problems onto neural network. *International Journal of Neural Systems*, *1*, 3–22.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. In *Progress in brain research: Vol. 165. Computational neuroscience: theoretical insights into brain function* (pp. 33–56).

Utgoff, P. E., & Stracuzzi, D. J. (2002). Many-layered learning. *Neural Computation*, *14*, 2497–2539.

Widrow, B. (1962). Layered neural nets for pattern recognition. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *36*, 1109–1118.

Widrow, B., & Lehr, M. A. (1990). 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, *78*, 1415–1442.