How Can Deep Rectifier Networks Achieve Linear Separability and Preserve Distances?

Senjian An SENJIAN.AN@UWA.EDU.AU
School of Computer Science and Software Engineering, The University of Western Australia, Australia

Farid Boussaid FARID.BOUSSAID@UWA.EDU.AU
School of Electrical, Electronic and Computer Engineering, The University of Western Australia, Australia

Mohammed Bennamoun MOHAMMED.BENNAMOUN@UWA.EDU.AU
School of Computer Science and Software Engineering, The University of Western Australia, Australia

Abstract

This paper investigates how hidden layers of deep rectifier networks are capable of transforming two or more pattern sets to be linearly separable while preserving the distances with a guaranteed degree, and proves the universal classification power of such distance preserving rectifier networks. Through the nearly isometric nonlinear transformation in the hidden layers, the margin of the linear separating plane in the output layer and the margin of the nonlinear separating boundary in the original data space can be closely related so that the maximum margin classification in the input data space can be achieved approximately via the maximum margin linear classifiers in the output layer. The generalization performance of such distance preserving deep rectifier neural networks can be well justified by the distance-preserving properties of their hidden layers and the maximum margin property of the linear classifiers in the output layer.

1. Introduction

With the exponential increase in computing power and the development of efficient training techniques (Hinton et al., 2006; Glorot & Bengio, 2010; Sutskever, 2013), deep learning networks have achieved impressive successes across a wide variety of domains such as speech recognition ((Seide et al., 2011; Hinton et al., 2012; Deng et al., 2013), handwritten digit recognition (Ciresan et al., 2012), object recognition (Krizhevsky et al., 2012; Zeiler & Fergus, 2014; Lee et al., 2014; He et al., 2015) and face verification (Taigman et al., 2014; Sun et al., 2014).

Deep rectifier networks, wherein the rectifier (i.e. $\max(0, x)$) acts as the nonlinear activation function, are among the most successful deep learning networks. The training advantages and improved performance of rectifiers over sigmodal activation functions have been shown in a number of recent deep learning networks (Nair & Hinton, 2010; Zeiler et al., 2013; Krizhevsky et al., 2012; Bengio, 2013; Maas et al., 2013; Glorot et al., 2011), with rectifier networks providing some of the best results on several benchmark problems for object classification (Krizhevsky et al., 2012) and speech recognition (Dahl et al., 2013).

While there is a vast body of empirical evidence for the excellent generalization performance of deep neural networks, there is only a limited body of works that have sought to provide a theoretical justification of such performance. Most of the theoretical works have focused on the universal approximation power of deep neural networks for functions (Hornik et al., 1989) or for probability distributions (Le Roux & Bengio, 2010; Montufar & Ay, 2011). Recently, several publications have investigated the superior expressive power of deep networks against shallow networks (i.e., with a single hidden layer). (Delalleau & Bengio, 2011) showed that the deep network representation of a certain family of polynomials can be much more compact (i.e., with less hidden units) than that provided by a shallow network. Similarly, with the same number of hidden units, deep networks are able to separate their input space into much more regions of linearity than their shallow counterparts (Pascanu et al., 2014; Montúfar et al., 2014). However, the universal approximation power and the superior expressive power are not enough to explain the superior generalization performance of deep neural networks. In fact, there is currently no clear theoretical justification of

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

the excellent empirical performance of deep networks with a huge number of parameters.

Motivated by the fact that rectifier neural networks perform linear classification in the output layer and the generalization performance of linear classifiers (e,g. Support Vector Machines (SVM)(Cortes & Vapnik, 1995)) can be well justified by their maximum margin property, this paper investigates the distance preserving properties of the rectifier hidden layers in achieving linear separability, and establishes the link between the margin of the linear separating boundary in the output layer and the margin of the nonlinear separating boundary in the input data space. If the distances are preserved perfectly in the transformation from the data space to the output of the topmost hidden layer, the area bounded by two parallel hyperplanes in the output layer would correspond to an area bounded by two parallel manifolds in the input data space.

The link between the output and input of a hidden layer is a rectified linear transformation (RLT), namely, $\max(0, W^T\mathbf{x} + \mathbf{b})$. The only difference between linear transformations and RLTs lies in that the rectifier forces all the negative outputs to be zero. However, this seemingly small change makes all the big differences between linear transformations and RLTs. We will prove that RLT can make any disjoint data linearly separable through a cascade of two RLTs, and consequently, two-hidden-layer rectified feedforward networks are universal classifiers. Our proof is constructive, with the aid of a new proposed data model, and explains the strategies of RLTs in transforming linearly inseparable data to be linearly separable. Furthermore, we will show how RLT can preserve at least $\frac{\sqrt{2}}{2} \approx 70.7\%$ of the distance of any two vectors in the input space, and for two-hidden-layer rectifier networks, half of the distances can thus be preserved. The separating boundary area with a margin γ, bounded by two parallel hyperplanes from a linear SVM in the output layer, is related to a separating boundary area, in the original data space, bounded by two manifolds with margins varying from γ to 2γ. The generalization performance of such deep rectifier networks can be well justified by their approximate distance preserving properties and the maximum margin properties of the linear classifiers in the output layer.

The main contributions of this paper include: **1) Disjoint Convex Hull Decompositions of Data**−A new data model is proposed to construct the rectified linear units that can transform linearly inseparable data to be linearly separable; **2) Bidirectional RLT**−A new type of ReLU, which splits the positive and negative components of linear units into two separate features, is introduced for sake of distance preservation; **3) Distance Preserving Rectified Networks**–A special type of rectifier networks is identified to have both universal classification power and distance preserving properties. It is shown that, through a cascade of two orthogonal bidirectional RLTs, any two or more disjoint pattern sets can be transformed to be linearly separable under the constraint that the distance distortions are within factors from 0.5 to 1.

Notations: Throughout the paper, we use capital letters to denote matrices, lower letters for scalar numbers, and bold lower letters for vectors. Given an integer m, we use $[m]$ to denote the integer set from 1 to m, I the identity matrix with proper dimensions and $\mathbf{0}$ a vector with all elements being 0. Given a finite number of points $\mathbf{x}_i$ ($i \in [m]$) in $\mathbb{R}^n$, a convex combination of these points is a linear combination of them in which all coefficients are non-negative and sum to 1. The convex hull of a set $\mathcal{X}$, denoted by $\mathrm{CH}(\mathcal{X})$, is a set of all convex combinations of the points in $\mathcal{X}$.

The rest of this paper is organised as follows. In Section 2, we introduce the disjoint convex hull decomposition models of data, and then use this model, in Section 3, to address the power of RLTs in transforming linearly inseparable data to be linearly separable. The bidirectional RLTs are introduced and their distance preserving properties are investigated in Section 4. Section 5 addresses distance preserving rectifier networks and their universal classification power, while Section 6 concludes the paper with a discussion on the related works and future research directions.

2. Decomposition of Multiple Pattern Sets

Two pattern sets, namely $\mathcal{X}_1$ and $\mathcal{X}_2$, are called linearly separable if there exist $\mathbf{w}$ and b such that $\mathbf{w}^T\mathbf{x}+b > 0,\ \forall\ \mathbf{x} \in \mathcal{X}_1$ and $\mathbf{w}^T\mathbf{x}+b \leq 0,\ \forall\ \mathbf{x} \in \mathcal{X}_2$. It is well known that two pattern sets are linearly separable if and only if their convex hulls are disjoint. In order to investigate how linearly inseparable pattern sets can be transformed to be linearly separable, we model each pattern set with several subsets so that the convex hulls of these subsets are disjoint across different classes of patterns. More precisely, we define the disjoint convex hull decomposition model of data as below.

Definition 1 (Disjoint Convex Hull Decomposition) *Let $\mathcal{X}_k, (k \in [m])$, be m disjoint subsets in $\mathbb{R}^n$. A decomposition of $\mathcal{X}_k$, namely, $\mathcal{X}_k = \bigcup_{i=1}^{L_k} \mathcal{X}_k^i$, is called a disjoint convex hull decomposition if the unions of the convex hulls of $\mathcal{X}_k^i$, denoted by $\hat{\mathcal{X}}_k \triangleq \bigcup_{i=1}^{L_k} \mathrm{CH}(\mathcal{X}_k^i)$, are still disjoint, i.e.,*

$$\hat{\mathcal{X}}_k \cap \hat{\mathcal{X}}_l = \emptyset,\ \forall\ k \neq l \tag{1}$$

or equivalently, for all $i \in [L_k], j \in [L_l]$,

$$\mathrm{CH}(\mathcal{X}_k^i) \cap \mathrm{CH}(\mathcal{X}_l^j) = \emptyset,\ \forall\ k \neq l. \tag{2}$$

For finite pattern sets $\mathcal{X}_k$, a trivial disjoint convex hull decomposition is to select each point as a subset. Hence,

any disjoint pattern sets have at least one disjoint convex hull decomposition. The complexity of the decomposition model can be characterised by its size, i.e., the number of involved linearly separable subsets. To generate a small size disjoint convex hull decomposition, one can proceed as follows. First, we decompose $\mathcal{X}_1$ as $\mathcal{X}_1 = \bigcup_{i=1}^{L_1} \mathcal{X}_1^i$ so that the size L_1 is minimal and

$$\mathrm{CH}(\mathcal{X}_1^i) \bigcap \left(\bigcup_{l=2}^{m} \mathcal{X}_l \right) = \emptyset, \forall\, i \in [L_1]. \tag{3}$$

Then, we decompose $\mathcal{X}_2$ as $\mathcal{X}_2 = \bigcup_{j=1}^{L_2} \mathcal{X}_2^j$ so that the size L_2 is minimal and

$$\begin{aligned} \mathrm{CH}(\mathcal{X}_2^j) \quad \bigcap \quad & \left(\bigcup_{l=3}^{m} \mathcal{X}_l \right) \quad = \quad \emptyset, \forall\, j \in [L_2] \\ \mathrm{CH}(\mathcal{X}_2^j) \quad \bigcap \quad & \left(\bigcup_{i=1}^{L_1} \mathrm{CH}(\mathcal{X}_1^i) \right) \quad = \quad \emptyset, \forall\, j \in [L_2]. \end{aligned} \tag{4}$$

Sequentially, we can obtain the decompositions for all $\mathcal{X}_k$ and form a disjoint convex hull decomposition.

According to the characteristics of disjoint convex hull decomposition models, pattern sets can be categorized into three typical categories:

1) *Linearly Separable Pattern Sets*: Two linearly separable pattern sets have a disjoint decomposition convex model with $L_1 = L_2 = 1$, i.e., $\mathrm{CH}(\mathcal{X}_1) \cap \mathrm{CH}(\mathcal{X}_2) = \emptyset$;

2) *Convexly Separable Pattern Sets*: Two pattern sets are called convexly separable if they have a disjoint convex hull decomposition with $\min(L_1, L_2) = 1$. They are referred to as convexly separable because there exists a convex region which can separate one class from the other, i.e., $\mathrm{CH}(\mathcal{X}_1) \cap \mathcal{X}_2 = \emptyset$ or $\mathrm{CH}(\mathcal{X}_2) \cap \mathcal{X}_1 = \emptyset$;

3) *Convexly Inseparable Pattern Sets*: If any disjoint convex hull decomposition of $\mathcal{X}_1$ and $\mathcal{X}_2$ satisfies $\min(L_1, L_2) > 1$, then they are called convexly inseparable.

Given that the classification of linear inseparable patterns is more challenging, it is desirable to transform the linearly inseparable pattern sets into linearly separable ones since the latter have been well investigated and there are many well-developed solutions (e.g linear SVM) for them.

3. From Linear Inseparability to Linear Separability: The Roles of the Rectifier

In this section, we investigate how RLTs can transform linearly inseparable pattern sets to be linearly separable.

3.1. From Convex Separability to Linear Separability

Theorem 2 *Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two convexly separable pattern sets with a finite number of points in $\mathbb{R}^n$, $\mathrm{CH}(\mathcal{X}_1) \cap \mathcal{X}_2 = \emptyset$, $\mathcal{X}_2 = \bigcup_{j=1}^{L_2} \mathcal{X}_2^j$ with $\mathrm{CH}(\mathcal{X}_2^j) \cap \mathrm{CH}(\mathcal{X}_1) = \emptyset$, and let $\mathbf{w}_j^T \mathbf{x} + b_j$ be the linear classifiers of $\mathcal{X}_2^j$ and $\mathcal{X}_1$ such that, for any $j \in [L_2]$,*

$$\begin{array}{rcl} \mathbf{w}_j^T \mathbf{x} + b_j & \leq & 0,\ \forall\, \mathbf{x} \in \mathcal{X}_1 \\ \mathbf{w}_j^T \mathbf{x} + b_j & > & 0,\ \forall\, \mathbf{x} \in \mathcal{X}_2^j. \end{array} \tag{5}$$

Denote

$$\begin{array}{rcl} W & \triangleq & [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{L_2}] \\ \mathbf{b} & \triangleq & [b_1, b_2, \cdots, b_{L_2}]^T \\ \mathcal{Z}_k & \triangleq & \{\mathbf{z} = \max(0, W^T \mathbf{x} + \mathbf{b}) : \mathbf{x} \in \mathcal{X}_k\}, k = 1, 2. \end{array} \tag{6}$$

Then $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are linearly separable.

Proof: From the definition of $\mathcal{Z}_1$ in (6) and(5), we have $\mathcal{Z}_1 = \{\mathbf{0}\}$. Next we show that $\mathbf{0} \notin \mathrm{CH}(\mathcal{Z}_2)$. Let $\mathbf{z}$ be any vector in $\mathcal{Z}_2$. From the definition of $\mathcal{Z}_2$ in (6) and(5), it follows that $\mathbf{z} \neq \mathbf{0}$, all the entries of $\mathbf{z}$ are non-negative and at least one of them is strictly positive. Note that $\mathcal{Z}_2$ is a finite set, we have $\mathbf{0} \notin \mathrm{CH}(\mathcal{Z}_2)$, and therefore, $\mathrm{CH}(\mathcal{Z}_1) \cap \mathrm{CH}(\mathcal{Z}_2) = \emptyset$. That is, $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are linearly separable and the proof is completed.

□

From Theorem 2, we can see that two convexly separable pattern sets can be transformed to be linearly separable by squeezing one pattern set into one point while keeping this point away from the convex hull of the other class patterns. The minimal number of required rectified linear units is no larger than the minimal number of subsets into which $\mathcal{X}_2$ can be decomposed so that each subset is linearly separable from $\mathcal{X}_1$.

3.2. From Convex Inseparability to Convex Separability

Theorem 3 *Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two convexly inseparable pattern sets, and*

$$\mathcal{X}_1 \quad = \quad \bigcup_{i=1}^{L_1} \mathcal{X}_1^i, \quad \mathcal{X}_2 \quad = \quad \bigcup_{j=1}^{L_2} \mathcal{X}_2^j \tag{7}$$

be one of their disjoint convex hull decompositions ($L_1 > 1$, $L_2 > 1$), and let $\mathbf{w}_{ij}^T \mathbf{x} + b_{ij}$ be the linear classifiers of $\mathcal{X}_2^j$ and $\mathcal{X}_1^i$ such that, for any $i \in [L_1]$ and $j \in [L_2]$,

$$\begin{array}{rcl} \mathbf{w}_{ij}^T \mathbf{x} + b_{ij} & \leq & 0,\ \forall\, \mathbf{x} \in \mathcal{X}_1^i \\ \mathbf{w}_{ij}^T \mathbf{x} + b_{ij} & > & 0,\ \forall\, \mathbf{x} \in \mathcal{X}_2^j. \end{array} \tag{8}$$

Denote

$$\begin{array}{rcl} W_i & \triangleq & [\mathbf{w}_{i1}, \mathbf{w}_{i2}, \cdots, \mathbf{w}_{iL_2}] \\ \mathbf{b}_i & \triangleq & [b_{i1}, b_{i2}, \cdots, b_{iL_2}]^T \\ W & \triangleq & [W_1, W_2, \cdots, W_{L_1}] \\ \mathbf{b} & \triangleq & [\mathbf{b}_1^T, \mathbf{b}_2^T, \cdots, \mathbf{b}_{L_1}^T]^T \\ \mathcal{Z}_k & \triangleq & \{\mathbf{z} = \max(0, W^T\mathbf{x} + \mathbf{b}) : \mathbf{x} \in \mathcal{X}_k\}, k = 1, 2. \\ \mathcal{Z}_1^i & \triangleq & \{\mathbf{z} = \max(0, W^T\mathbf{x} + \mathbf{b}) : \mathbf{x} \in \mathcal{X}_1^i\}, i \in [L_1]. \end{array} \tag{9}$$

Then

$$\mathrm{CH}(\mathcal{Z}_2) \bigcap \left(\bigcup_{i=1}^{L_1} \mathrm{CH}(\mathcal{Z}_1^i) \right) = \emptyset \tag{10}$$

which implies that $\mathcal{Z}_1, \mathcal{Z}_2$ are convexly separable.

Proof: Denote, for $i \in [L_1]$ and $t \in [L_1]$,

$$\begin{array}{rcl} \mathcal{Z}_{2t} & \triangleq & \{\mathbf{z} = \max(0, W_t^T\mathbf{x} + \mathbf{b}_t) : \mathbf{x} \in \mathcal{X}_2\} \\ \mathcal{Z}_{1t}^i & \triangleq & \{\mathbf{z} = \max(0, W_t^T\mathbf{x} + \mathbf{b}_t) : \mathbf{x} \in \mathcal{X}_1^i\}. \end{array} \tag{11}$$

Note that $\mathrm{CH}(\mathcal{X}_1^i) \cap \mathcal{X}_2 = \emptyset$. Apply Theorem 2 on $\mathcal{X}_1^i, \mathcal{X}_2$ and their images, $\mathcal{Z}_{1i}^i$ and $\mathcal{Z}_{2i}$ respectively, under the transformation $\max(0, W_i^T + \mathbf{b}_i)$. Then we have

$$\mathrm{CH}(\mathcal{Z}_{1i}^i) \cap \mathrm{CH}(\mathcal{Z}_{2i}) = \emptyset, i \in [L_1] \tag{12}$$

which implies that

$$\mathrm{CH}(\mathcal{Z}_1^i) \cap \mathrm{CH}(\mathcal{Z}_2) = \emptyset, i \in [L_1]. \tag{13}$$

This implication is due to the fact that, according to the definitions in (9), W_i is a submatrix of W, $\mathbf{b}_i$ is a subvector of $\mathbf{b}$ and therefore the points in the sets $\mathcal{Z}_{1i}^i$ and $\mathcal{Z}_{2i}$ are the projections, into a lower dimensional subspace, of the points in the sets $\mathcal{Z}_1^i$ and $\mathcal{Z}_2$ respectively.

Note that $\mathcal{Z}_1 \subset \cup_{i=1}^{L_1} \mathrm{CH}(\mathcal{Z}_1^i)$, we have $\mathcal{Z}_1 \cap \mathrm{CH}(\mathcal{Z}_2) = \emptyset$ and thus, $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are convexly separable.

□

Theorem 3 shows that any two disjoint subsets $\mathcal{X}_1, \mathcal{X}_2$ in $\mathbb{R}^n$ can be transformed convexly separable through an RLT. The minimal number of required rectified linear units is no larger than the minimal value of $L_1 L_2$ such that a disjoint convex hull decomposition of $\mathcal{X}_1, \mathcal{X}_2$ exists with L_1 subsets of $\mathcal{X}_1$ and L_2 subsets of $\mathcal{X}_2$.

3.3. From Convex Inseparability to Linear Separability

Let $W, \mathbf{b}, W_i, \mathbf{b}_i, \mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_1^i, i \in [L_1]$, be defined as in Theorem 3, and $\mathbf{z} = \max(0, W^T\mathbf{x} + \mathbf{b})$ be the RLT. From (10), $\mathcal{Z}_2$ and $\mathcal{Z}_1^i$ are linearly separable. Let $\mathbf{v}_i^T\mathbf{z} + c_i$ be the linear classifiers of $\mathcal{Z}_2$ and $\mathcal{Z}_1^i$ such that

$$\begin{array}{rcl} \mathbf{v}_i^T\mathbf{z} + c_i & \leq & 0, \ \forall \ \mathbf{z} \in \mathcal{Z}_2 \\ \mathbf{v}_i^T\mathbf{z} + c_i & > & 0, \ \forall \ \mathbf{z} \in \mathcal{Z}_{1i}. \end{array} \tag{14}$$

Now let $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{L_1}], \mathbf{c} = [c_1, c_2, \cdots, c_{L_1}]^T$, $\mathbf{y} = \max(0, V^T\mathbf{z} + \mathbf{c})$ and define

$$\begin{array}{rcl} \mathcal{Y}_k & \triangleq & \{\mathbf{y} = \max(0, V^T\mathbf{z} + \mathbf{c}) : \mathbf{z} \in \mathcal{Z}_k\} \\ & = & \left\{\mathbf{y} = \max\left(0, V^T \max\{0, W^T\mathbf{x} + \mathbf{b}\} + \mathbf{c}\right) : \mathbf{x} \in \mathcal{X}_k\right\} \end{array} \tag{15}$$

for $k = 1, 2$. Then from Theorem 2, we have $\mathrm{CH}(\mathcal{Y}_1) \cap \mathrm{CH}(\mathcal{Y}_2) = \emptyset$ and thus $\mathcal{Y}_1, \mathcal{Y}_2$ are linearly separable.

Hence, any two disjoint subsets, namely $\mathcal{X}_1$ and $\mathcal{X}_2$, can be transformed to be linearly separable through a cascade of two RLTs, which require $L_1(L_2+1)$ or less rectified linear units if $\mathcal{X}_1$ and $\mathcal{X}_2$ have a disjoint convex hull decomposition with L_1 subsets of $\mathcal{X}_1$ and L_2 subsets of $\mathcal{X}_2$. The above results can be summarised by the following Theorem:

Theorem 4 *Any two disjoint subsets in $\mathbb{R}^n$ can be transformed to be linearly separable through a cascade of two RLTs.*

3.4. Multiple Sets with Pairwise Linear Separability

Given m pattern sets $\mathcal{X}_i$ in $\mathbb{R}^n$, they are said to be linearly separable if each set, namely $\mathcal{X}_i$, is linearly separable from the union of the other sets. They are said to be pairwise linearly separable if every pair of them, namely $\mathcal{X}_i$ and $\mathcal{X}_j$, are linearly separable. Linear separability is much stronger than pairwise linear separability for multiple sets. Next, we will show that any multiple sets with pairwise linear separability can be transformed to be linearly separable through an RLT, and Section 3.5 will show that any disjoint multiple sets can be transformed to be linearly separable by a cascade of two RLTs.

Let $\mathcal{X}_k$ be m pattern sets with pairwise linear separability, i.e., $\mathrm{CH}(\mathcal{X}_i) \cap \mathrm{CH}(\mathcal{X}_j) = \emptyset, \ \forall \ i \neq j$, and let $\mathbf{w}_{i,j}^T\mathbf{x} + b_{i,j}$ be the linear classifiers of $\mathcal{X}_i$ and $\mathcal{X}_j$, satisfying $\mathbf{w}_{j,i} = -\mathbf{w}_{i,j}, b_{j,i} = -b_{i,j}$ and

$$\begin{array}{rcl} \mathbf{w}_{i,j}^T\mathbf{x} + b_{i,j} & \leq & 0, \ \forall \ \mathbf{x} \in \mathcal{X}_i, \\ \mathbf{w}_{i,j}^T\mathbf{x} + b_{i,j} & > & 0, \ \forall \ \mathbf{x} \in \mathcal{X}_j. \end{array} \tag{16}$$

Denote

$$\begin{array}{rcl} W_i & \triangleq & [\mathbf{w}_{i,1}, \cdots, \mathbf{w}_{i,i-1}, \mathbf{w}_{i,i+1}, \cdots, \mathbf{w}_{i,m}] \\ \mathbf{b}_i & \triangleq & [b_{i,1}, \cdots, b_{i,i-1}, b_{i,i+1}, \cdots, b_{i,m}] \\ W & \triangleq & [W_1, W_2, \cdots, W_m] \\ \mathbf{b} & \triangleq & [\mathbf{b}_1^T, \mathbf{b}_2^T, \cdots, \mathbf{b}_m^T]^T \\ \mathcal{Z}_j^i & \triangleq & \{\max(0, W_i^T\mathbf{x} + \mathbf{b}_i) : \mathbf{x} \in \mathcal{X}_j\} \\ \mathcal{Z}_i & \triangleq & \{\max(0, W^T\mathbf{x} + \mathbf{b}) : \mathbf{x} \in \mathcal{X}_i\}. \end{array} \tag{17}$$

for $i \in [m]$ and $j \in [m]$. Apply Theorem 2 on $\mathcal{X}_i, \cup_{j \neq i}\mathcal{X}_j$ (as two convexly separable sets in Theorem 2) with the RLT $\mathbf{z} = \max(0, W_i^T\mathbf{x} + \mathbf{b}_i)$. Then we have

$$\mathrm{CH}\left(\mathcal{Z}_i^i\right) \cap \mathrm{CH}\left(\cup_{j \neq i}\mathcal{Z}_j^i\right) = \emptyset \tag{18}$$

which, similar to the implication from (12) to (13), implies that

$$\mathrm{CH}\left(\mathcal{Z}_i\right) \cap \mathrm{CH}\left(\cup_{j\neq i}\mathcal{Z}_j\right) = \emptyset. \tag{19}$$

Hence, the multiple sets $\mathcal{Z}_i$, transformed from $\mathcal{X}_i$, are linearly separable and we have the following Theorem

Theorem 5 *Any m subsets in $\mathbb{R}^n$ with pairwise linear separability can be transformed to be linearly separable through an $m(m-1)$ dimensional RLT.*

3.5. Multiple Disjoint Sets

Let $\mathcal{X}_k$ be m disjoint subsets, $\hat{\mathcal{X}}_k \triangleq \bigcup_{l=1,l\neq k}^{m} \mathcal{X}_l$, and

$$\mathcal{X}_k \;=\; \bigcup_{i=1}^{L_k} \mathcal{X}_k^i, \quad \hat{\mathcal{X}}_k \;=\; \bigcup_{j=1}^{\hat{L}_k} \hat{\mathcal{X}}_k^j \tag{20}$$

be a disjoint convex hull decomposition of $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$. That is, $\mathrm{CH}(\mathcal{X}_k^i)\cap\mathrm{CH}(\hat{\mathcal{X}}_k^j) = \emptyset$ and there exist linear classifiers $\mathbf{w}_{kij}^T\mathbf{x} + b_{kij}$ such that

$$\begin{array}{rcl} \mathbf{w}_{kij}^T\mathbf{x} + b_{kij} & \leq & 0,\ \forall\, \mathbf{x} \in \mathcal{X}_k^i \\ \mathbf{w}_{kij}^T\mathbf{x} + b_{kij} & > & 0,\ \forall\, \mathbf{x} \in \hat{\mathcal{X}}_k^j. \end{array} \tag{21}$$

Denote

$$\begin{array}{rcl} W_{ki} & \triangleq & [\mathbf{w}_{ki1}, \mathbf{w}_{ki2}, \cdots, \mathbf{w}_{ki\hat{L}_k}] \\ \mathbf{b}_{ki} & \triangleq & [b_{ki1}, b_{ki2}, \cdots, b_{ki\hat{L}_k}]^T \\ W_k & \triangleq & [W_{k1}, W_{k2}, \cdots, W_{kL_k}] \\ \mathbf{b}_k & \triangleq & [\mathbf{b}_{k1}^T, \mathbf{b}_{k2}^T, \cdots, \mathbf{b}_{kL_k}^T]^T \\ W & \triangleq & [W_1, W_2, \cdots, W_m] \\ \mathbf{b} & \triangleq & [\mathbf{b}_1^T, \mathbf{b}_2^T, \cdots, \mathbf{b}_m^T]^T \end{array} \tag{22}$$

and define the following sets, for $i \in [L_k], k \in [m]$,

$$\begin{array}{rcl} \mathcal{Z}_k^k & \triangleq & \{\mathbf{z} = \max(0, W_k^T\mathbf{x} + \mathbf{b}_k) : \mathbf{x} \in \mathcal{X}_k\} \\ \mathcal{Z}_k^{ki} & \triangleq & \{\mathbf{z} = \max(0, W_k^T\mathbf{x} + \mathbf{b}_k) : \mathbf{x} \in \mathcal{X}_k^i\} \\ \hat{\mathcal{Z}}_k^k & \triangleq & \{\mathbf{z} = \max(0, W_k^T\mathbf{x} + \mathbf{b}_k) : \mathbf{x} \in \hat{\mathcal{X}}_k\} \\ \mathcal{Z}_k & \triangleq & \{\mathbf{z} = \max(0, W^T\mathbf{x} + \mathbf{b}) : \mathbf{x} \in \mathcal{X}_k\} \\ \mathcal{Z}_k^i & \triangleq & \{\mathbf{z} = \max(0, W^T\mathbf{x} + \mathbf{b}) : \mathbf{x} \in \mathcal{X}_k^i\} \\ \hat{\mathcal{Z}}_k & \triangleq & \{\mathbf{z} = \max(0, W^T\mathbf{x} + \mathbf{b}) : \mathbf{x} \in \hat{\mathcal{X}}_k\}. \end{array} \tag{23}$$

By applying Theorem 3 on $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$ (corresponding to $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively) with the transformation $\max(0, W_k^T\mathbf{x} + \mathbf{b}_k)$, we have

$$\mathrm{CH}(\hat{\mathcal{Z}}_k^k) \bigcap \left(\bigcup_{i=1}^{L_1} \mathrm{CH}(\mathcal{Z}_k^{ki}) \right) = \emptyset \tag{24}$$

which, similar to the implication from (12) to (13), implies that

$$\mathrm{CH}(\hat{\mathcal{Z}}_k) \bigcap \left(\bigcup_{i=1}^{L_1} \mathrm{CH}(\mathcal{Z}_k^{i}) \right) = \emptyset. \tag{25}$$

Therefore, $\hat{\mathcal{Z}}_k$ is linearly separable from $\mathcal{Z}_k^i$ for each $i \in [L_k]$, and thus convexly separable from $\mathcal{Z}_k$.

Next, we construct another RLT on the sets $\mathcal{Z}_k$ and $\hat{\mathcal{Z}}_k$ to transform the pattern sets linearly separable. Let $\mathbf{v}_{ki}^T\mathbf{z} + c_{ki}$ be the linear separator of $\hat{\mathcal{Z}}_k$ and $\mathcal{Z}_k^i$ such that

$$\begin{array}{rcl} \mathbf{v}_{ki}^T\mathbf{z} + c_{ki} & \leq & 0,\ \forall\, \mathbf{z} \in \hat{\mathcal{Z}}_k \\ \mathbf{v}_{ki}^T\mathbf{z} + c_{ki} & > & 0,\ \forall\, \mathbf{z} \in \mathcal{Z}_k^i \end{array} \tag{26}$$

and define

$$\begin{array}{rcl} V_k & \triangleq & [\mathbf{v}_{k1}, \mathbf{v}_{k2}, \cdots, \mathbf{v}_{kL_k}] \\ \mathbf{c}_k & \triangleq & [c_{k1}, c_{k2}, \cdots, c_{kL_k}]^T \\ V & \triangleq & [V_1, V_2, \cdots, V_m] \\ \mathbf{c} & \triangleq & [\mathbf{c}_1^T, \mathbf{c}_2^T, \cdots, \mathbf{c}_m^T]^T \\ \mathcal{Y}_k^k & \triangleq & \{\mathbf{y} = \max(0, V_k^T\mathbf{z} + \mathbf{c}_k) : \mathbf{z} \in \mathcal{Z}_k\} \\ \hat{\mathcal{Y}}_k^k & \triangleq & \{\mathbf{y} = \max(0, V_k^T\mathbf{z} + \mathbf{c}_k) : \mathbf{z} \in \hat{\mathcal{Z}}_k\} \\ \mathcal{Y}_k & \triangleq & \{\mathbf{y} = \max(0, V^T\mathbf{z} + \mathbf{c}) : \mathbf{z} \in \mathcal{Z}_k\} \\ \hat{\mathcal{Y}}_k & \triangleq & \{\mathbf{y} = \max(0, V^T\mathbf{z} + \mathbf{c}) : \mathbf{z} \in \hat{\mathcal{Z}}_k\}. \end{array} \tag{27}$$

Then from Theorem 2, we have

$$\mathrm{CH}(\mathcal{Y}_k^k) \cap \mathrm{CH}(\hat{\mathcal{Y}}_k^k) = \emptyset \tag{28}$$

which implies that

$$\mathrm{CH}(\mathcal{Y}_k) \cap \mathrm{CH}(\hat{\mathcal{Y}}_k) = \emptyset. \tag{29}$$

By the definitions of $\mathcal{Y}_k, \hat{\mathcal{Y}}_k$ in (27) and the definitions of $\mathcal{Z}_k, \hat{\mathcal{Z}}_k$ in (23), we know that the points of $\mathcal{Y}_k, \hat{\mathcal{Y}}_k$ correspond to those of $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$ through the following transformation: $\mathbf{y} = \max\left(0, V^T \max(0, W^T\mathbf{x} + \mathbf{b}) + \mathbf{c}\right)$. Thus, we have the following Theorem:

Theorem 6 *Any multiple disjoint sets can be transformed to be linearly separable through a cascade of two RLTs.*

4. Bidirectional RLTs and Their Distance Preserving Properties

Since the rectifier of a vector discards the information of the negative elements, it does not preserve distances and can transform two very different vectors into an identical vector. For sake of distance preservation, we introduce a new type of rectifier, which keeps both the information of the positive and the negative elements.

Definition 7 *The bidirectional rectifier of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined by*

$$\mathbf{z} = \begin{bmatrix} \max(\mathbf{0}, \mathbf{x}) \\ \max(\mathbf{0}, -\mathbf{x}) \end{bmatrix}. \tag{30}$$

Unlike the rectifier, the bidirectional rectifier can preserve distances with guaranteed degrees. Let $\mathbf{x}_1, \mathbf{x}_2$ be any two vectors in $\mathbb{R}^n$ and $\mathbf{z}_1, \mathbf{z}_2$ be their bidirectional rectifications. For brevity of notations, hereafter, we denote $\mathbf{x}^+ = \max(0, \mathbf{x}), \mathbf{x}^- = \max(0, -\mathbf{x})$, and denote the k^{th} element of $\mathbf{x}$ by $\mathbf{x}(k)$. Then $\mathbf{x}^+, \mathbf{x}^-$ are non-negative vectors, $\mathbf{x} = \mathbf{x}^+ - \mathbf{x}^-$, $(\mathbf{x}^+)^T\mathbf{x}^- = 0$, and therefore

$$\begin{aligned} \|\mathbf{z}_1 - \mathbf{z}_2\|^2 &= \|\mathbf{x}_1^+ - \mathbf{x}_2^+\|^2 + \|\mathbf{x}_1^- - \mathbf{x}_2^-\|^2 \\ \|\mathbf{x}_1 - \mathbf{x}_2\|^2 &= \|(\mathbf{x}_1^+ - \mathbf{x}_1^-) - (\mathbf{x}_2^+ - \mathbf{x}_2^-)\|^2 \\ &= \|\mathbf{x}_1^+ - \mathbf{x}_2^+\|^2 + \|\mathbf{x}_1^- - \mathbf{x}_2^-)\|^2 \\ &\quad -2(\mathbf{x}_1^+ - \mathbf{x}_2^+)^T(\mathbf{x}_1^- - \mathbf{x}_2^-) \\ &= \|\mathbf{z}_1 - \mathbf{z}_2\|^2 + 2(\mathbf{x}_1^+)^T\mathbf{x}_2^- + 2(\mathbf{x}_1^-)^T\mathbf{x}_2^+ \\ &\geq \|\mathbf{z}_1 - \mathbf{z}_2\|^2. \end{aligned} \tag{31}$$

The equality $\|\mathbf{z}_1 - \mathbf{z}_2\| = \|\mathbf{x}_1 - \mathbf{x}_2\|$ holds if and only if $(\mathbf{x}_1^+)^T\mathbf{x}_2^- = 0$ and $(\mathbf{x}_1^-)^T\mathbf{x}_2^+ = 0$, or equivalently $\mathbf{x}_1(k)\mathbf{x}_2(k) \geq 0, k \in [n]$, that is, the k^{th} elements of $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same sign for all $k \in [n]$. Only the elements of $\mathbf{x}_1, \mathbf{x}_2$ with different signs contribute to the loss of the Euclidean distance in bidirectional rectification.

Now we derive the upper bound of $\|\mathbf{x}_1 - \mathbf{x}_2\|$.

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}_2\|^2 &= \|(\mathbf{x}_1^+ - \mathbf{x}_1^-) - (\mathbf{x}_2^+ - \mathbf{x}_2^-)\|^2 \\ &= \|(\mathbf{x}_1^+ - \mathbf{x}_2^+) - (\mathbf{x}_1^- - \mathbf{x}_2^-)\|^2 \\ &= \|(\mathbf{x}_1^+ - \mathbf{x}_2^+)\|^2 + \|(\mathbf{x}_1^- - \mathbf{x}_2^-)\|^2 \\ &\quad -2(\mathbf{x}_1^+ - \mathbf{x}_2^+)^T(\mathbf{x}_1^- - \mathbf{x}_2^-) \\ &= 2\|(\mathbf{x}_1^+ - \mathbf{x}_2^+)\|^2 + 2\|(\mathbf{x}_1^- - \mathbf{x}_2^-)\|^2 \\ &\quad -\|(\mathbf{x}_1^+ - \mathbf{x}_2^+) + (\mathbf{x}_1^- - \mathbf{x}_2^-)\|^2 \\ &\leq 2\|\mathbf{x}_1^+ - \mathbf{x}_2^+\|^2 + 2\|\mathbf{x}_1^- - \mathbf{x}_2^-\|^2 \\ &= 2\|\mathbf{z}_1 - \mathbf{z}_2\|^2. \end{aligned} \tag{32}$$

Furthermore, the equality $2\|\mathbf{z}_1 - \mathbf{z}_2\|^2 = \|\mathbf{x}_1 - \mathbf{x}_2\|^2$ holds if and only if $(\mathbf{x}_1^+ + \mathbf{x}_1^-) - (\mathbf{x}_2^+ + \mathbf{x}_2^-) = 0$, or equivalently $|\mathbf{x}_1(k)| = |\mathbf{x}_2(k)|$ for all $k \in [n]$.

The distance preserving properties of the bidirectional rectifier can be summarised as below.

Proposition 8 *Let $\mathbf{x}_1 \neq \mathbf{x}_2$ be any two different vectors in $\mathbb{R}^n$ and $\mathbf{z}_1, \mathbf{z}_2$ be their corresponding bidirectional rectifications. Then we have*

$$\tfrac{\sqrt{2}}{2}\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\mathbf{z}_1 - \mathbf{z}_2\| \leq \|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{33}$$

That is, the bidirectional rectifier is a contract mapping and preserves at least $\frac{\sqrt{2}}{2} \approx 70.7\%$ of the Euclidean distance of any two different vectors.

Next, we define the bidirectional rectified linear transformation and investigate its distance preserving properties.

Definition 9 *A bidirectional RLT is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^{2d}$ and is defined as*

$$\mathbf{z} = \begin{bmatrix} \max(\mathbf{0}, W^T\mathbf{x} + \mathbf{b}) \\ \max(\mathbf{0}, -W^T\mathbf{x} - \mathbf{b}) \end{bmatrix} \tag{34}$$

where $W \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^d$.

A bidirectional RLT is called singular, nonsingular, or orthogonal if $Q = WW^T$ is singular, nonsingular and orthogonal respectively.

Note that

$$\|W^T\mathbf{x}_1 - W^T\mathbf{x}_2\|^2 = (\mathbf{x}_1 - \mathbf{x}_2)^T Q(\mathbf{x}_1 - \mathbf{x}_2) \tag{35}$$

where $Q = WW^T$. Then $\|W^T\mathbf{x}_1 - W^T\mathbf{x}_2\| = \|\mathbf{x}_1 - \mathbf{x}_2\|$ if Q is orthogonal (i.e., $Q^TQ = I$); and $W^T\mathbf{x}_1 \neq W^T\mathbf{x}_2 \Leftrightarrow \mathbf{x}_1 \neq \mathbf{x}_2$ if Q is nonsingular. Then, from Proposition 8, we have

Proposition 10 *Let the bidirectional RLT be defined as in (34), $\mathbf{x}_1, \mathbf{x}_2$ be any two different vectors in $\mathbb{R}^n$ and $\mathbf{z}_1, \mathbf{z}_2$ be their responses of the transform. Then the following statements are correct:*

1). Orthogonal bidirectional RLT is a contract mapping and preserves at least $\frac{\sqrt{2}}{2} \approx 70.7\%$ of the Euclidean distance of any two different vectors, more precisely, the inequalities

$$\frac{\sqrt{2}}{2}\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\mathbf{z}_1 - \mathbf{z}_2\| \leq \|\mathbf{x}_1 - \mathbf{x}_2\| \tag{36}$$

hold if $Q = WW^T$ is orthogonal.

2). Nonsingular bidirectional RLT preserves the disjointness of any two disjoint points, i.e.,

$$\mathbf{x}_1 \neq \mathbf{x}_2 \Leftrightarrow \mathbf{z}_1 \neq \mathbf{z}_2, \text{ if } Q = WW^T \text{ is nonsingular.} \tag{37}$$

5. Universal Classification Power of Distance Preserving Rectifier Networks

In Section 3, we have shown that a cascade of two RLTs are capable of transforming any disjoint pattern sets to be linearly separable. Next, we show that a cascade of two *orthogonal* bidirectional RLTs are also capable of achieving linear separability for any two or multiple disjoint pattern sets, with additional distance preserving property due to the bidirectional rectifier and the orthogonality constraint on the weight matrix.

Theorem 11 *The following three statements are true for orthogonal bidirectional RLTs:*

1). Any two convexly separable sets can be transformed to be linearly separable through an orthogonal bidirectional RLT.

2). Any multiple disjoint sets with pairwise linear separability can be transformed to be linearly separable through an orthogonal bidirectional RLT.

3). Any two or more disjoint sets can be transformed to be linearly separable through a cascade of two orthogonal bidirectional RLTs.

To prove Theorem 11, it suffices to prove the following Lemma 12, because Theorem 11 follows from Theorem 2, Theorem 4, Theorem 5, Theorem 6, and Lemma 12.

Lemma 12 *Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be any two disjoint subsets in $\mathbb{R}^n$. The following statements are true:*

1). If $\mathcal{X}_1$ and $\mathcal{X}_2$ can be transformed to be linearly separable by an RLT, then an orthogonal bidirectional RLT exists to transform them to be linear separable.

2). If $\mathcal{X}_1$ and $\mathcal{X}_2$ can be transformed to be linearly separable by a cascade of RLTs, then a cascade of orthogonal bidirectional RLTs exist to transform them to be linear separable.

Proof: The proofs of 1) and 2) are proceeded as follows: first, we scale the RLTs, which are assumed to transform the two pattern sets to be linearly separable, so that the resulted RLTs are contract mappings; then we add some more linear units so that the resulted bidirectional RLTs are orthogonal. Note that these operations do not change the linear separability of pattern sets, the constructed orthogonal bidirectional RLTs are capable of transforming the data to be linearly separable.

1). Assume that the transformation $\mathbf{z} = \max(0, W^T\mathbf{x}+\mathbf{b})$ transforms $\mathcal{X}_1$ and $\mathcal{X}_2$ to be linearly separable and $\mathbf{w}^T\mathbf{z}+b$ be the linear classifier such that

$$\begin{array}{rcl} \mathbf{w}^T \max(0, W^T\mathbf{x}+\mathbf{b}) + b & \leq & 0;\ \forall\, \mathbf{x} \in \mathcal{X}_1 \\ \mathbf{w}^T \max(0, W^T\mathbf{x}+\mathbf{b}) + b & > & 0;\ \forall\, \mathbf{x} \in \mathcal{X}_2. \end{array} \tag{38}$$

Let Σ be a diagonal matrix with positive diagonals such that the largest eigenvalue of $W\Sigma^2W^T$ be less than 1. Denote $\hat{W} = W\Sigma, \hat{\mathbf{b}} = \Sigma\mathbf{b}$ and $\hat{\mathbf{w}} = \Sigma^{-1}\mathbf{w}$. Then it follows that $\mathbf{w}^T \max(0, W^T\mathbf{x}+\mathbf{b}) = \hat{\mathbf{w}}^T \max(0, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}})$ and therefore

$$\begin{array}{rcl} \hat{\mathbf{w}}^T \max(0, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}}) + b & \leq & 0;\ \forall\, \mathbf{x} \in \mathcal{X}_1 \\ \hat{\mathbf{w}}^T \max(0, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}}) + b & > & 0;\ \forall\, \mathbf{x} \in \mathcal{X}_2 \end{array} \tag{39}$$

which imply that the transformation $\mathbf{z} = \hat{W}^T\mathbf{x} + \hat{\mathbf{b}}$ turns the pattern sets linearly separable.

Since the largest eigenvalue of $\hat{W}\hat{W}^T$ is less than 1, $I - \hat{W}\hat{W}^T$ is positive definite and there exists $U \in \mathbb{R}^{n\times n}$ such that $UU^T = I - \hat{W}\hat{W}^T$. Denote $\bar{W} = [\hat{W}, U], \bar{\mathbf{b}} = [\hat{\mathbf{b}}^T, \mathbf{0}_n^T]^T$ and define the following bidirectional RLT

$$\mathbf{z} = \begin{bmatrix} \max(\mathbf{0}, \bar{W}^T\mathbf{x}+\bar{\mathbf{b}}) \\ \max(\mathbf{0}, -\bar{W}^T\mathbf{x}-\bar{\mathbf{b}}) \end{bmatrix} = \begin{bmatrix} \max(\mathbf{0}, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}}) \\ \max(\mathbf{0}, U^T\mathbf{x}) \\ \max(\mathbf{0}, -\hat{W}^T\mathbf{x}-\hat{\mathbf{b}}) \\ \max(\mathbf{0}, -U^T\mathbf{x}) \end{bmatrix} \tag{40}$$

Note that $\bar{W}\bar{W}^T = I$ and $\max(0, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}})$ is a subvector of $\mathbf{z}$, the bidirectional RLT (40) is orthogonal and transforms the data to be linearly separable.

2). Let $\mathbf{y} = \max\{0, V^T \max(0, W^T\mathbf{x}+\mathbf{b}) + \mathbf{c}\}$ be a cascade of RLTs, which transform the disjoint sets $\mathcal{X}_k (k \in [m])$ to be linearly separable. Assume that $W \in \mathbb{R}^{n\times d_1}$ and $V \in \mathbb{R}^{d_1\times d_2}$. Let Σ_1 be a diagonal matrix with positive diagonals such that $I - W\Sigma_1^2W^T$ is positive definite and there exists $U \in \mathbb{R}^{n\times n}$ such that $W\Sigma_1^2W^T + UU^T = I$. Denote $\hat{W} = W\Sigma_1, \hat{\mathbf{b}} = \Sigma_1\mathbf{b}, \bar{W} = [\hat{W}, U] \in \mathbb{R}^{n\times(d_1+n)}$. Then $\bar{W}\bar{W}^T = I$ and the following bidirectional RLT

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \mathbf{z}_4 \end{bmatrix} = \begin{bmatrix} \max(\mathbf{0}, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}}) \\ \max(\mathbf{0}, U^T\mathbf{x}) \\ \max(\mathbf{0}, -\hat{W}^T\mathbf{x}-\hat{\mathbf{b}}) \\ \max(\mathbf{0}, -U^T\mathbf{x}) \end{bmatrix} \tag{41}$$

is orthogonal. Let $\mathcal{Z}_k$ denote the response sets of $\mathcal{X}_k$ by the orthogonal bidirectional RLT (41). Note that $\mathbf{y} = \max\{0, V^T \max(0, W^T\mathbf{x}+\mathbf{b}) + \mathbf{c}\} = \max\{0, \hat{V}^T \max(0, \hat{W}^T\mathbf{x}+\hat{\mathbf{b}}) + \mathbf{c}\}$, where $\hat{V} = \Sigma_1^{-1}V$, is a cascade of two RLTs, which transform the sets $\mathcal{X}_k$ to be linearly separable. Consequently, the transformation $\mathbf{y} = \max(0, \hat{V}^T\mathbf{z}_1 + \mathbf{c}) = \max(0, \bar{V}^T\mathbf{z} + \mathbf{c})$ transforms the sets $\mathcal{Z}_k$ to be linearly separable, where $\mathbf{z}, \mathbf{z}_1$ are defined in (41) and $\bar{V} = [\hat{V}^T, \mathbf{0}^T, \mathbf{0}^T, \mathbf{0}^T]$. Then from statement 1) (already proved), $\mathcal{Z}_k$ can be transformed to be linearly separable by an orthogonal bidirectional RLT and therefore, $\mathcal{X}_k, k \in [m]$ can be transformed to be linearly separable by a cascade of two orthogonal bidirectional RLTs and the proof is completed.

□

Distance Preserving Rectifier Networks. An orthogonal bidirectional RLT is related to a hidden layer with a weight matrix $\hat{W} = [W, -W]$, satisfying $WW^T = I$, and a bias vector $\hat{\mathbf{b}} = [\mathbf{b}^T, -\mathbf{b}^T]^T$. Such hidden layers are referred to as distance preserving hidden layers since they have the same distance preserving property as the orthogonal bidirectional RLTs. One can formulate rectifier networks by using one or more distance preserving hidden layers, and these rectifier networks are referred to as distance preserving rectifier networks. From Theorem 11, we know that two-hidden-layer distance preserving rectifier neural networks are capable of separating any disjoint pattern sets and thus are universal classifiers. From Proposition 10, we know that each hidden layer preserves at least $\frac{\sqrt{2}}{2}$ of the distances and therefore two hidden layers preserve at least half of the distances in the original data space. That is, for any two vectors, namely $\mathbf{x}_1$ and $\mathbf{x}_2$, and their outputs $\mathbf{z}_1$ and $\mathbf{z}_2$ in the second hidden layer, we have

$$\frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\mathbf{z}_1 - \mathbf{z}_2\| \leq \|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{42}$$

Remarks. 1). It is worth noting that the orthogonality of the weight matrices W in the hidden layers is not the orthogonality of the linear units (i.e., $W^TW = I$), but the orthogonality of the weight matrix across data dimensions (i.e., $WW^T = I$). While there are at most n orthogonal linear units for n dimensional input data, one can have any large number of linear units which satisfy the orthogonality constraint $WW^T = I$. Moreover, for a randomly selected matrix W, the more columns (corresponding to linear units) it has, the closer it is to satisfy $WW^T = I$. 2). Motivated by the fact that the large distances of patterns in the input data space may not worth perfect preservation, and for sake of computational efficiency, one may conduct locality preserving projection (Niyogi, 2004) on the input data and/or the hidden layer outputs so that the number of hidden nodes can be significantly reduced while preserving the most important distances of the nearby points and ensuring that the hidden layers can still be able to transform the data to be linearly separable. To achieve the distance preservation in a certain subspace spanned by H with $H^TH = I_r$, where $r < n$ and n is the number of rows of W, one can choose the weight matrix $W = HV$ such that $VV^T = I_r$. By this selection of W, only the distances in the space spanned by H will be preserved.

6. Discussion

This paper has shown how two hidden layers of rectifier networks can transform any two or more disjoint pattern sets to be linearly separable while preserving the distance of any two vectors in the data space with a factor ranging from 0.5 to 1. This nearly isometric property makes the maximum margins achieved by the linear classifiers in the output layer closely related to the separating margins in the original data space, and makes the generalization performance of such rectifier networks well justified.

Related Work. Our work on the universal classification power of deep rectifier networks is related to the works on the universal approximation powers of deep neural networks (Hornik et al., 1989; Le Roux & Bengio, 2010; Montufar & Ay, 2011). These works consider the universal approximation power of arbitrary neural networks with one or more hidden layers. In particular, (Hornik et al., 1989) proves that any Borel measurable function can be approximated by a single hidden layer neural network. However, this proof only show the existence of such approximations for classifiers. In fact, it is not yet clear whether any disjoint pattern sets can be perfectly separated by a single hidden layer neural network. In contrast, our proof is constructive and this paper identifies a special type of two-hidden-layer deep rectifier networks which can serve as universal classifiers and whose generalization performance can also be well justified by the distance preserving property of the hidden layers and the maximum margin property of the linear classifiers in the output layer. Our effort to provide theoretical justifications for deep rectifier network's generalization performance is related to several recent publications (Delalleau & Bengio, 2011; Pascanu et al., 2014; Montúfar et al., 2014) on the superior expressive powers of deep networks against shallow networks (i.e., with a single hidden layer). These works apply the Occam's razor rule, which favours simple solutions over complex ones, and use the smaller number of required hidden units to justify the superior performance of deep rectifier networks against their shallow counterpart. However, this cannot explain why many practical deep neural networks have thousands of hidden units, with millions of parameters, but exhibit excellent generalization performance. Our work is the first to justify the generalization performance by exploring the distance preserving properties of hidden layers and establishing the link between the separating boundary margins in the output layer and those in the input data space. Although the state-of-the-art learnt deep rectifier networks may not preserve the distances as the proposed distance preserving rectifier network does, they usually have a large number of hidden units and the weight matrices are likely to be approximately orthogonal so that the hidden layers are capable of preserving the distances in certain degrees. The way of the splitting in bidirectional rectifiers was used in (Coates & Ng, 2011) to split the positive and negative components of the sparse codes into separate features and allow classifiers to weigh positive and negative responses differently.

Implications to Practical Training of Rectified Networks. Our analysis suggests that the distance preserving rectifier networks' generalization performance can be well justified even though they may have a large number of hidden units. In practical training of rectifier networks, one may add constraints on the weight matrix or add a cost function for each hidden layer to promote distance preservations.

Limitations and Future Directions. This work has focused on theoretically analysing the universal classification power and the distance preserving properties of rectifier networks, and providing an insightful explanation for the recent successes of rectifier networks in practice. However, there are other factors involved in their generalization performance. Further work is needed to investigate the properties of convolutional neural networks, as well as the training of distance preserving deep rectifier networks.

Acknowledgements: This work was supported by ARC grants (DP150100294, DP150104251), and by the Western Australia Office of Science through Applied Research Program (ARP) Shark Hazard Mitigation.

References

Bengio, Yoshua. Deep learning of representations: Looking forward. In *Statistical Language and Speech Processing*, 2013.

Ciresan, Dan, Meier, Ueli, and Schmidhuber, Jürgen. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3642–3649. IEEE, 2012.

Coates, Adam and Ng, Andrew Y. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 921–928, 2011.

Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dahl, George E, Sainath, Tara N, and Hinton, Geoffrey E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8609–8613. IEEE, 2013.

Delalleau, Olivier and Bengio, Yoshua. Shallow vs. deep sum-product networks. In *Advances in Neural Information Processing Systems*, pp. 666–674, 2011.

Deng, Li, Li, Jinyu, Huang, Jui-Ting, Yao, Kaisheng, Yu, Dong, Seide, Frank, Seltzer, Michael, Zweig, Geoffrey, He, Xiaodong, Williams, Jason, et al. Recent advances in deep learning for speech research at microsoft. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8604–8608. IEEE, 2013.

Glorot, Xavier and Bengio, Yoshua. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, volume 15, pp. 315–323, 2011.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

Hinton, Geoffrey, Osindero, Simon, and Teh, Yee-Whye. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29 (6):82–97, 2012.

Hornik, Kurt, Stinchcombe, Maxwell, and White, Halbert. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Le Roux, Nicolas and Bengio, Yoshua. Deep belief networks are compact universal approximators. *Neural computation*, 22(8):2192–2207, 2010.

Lee, Chen-Yu, Xie, Saining, Gallagher, Patrick, Zhang, Zhengyou, and Tu, Zhuowen. Deeply-supervised nets. *arXiv preprint arXiv:1409.5185*, 2014.

Maas, Andrew L, Hannun, Awni Y, and Ng, Andrew Y. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

Montufar, Guido and Ay, Nihat. Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *Neural Computation*, 23 (5):1306–1319, 2011.

Montúfar, Guido, Pascanu, Razvan, Cho, Kyunghyun, and Bengio, Yoshua. On the number of linear regions of deep neural networks. *arXiv preprint arXiv:1402.1869*, 2014.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

Niyogi, X. Locality preserving projections. In *Neural information processing systems*, volume 16, pp. 153, 2004.

Pascanu, Razvan, Montufar, Guido, and Bengio, Yoshua. On the number of inference regions of deep feed forward networks with piece-wise linear activations. In *International Conference on Learning Representations 2014(Conference Track)*, April 2014. URL http://arxiv.org/abs/1312.6026.

Seide, Frank, Li, Gang, and Yu, Dong. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech*, pp. 437–440, 2011.

Sun, Yi, Chen, Yuheng, Wang, Xiaogang, and Tang, Xiaoou. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pp. 1988–1996, 2014.

Sutskever, Ilya. *Training recurrent neural networks*. PhD thesis, University of Toronto, 2013.

Taigman, Yaniv, Yang, Ming, Ranzato, Marc'Aurelio, and Wolf, Lior. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1701–1708. IEEE, 2014.

Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

Zeiler, Matthew D, Ranzato, M, Monga, Rajat, Mao, M, Yang, K, Le, Quoc Viet, Nguyen, Patrick, Senior, A, Vanhoucke, Vincent, Dean, Jeffrey, et al. On rectified linear units for speech processing. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 3517–3521. IEEE, 2013.