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

Senjian An, Farid Boussaid, Mohammed Bennamoun

The University of Western Australia

Outline

- **Backgrounds** on Nonlinear Maximum Margin Classification and Deep Rectifier Networks
- **From Nonlinear to Linear Separability**
  - **Decomposition** of Pattern Sets
  - **Categories** of Separable Pattern Sets
  - **Transform** the Data to be Linearly Separable
- Bidirectional Rectification for Distance Preservation
- Distance Preserving Rectifier Networks
- Random Weights and Distance Preservation
- Conclusions

Nonlinear Classification

- **Nonlinear vs Linear**
  - **High Separating Power**
  - **High Risk of Overfitting**
- **Maximum Margin Separating Boundary**
  - **Intuitive, Geometrically Interpretable**
  - **Linear Case**: Linear SVM
  - **Nonlinear Case**: Metric Distortion Problem
- **An Ideal Solution**
  - **Transforms the data linearly separable while preserving the metric**
  - **Apply linear SVM on the transformed data**

Deep Rectifier Network for Classification

- **Characteristics**
  - Classifier: *nonlinear* (universal approximation capacity)
  - Training: *Large amount of parameters*
- **Empirical Successes**
  - Image Classification, Speech Recognition
  - Generalize well , *Very successful in practice*
- **Missing Theory**
  - *Why generalize well with millions of free parameters*?
  - *Why random weights do a good job*?
  - Regularization, Kernel Methods.
- **A Novel Explanation**: Metric can be well preserved in rectifier hidden layers when they transform data to be linearly separable.

Disjoint Convex Hull Decomposition

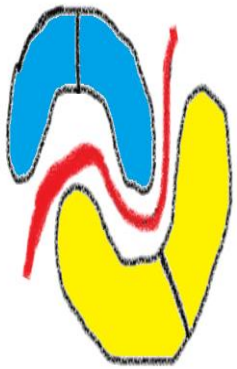
- The decomposition of two pattern sets

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

is a **disjoint convex hull decomposition** if

$$\mathrm{CH}(\mathcal{X}_1^i) \cap \mathrm{CH}(\mathcal{X}_2^j) = \emptyset, \ \forall\, i, j.$$

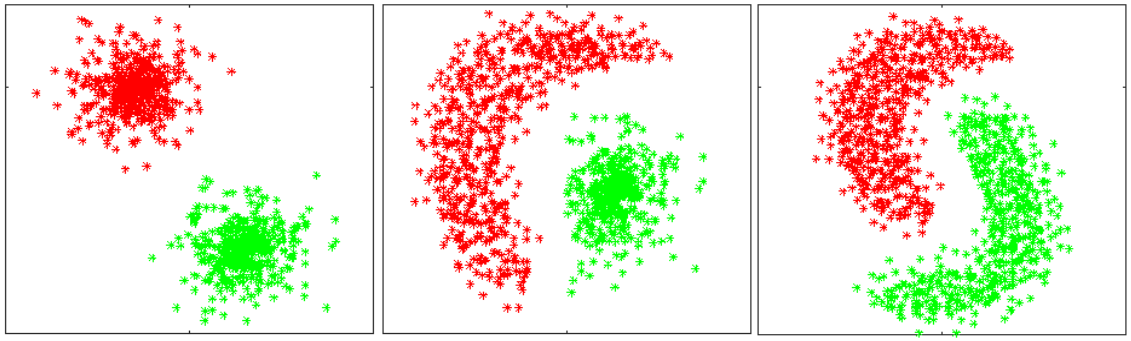
- **The subsets are linearly separable across classes**



Categories of Pattern Sets

- ***Linearly Separable:*** A separating hyper-plane exists
- ***Linearly inseparable***: No separating hyper-plane

- ***Convexly Separable*** *:* A convex region exists to include one pattern set while excluding the other
- **Convexly Inseparable**: No such convex region exists



Transformation of Data Separability

- Convexly separable pattern sets can be transformed to be linearly separable *through one hidden layer*.

- Convexly inseparable pattern sets can be transformed to be convexly separable *through one hidden layer*.

- Any Disjoint pattern sets: *two hidden layers*. The first achieves convex separability while the second achieves linear separability.

From Convexly Separable to Linearly Separable

- One pattern set can be decomposed into a few subsets each linearly separable from the other pattern set

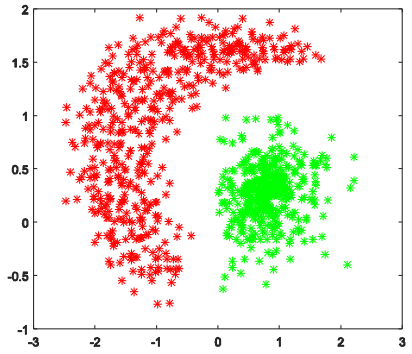
$$\mathcal{X}_2 = \bigcup_{j=1}^{L_2} \mathcal{X}_2^j$$

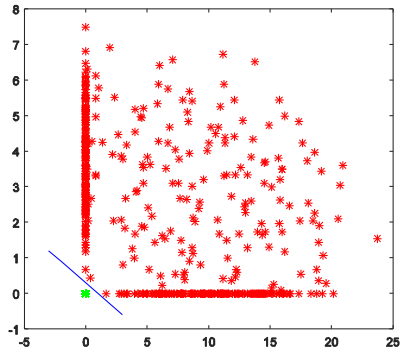
$$\mathrm{CH}(\mathcal{X}_2^j) \cap \mathrm{CH}(\mathcal{X}_1) = \emptyset,$$

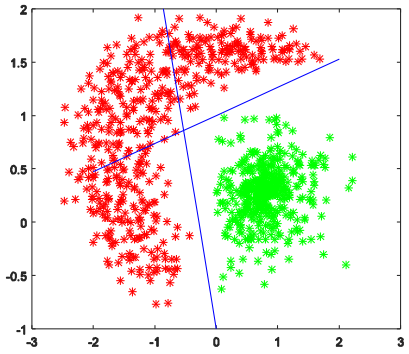
- The hidden layer, with these linear classifiers and rectification, transforms data to be linearly separable

$$\mathrm{CH}(\mathcal{Z}_1) \cap \mathrm{CH}(\mathcal{Z}_2) = \emptyset$$

An Illustrative Example







From Convexly Inseparable to Convexly Separable

- Using disjoint convex hull decomposition

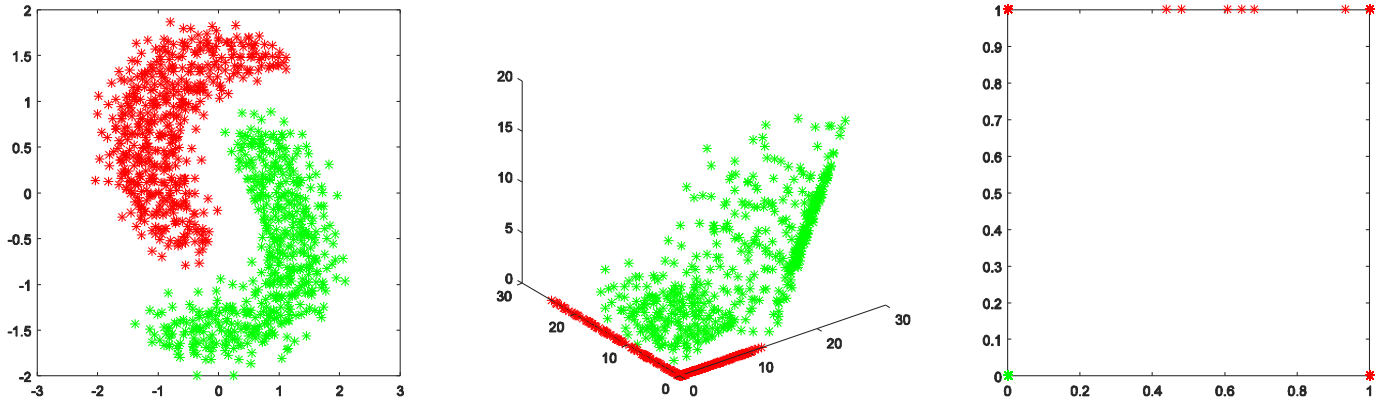
$$\mathcal{X}_1 = \bigcup_{i=1}^{L_1} \mathcal{X}_1^i, \quad \mathcal{X}_2 = \bigcup_{j=1}^{L_2} \mathcal{X}_2^j$$

- Subset $\mathcal{X}_1^i$ is convexly separable from $\mathcal{X}_2$ and thus can be transformed to be linearly separable through one hidden layer. $\mathrm{CH}(\mathcal{Z}_1^i) \cap \mathrm{CH}(\mathcal{Z}_2) = \emptyset,$
- After the transform, the two sets are convexly separable

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

From Convexly Inseparable to Linearly Separable

- First hidden layer transforms the data convexly separable;
- Second hidden layer transforms the data linearly separable.



Bidirectional Rectifier

- Rectifier discards the negative component

$$\mathbf{z} = \max(\mathbf{0}, \mathbf{x})$$

- Bidirectional Rectifier

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

- Distance Preservation

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

Bidirectional Rectified Linear Transform

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

- **Nonsingular BRLT: Nonsingular Hidden Layer**

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

- **Orthogonal BRLT: Orthogonal Hidden Layer**

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

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

Distance Preserving Rectifier Networks

- **Each Hidden Layer: an orthogonal BRLT**

$$WW^T = I$$

- **Universal Classification Power**
  - Any two or more disjoint pattern sets can be transformed to be linearly separable through two orthogonal hidden layers.

- **Generalization**
  - If SVM is applied in the output layer, the generalization performance can be well justified by the maximum margin property of SVM and the distance preserving property of the hidden layers.

Basic Lemmas for Proofs

- If two pattern sets can be transformed to be linearly separable *by one RLT*, then they can be transformed linearly separable *by an orthogonal hidden layer*
  - First scale the weight matrix to be contractive

$$WW^T \preceq I$$

  - Then add more units for orthogonality

$$[W, W_1][W, W_1]^T = WW^T + W_1 W_1^T = I$$

- If two pattern sets can be transformed to be linearly separable *by a cascade of two RLTs*, then they can be transformed linearly separable by *two orthogonal Hidden Layers*

Random Weights

- Metric Distortion
  - If the number of units is much larger than dimension, then

$$WW^T \approx \alpha I$$

  - Scale does not change the metric
  - The metric can be well preserved within a factor of two
- Improve linear separability
  - The larger the number of neurons, the better of linear separability
  - Linear separability can be improved by random chosen weights *with guaranteed level of metric preservation.*

Conclusions

- **The metric can be well preserved, *within a factor of two in two hidden layers*, in rectifier networks;**
- **Even there are a *large amount free parameters*, the generalization performance of distance-preserving rectifier networks can be well justified;**
- ***Random weights* improve linear separability without significant metric distortion if large number of units is applied.**
- ***Nonlinear maximum margin classification* can be partly achieved through rectifier net with linear SVM in the output layer.**

Thanks for Your Attention !