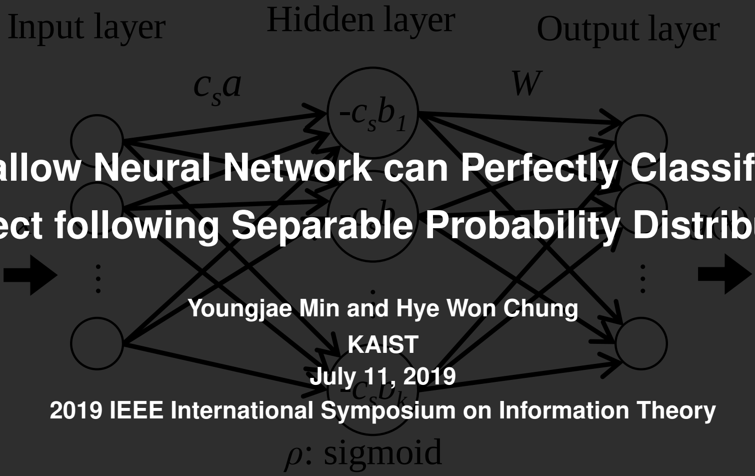
Shallow Neural Network can Perfectly Classify an Object following Separable Probability Distribution

Youngjae Min and Hye Won Chung

KAIST

July 11, 2019

2019 IEEE International Symposium on Information Theory



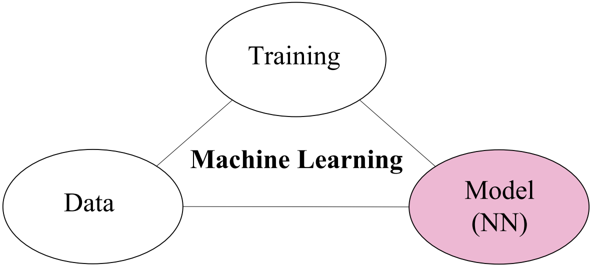
KAIST

ISIT 2019
PARIS JULY 7-12, 2019

Motivation

Problem in Machine Learning (ML)

- Choosing an architecture is very burdensome



Research Question

- From given data, can we find a proper architecture?
- What is a sufficient size of it?

Prior Works

- Universal Approximation Theorem
  "2-layer NN can approximate any function.."

→ **Just feasibility**

- C. Zhang et al., *Understanding deep learning requires rethinking generalization*, ICLR'17
  constructed 2-layer ReLU NN with $2n + d$ weights to fit a dataset with n finite samples in R^d
- H. Valvi and P. J. Ramadge, *An upper-bound on the required size of a neural network classifier*, ICASSP'18
  extended the result considering the separability of a finite dataset

→ **Just finite samples**

Can we guarantee the generalization beyond a finite dataset?

Our Purpose: Generalization

Can we guarantee the generalization beyond a finite dataset?

- An architecture which fits any datasets from a good distribution

For the rest,

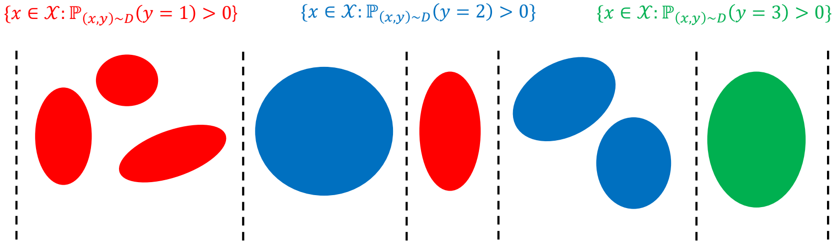
- Simple Separability
- 2-layer NN for Simple Separability
- Extended Separability
- 4-layer NN for Extended Separability

Simple Separability

Definition 1

*Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = [1:c]$. A distribution D over $\mathcal{X} \times \mathcal{Y}$ is **k-separable with δ-margin** (for some $\delta > 0$) if there exist a projection vector $a \in \mathbb{R}^d$ with $\|a\|_2 = 1$ and constants $b_1 < b_2 < \cdots < b_{k+1}$ such that, for $\mathcal{X}_i := \{x \in \mathcal{X} : b_i + \delta < a^T x < b_{i+1} - \delta\}$, $i \in [1:k]$,*

1. $\mathbb{P}_{(x,y)\sim D}(y = y_i \mid \mathcal{X}_i) = 1$ *for some* $y_i \in \mathcal{Y}$,
2. $\mathbb{P}_{(x,y)\sim D}\left(\bigcup_{i=1}^{k} \mathcal{X}_i\right) = 1$.

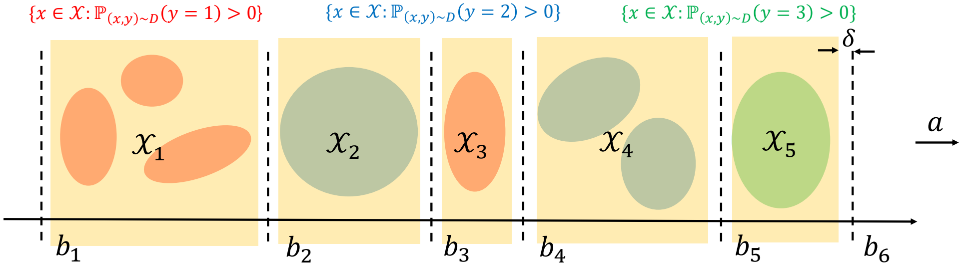


Simple Separability

Definition 1

*Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = [1:c]$. A distribution D over $\mathcal{X} \times \mathcal{Y}$ is **k-separable with δ-margin** (for some $\delta > 0$) if there exist a projection vector $a \in \mathbb{R}^d$ with $\|a\|_2 = 1$ and constants $b_1 < b_2 < \cdots < b_{k+1}$ such that, for $\mathcal{X}_i := \{x \in \mathcal{X} : b_i + \delta < a^T x < b_{i+1} - \delta\}$, $i \in [1:k]$,*

1. $\mathbb{P}_{(x,y)\sim D}\left(y = y_i \mid \mathcal{X}_i\right) = 1$ *for some* $y_i \in \mathcal{Y}$,
2. $\mathbb{P}_{(x,y)\sim D}\left(\bigcup_{i=1}^{k} \mathcal{X}_i\right) = 1.$



2-layer NN for Simple Separability

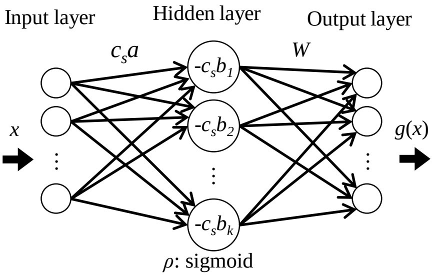
D: k-separable with δ-margin distribution, $a \in \mathbb{R}^d$: projection vector, $\{b_1, \dots, b_{k+1}\}$: boundary of intervals
For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, x: input, y: label, $f(y) \in \mathbb{R}^m$: desired output of NN ($f : \mathcal{Y} \to \mathbb{R}^m$ is injective)

Theorem 1

For any $\epsilon > 0$, the 2-layer neural network, $g : \mathcal{X} \to \mathbb{R}^m$ with parameters $a \in \mathbb{R}^d$, $\{b_1, \dots, b_k\}$,

$$W = \begin{bmatrix} f(y_1)^T \\ f(y_2)^T - f(y_1)^T \\ \vdots \\ f(y_k)^T - f(y_{k-1})^T \end{bmatrix} = [w_1\ w_2\ \cdots\ w_m], \text{ and}$$

$$c_s = (1/\delta) \log \left(\left(\sqrt{k} \cdot \left(\max_{1 \le j \le m} \|w_j\|_2 \right) \right) / \epsilon \right)$$

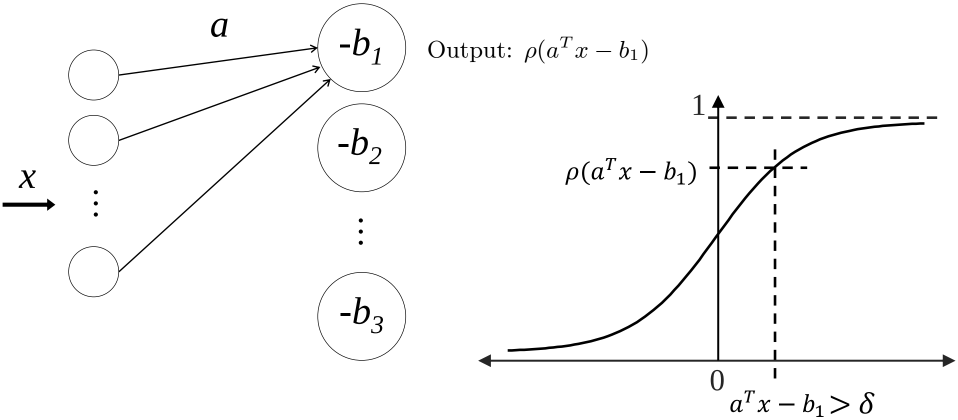


satisfies

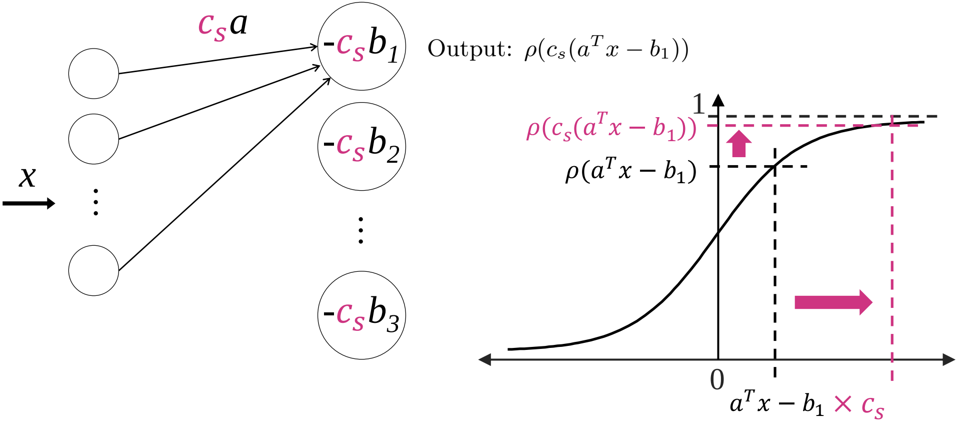
$$\mathbb{P}_{(x,y)\sim D} \left(\max_{1 \le j \le m} |g_j(x) - f_j(y)| > \epsilon \right) = 0$$

where f_j and g_j denote the j-th components of f and g, respectively.
This network is specified by total $(d + (m+1)k)$ parameters.

Main Idea: Saturation of Sigmoid through Scaling



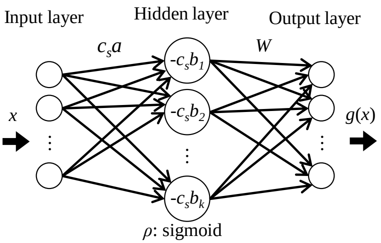
Main Idea: Saturation of Sigmoid through Scaling

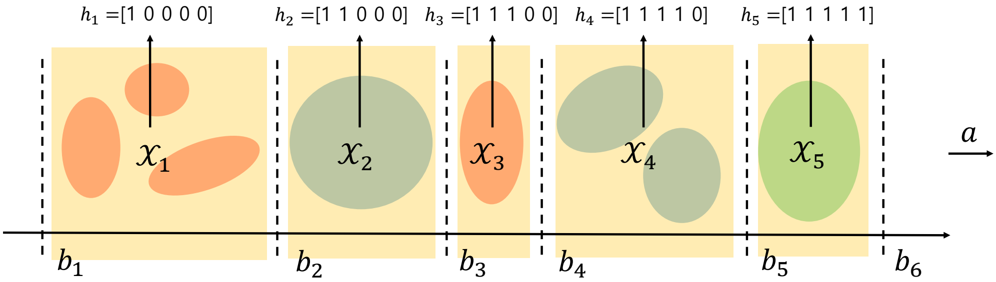


Group Behavior in Hidden Layer as $c_s \to \infty$

We can compute W s.t. $\begin{bmatrix} -h_1- \\ -h_2- \\ \vdots \\ -h_k- \end{bmatrix} W = \begin{bmatrix} -f(y_1)^T- \\ -f(y_2)^T- \\ \vdots \\ -f(y_k)^T- \end{bmatrix}$

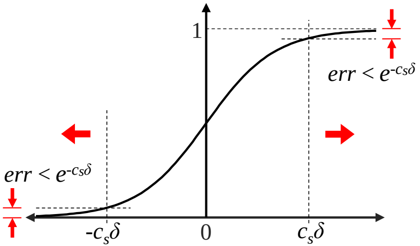
since the left matrix in LHS is invertible



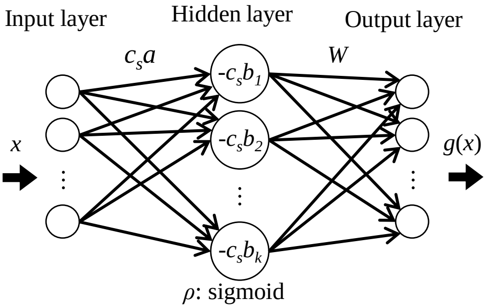


Allowing ϵ Errors in Output Layer

$c_s \to \infty$ **is impractical** $\Rightarrow$ **Can we confine** c_s **by allowing some error?**



$1 - e^{-t} < \rho(t) = 1/(1+e^{-t}) < e^{t}$



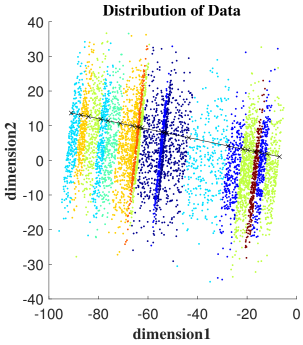
If $c_s \geq (1/\delta)\log\left(\left(\sqrt{k}\cdot(\max_{1\leq j\leq m}\|w_j\|_2)\right)/\epsilon\right)$,

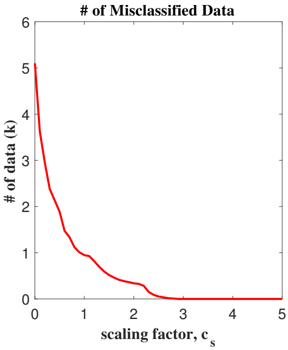
in each node of hidden layer, $err \leq \epsilon/\left(\sqrt{k}\cdot(\max_{1\leq j\leq m}\|w_j\|_2)\right)$

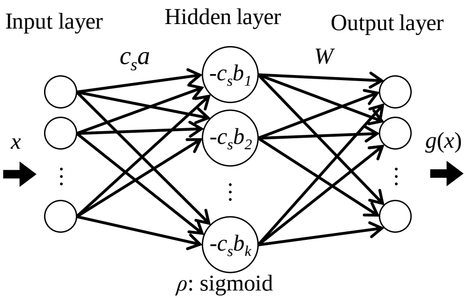
Then, in each node of output layer, $err \leq \epsilon$

2-layer NN for Simple Separability - Simulation

- f: one-hot encoding $\Rightarrow$ maximum allowable error: $\epsilon = 1/2$
- Synthetic data: 6k samples from a 20-separable with 0.1-margin distribution
- Sufficient $c_s = (1/\delta)\log\left(\left(\sqrt{k}\cdot(\max_{1\le j\le m}\|w_j\|_2)\right)/\epsilon\right) \approx 11.02$

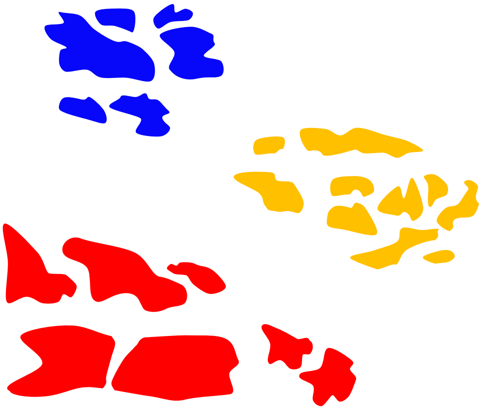






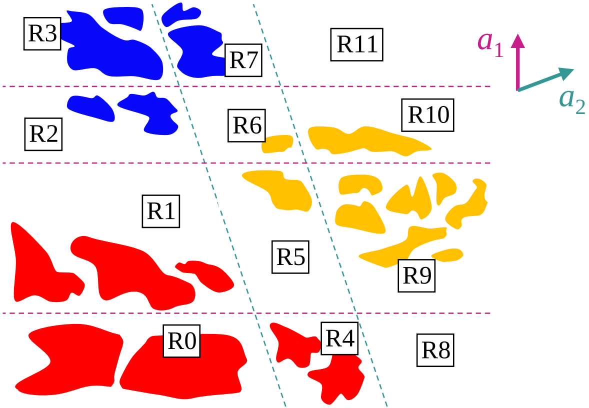
What if the data does not follow simple separability?

- Different colors for different labels



What if the data does not follow simple separability?

- Different colors for different labels

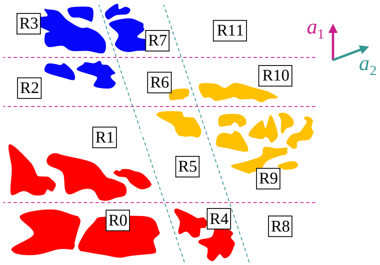


Extended Separability

Definition 2

Let $\mathcal{X} \subset \mathbb{R}^d$ *and* $\mathcal{Y}=[1 : c]$. *A distribution* D *over* $\mathcal{X} \times \mathcal{Y}$ *is* $(k_1, k_2, \cdots, k_n)$***-separable with*** δ***-margin*** *(for some* $\delta > 0$*) if there exist projection vectors* $a_1, a_2, \cdots, a_n \in \mathbb{R}^d$ *with* $\|a_s\|_2 = 1$ *and constants* $b_{s,1}<b_{s,2}<\cdots<b_{s,k_s+1}$ *for* $1\leq s\leq n$, *such that, for* $\mathcal{X}_{\mathbf{i}}=\{x \in \mathcal{X} : b_{s,i_s}+\delta < a_s^T x < b_{s,i_s+1}-\delta$ *for* $1\leq s\leq n\}$, $\mathbf{i}=(i_1, i_2, \cdots, i_n)$, *with* $i_s \in [1 : k_s]$ *for* $1\leq s\leq n$,

1. $\mathbb{P}_{(x,y)\sim D}\left(y = y_{\mathbf{i}} \mid \mathcal{X}_{\mathbf{i}}\right) = 1$ *for some* $y_{\mathbf{i}} \in \mathcal{Y}$,
2. $\mathbb{P}_{(x,y)\sim D}\left(\bigcup_{\mathbf{i}} \mathcal{X}_{\mathbf{i}}\right) = 1.$



4-layer NN for Extended Separability

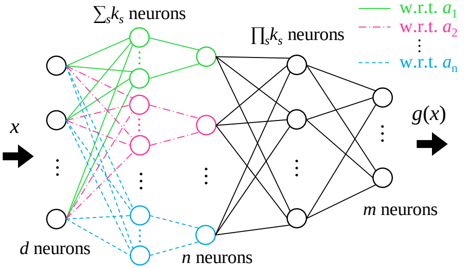
D: $(k_1, k_2, \cdots, k_n)$-separable with δ-margin distribution, $a_1, a_2, \cdots, a_n \in \mathbb{R}^d$: projection vectors
For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, x: input, y: label, $f(y) \in \mathbb{R}^m$: desired output of NN ($f: \mathcal{Y} \to \mathbb{R}^m$ is injective)

Theorem 2

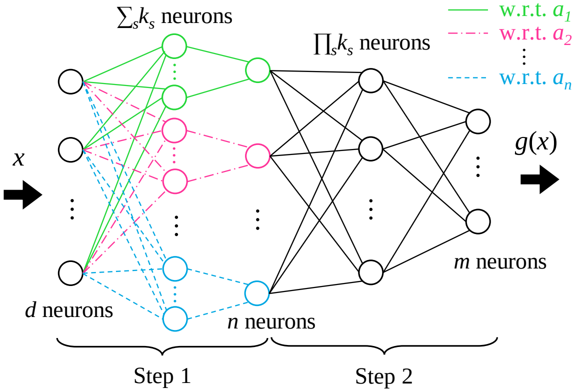
For any $\epsilon > 0$, there exists a 4-layer NN, $g: \mathcal{X} \to \mathbb{R}^m$, with $(n(d+1) + 2\sum_{s=1}^n k_s + (m+1)\prod_{s=1}^n k_s)$ parameters such that

$$\mathbb{P}_{(x,y)\sim D}\left(\max_{1\le j\le m} |g_j(x) - f_j(y)| > \epsilon\right) = 0$$

where f_j and g_j denote the j-th components of f and g, respectively.



2 Steps to Construct the 4-layer NN



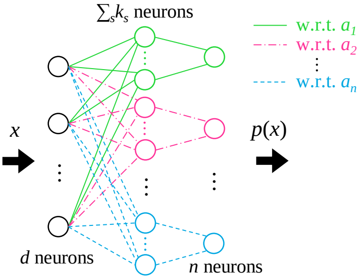
Step 1. Mapping to Simple Separable Data

Step 2. Constructing 2-layer NN for Simple Separability (Thm. 1)

Step 1. Mapping to Simple Separable Data

Lemma

For data (x, y) following a distribution D that is $(k_1, k_2, \ldots, k_n)$-separable with δ-margin by n projection vectors $(a_1, \ldots, a_n)$, there exists a 2-layer NN that implements $p : \mathcal{X} \to \mathbb{R}^n$ such that $(p(x), y)$ follows a distribution D' that is $\left(\prod_{s=1}^{n} k_s\right)$-separable with $\left(\frac{1}{4\sqrt{n}}\right)$-margin by a projection vector $a = \frac{1}{\sqrt{n}}[1, 1, \ldots, 1]^T \in \mathbb{R}^n$.



Main Idea

- Projection into $a = \frac{1}{\sqrt{n}}[1, 1, \ldots, 1]^T$ is a (scaled) component-wise summation
- Each parallel NN (approximately) outputs differently scaled integers
  ex) $\{0, 1, ..., k_1\}$ for a_1, $\{0 \times k_1, 1 \times k_1, ..., k_2 \times k_1\}$ for a_2, and so on

4-layer NN for Extended Separability

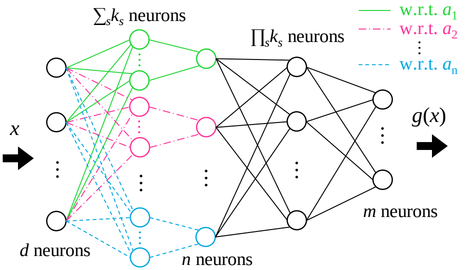
D: $(k_1, k_2, \cdots, k_n)$-separable with δ-margin distribution, $a_1, a_2, \cdots, a_n \in \mathbb{R}^d$: projection vectors
For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, x: input, y: label, $f(y) \in \mathbb{R}^m$: desired output of NN ($f : \mathcal{Y} \to \mathbb{R}^m$ is injective)

Theorem 2

For any $\epsilon > 0$, there exists a 4-layer NN, $g : \mathcal{X} \to \mathbb{R}^m$, with $(n(d+1) + 2\sum_{s=1}^{n} k_s + (m+1)\prod_{s=1}^{n} k_s)$ parameters such that

$$\mathbb{P}_{(x,y)\sim D}\left(\max_{1\le j\le m} |g_j(x) - f_j(y)| > \epsilon\right) = 0$$

where f_j and g_j denote the j-th components of f and g, respectively.



Conclusion

- Construct 4-layer sigmoid-type NN that could generalize to any datasets under the separable condition
- Demonstrate potential benefit of saturation of sigmoid func. in the generalization beyond finite samples

Remaining Questions

- How to find projection vectors and boundaries for given separable dataset?
- Can we approximate a general dataset as a separable one?
  - Error for approximating a Gaussian mixture

Full paper in arXiv:1904.09109