AN UPPER-BOUND ON THE REQUIRED SIZE OF A NEURAL NETWORK CLASSIFIER

Hossein Valavi Peter J. Ramadge

Department of Electrical Engineering
Princeton University
NJ, USA

ICASSP, 2018

Motivation

- Neural networks have been very successful
- But use many free parameters

LeNet-5: ~ 60K (*)
AlexNet: ~ 61M
VGG-16: ~ 138M
VGG-19:~ 144M
GoogleNet: ~ 5M

FractalNet: ~ 38.6M
ResNet: ~ 1.7M
Wide-ResNet: ~ 36.5M
DenseNet: ~ 1M
SqueezeNet: ~ 1.2M

- Typically size of training data << # of parameters

- We want to build a NN chip – needs to compact and lower power
- Need small Neural Nets…

Number of Parameters

CIFAR10, 50,000 training examples. Our results.

Network	#Params	Training err	Test err
ResNet-20	270K	~0%	~8.75%
ResNet-56	850K	~3.8%	~6.97%
ResNet-110	1700K	~2.2%	~6.43%
ResNet-1001	10200K	~0%	~4%
ResNet-1202	19400K	~0%	~7.93%
DenseNet-100	800K	~0%	~4.51%
DenseNet-250	15300K	~0%	~3.62%
DenseNet-190	25600K	~0%	~3.46%
Wide-ResNet-28-10	36500K	~0%	~4%

0% training error indicates interpolation of training data

The large size of NN gives rise to many questions

Why do these NN generalize well?

(G. Dziugate, et al., "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data")

Can a trained NN be compressed?

(S. Han, et al., "Deep Compression")

Can interpolators replace deep neural nets?

(M. Belkin, et al., "To understand deep learning we need to understand kernel learning")

(S. Ma, et al., "The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning")

What is the smallest interpolating neural network?

Metric: classifying the training data with 100% accuracy.
For the moment, don't worry about generalization,
and don't worry about ease of training.

Prior Work

Functional Analysis point of view:

- 2-layer (very wide) neural nets are universal function approximators. (Hornik, 1989, Not constructive)
- Adding layers increases neural net expressivity
- (Montufar, et al., 2014, total number of linearly bounded regions). Also not constructive.

Our Research Questions:

- What size network is required to interpolate the training data? (Width, Depth, # of Layers)
- Is there a constructive solution?

Start Very Simple: First Result

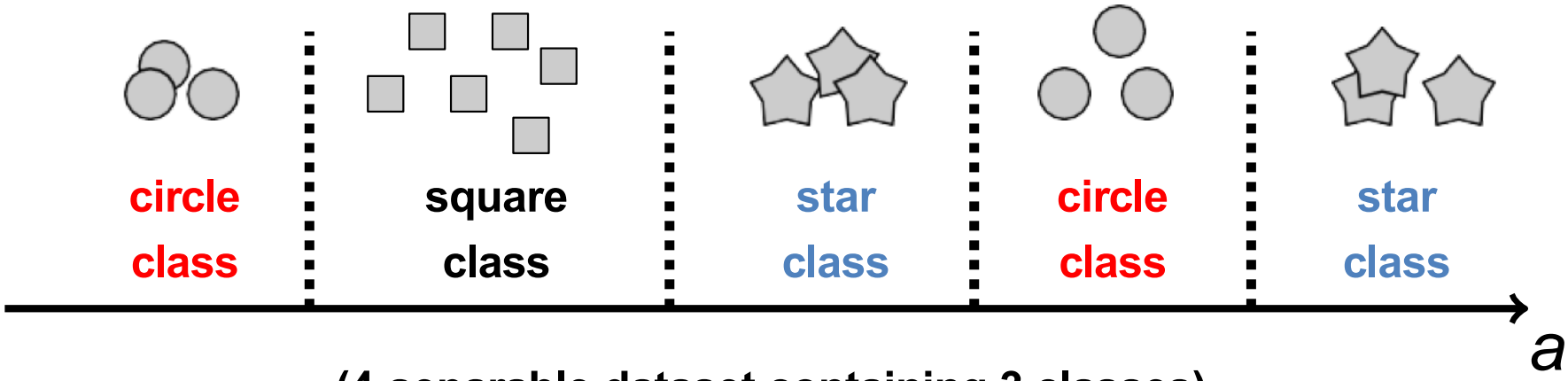
- Introduce a simple Neural Network construction under a simplifying assumption
- The construction gives an upper bound on size of a NN interpolator
- Then we set out to generalize this result
- Motivated by:

  C. Zhang et al., “Understanding deep learning requires rethinking generalization”

Definition: s-Separable Dataset

s-Separable Training Data

Let $\{(x_i, y_i)\}_{i=1}^{p}$ be a dataset, $x_i \in \mathbb{R}^d$ & $y_i \in [1:c]$.
dataset is s-separable if $\exists\ a \in \mathbb{R}^d$ s.t. projecting the data onto a creates $(s+1)$ homogeneous-label intervals.

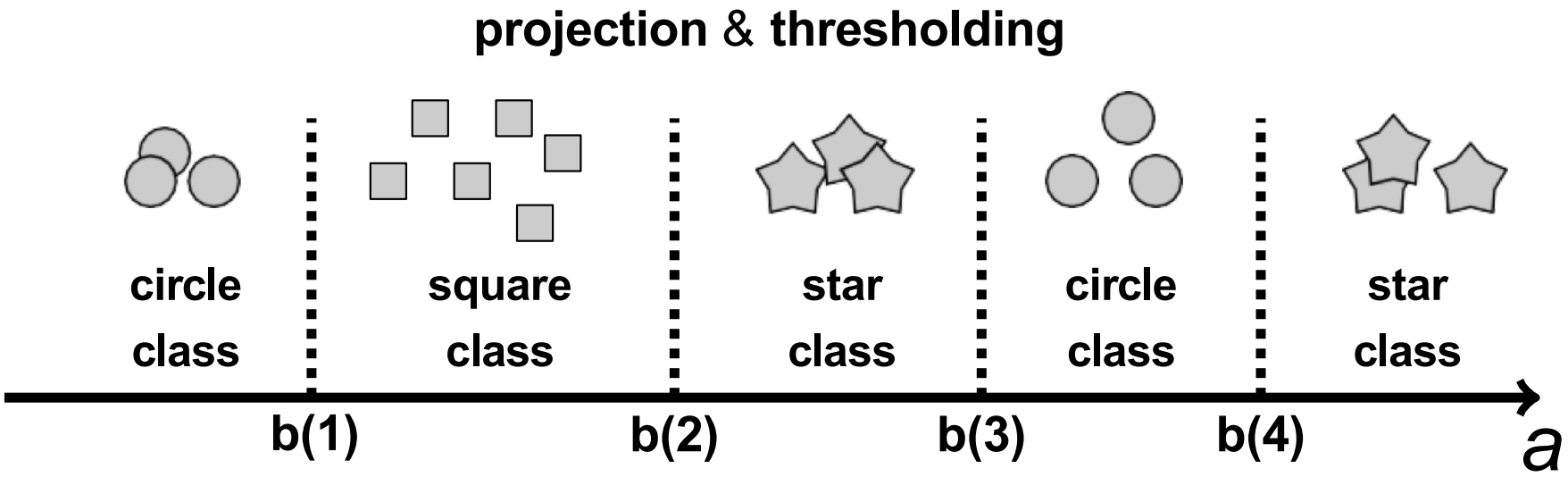


(4-separable dataset containing 3 classes)

Construction Algorithm for s-Separable Data

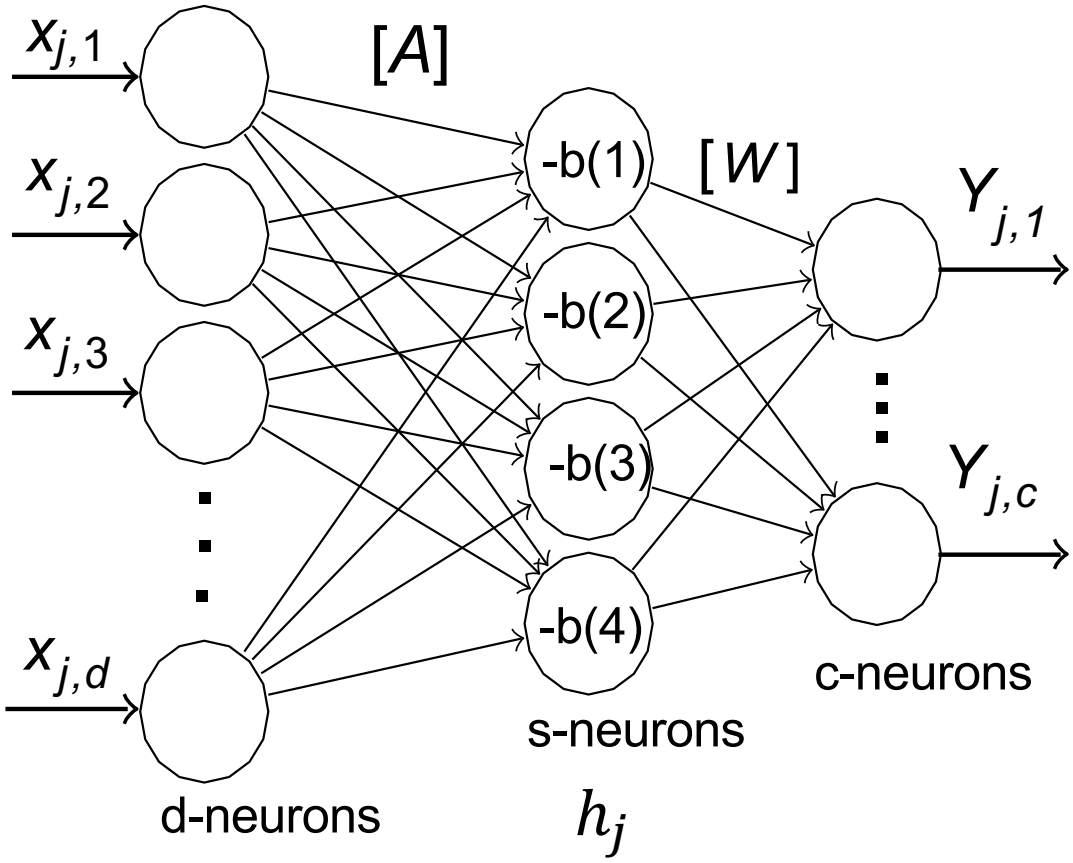
Result

For an s-separable dataset, there exists 2-layer neural net with s-hidden units that interpolates the training data.



Construction for s-Separable Data, cont.

- one-hot encoding class labels creates matrix Y$\in \mathbb{R}^{p\times c}$
- Let $A = a\mathbf{1}^T$, after applying ReLU want Y$\subset$ Range(H)

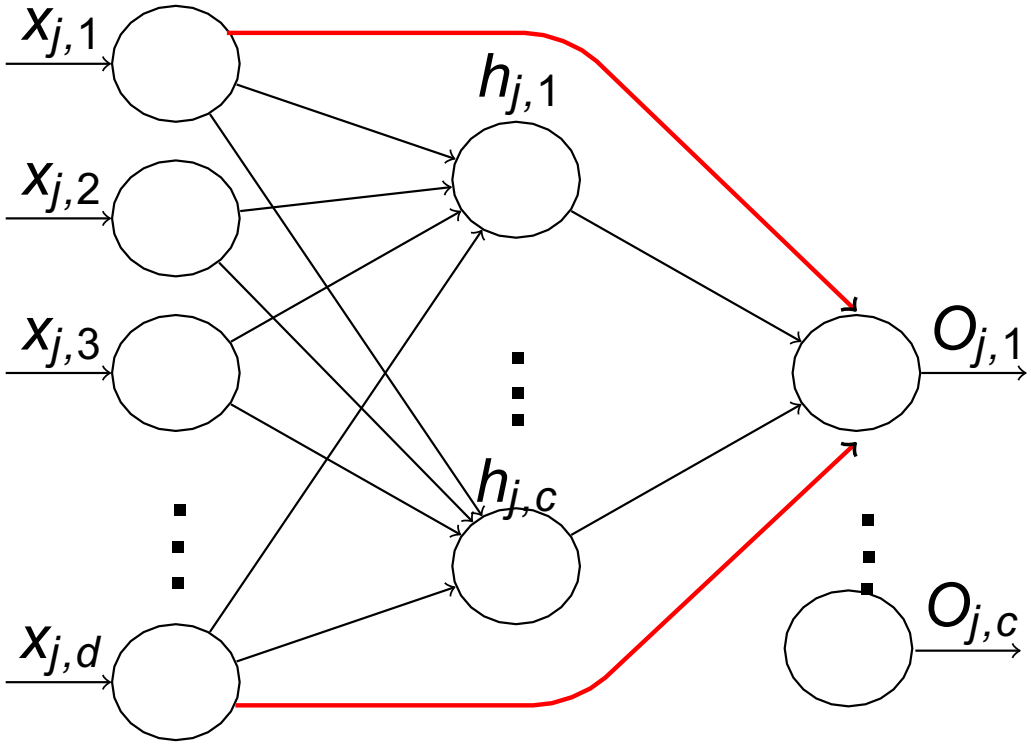


$$H = \begin{pmatrix} -- h_1^T -- \\ \vdots \\ -- h_p^T -- \end{pmatrix} \in \mathbb{R}^{p\times s}$$

Aside: ResNet & DenseNet

- Bypass connections reduces the required size of neural nets.
  (K. He, et al., “Deep Residual Learning for Image Recognition”)
  (G. Huang, et al., “Densely Connected Convolutional Networks”)

- Bypass connections increase Range(H), and hence, expressivity.



Summary so far

- Construction for a multi-class problem (*#classes* ≥ 2)

  But training data needs to be *s*-separable

- # free parameters = $d + 2s$
  - for binary-class linearly separable (s=1) data:

  $(d + 2)$ parameters; SVM $(d + 1)$ parameters

- If dataset with p examples and c classes is s-separable

$$(c - 1) \leq s \leq (p - 1)$$

Maximum Number of Parameters Needed

CIFAR10: p = 50,000 (# training examples)

In the worst case: $s = (p - 1)$

Need d+2s $\sim$ 101K parameters to interpolate training examples.

Network	#Params	Training err	Test err
ResNet-20	270K	~0%	~8.75%
ResNet-56	850K	~3.8%	~6.97%
ResNet-110	1700K	~2.2%	~6.43%
ResNet-1001	10200K	~0%	~4%
ResNet-1202	19400K	~0%	~7.93%
DenseNet-100	800K	~0%	~4.51%
DenseNet-250	15300K	~0%	~3.62%
DenseNet-190	25600K	~0%	~3.46%
Wide-ResNet-28-10	36500K	~0%	~4%

This table: Our computed errors

Other useful results that follow:

- Can Implement any quantizer with a neural net.
- Can Implement any truth-table (classic problem).

  Entries of the truth-table are the training data.

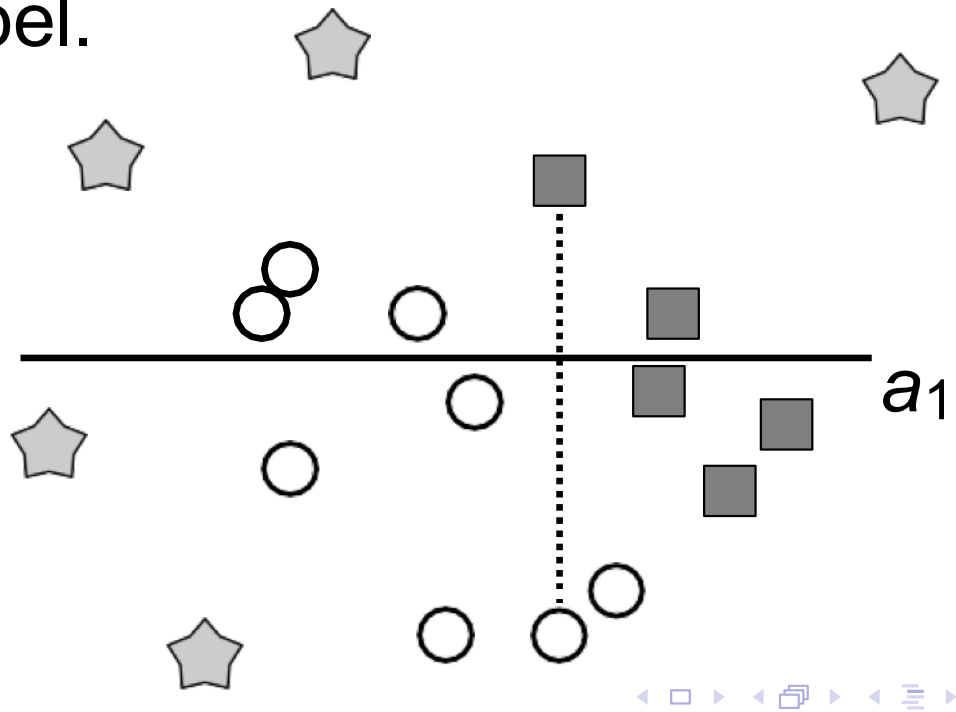
Input $\in R^{2^b}$	Output $\in R^m$
R1	Y1
R2	Y2
R3	Y3
:	:

This is useful in what follows

Construction for non s-separable Data

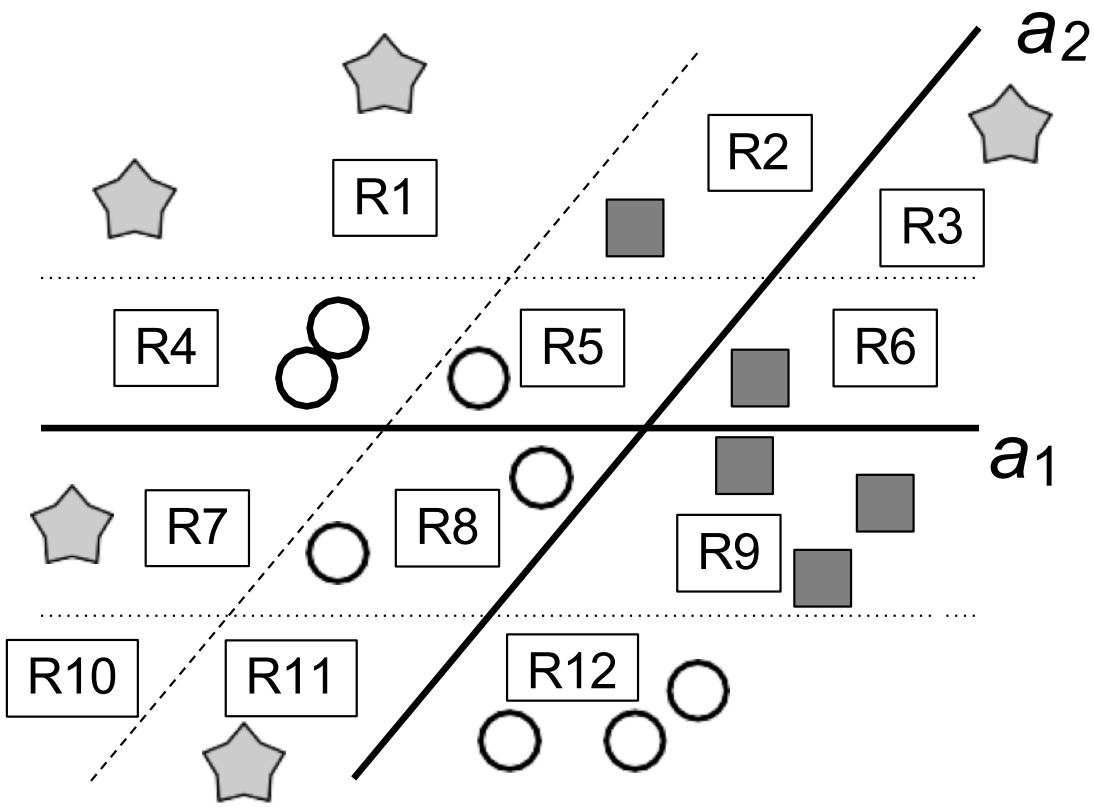
Non s-separability:

- k-points (k>1) from distinct classes are projected to the same point
- projected data cannot be clustered into (at most) s intervals, all containing elements with the same label.

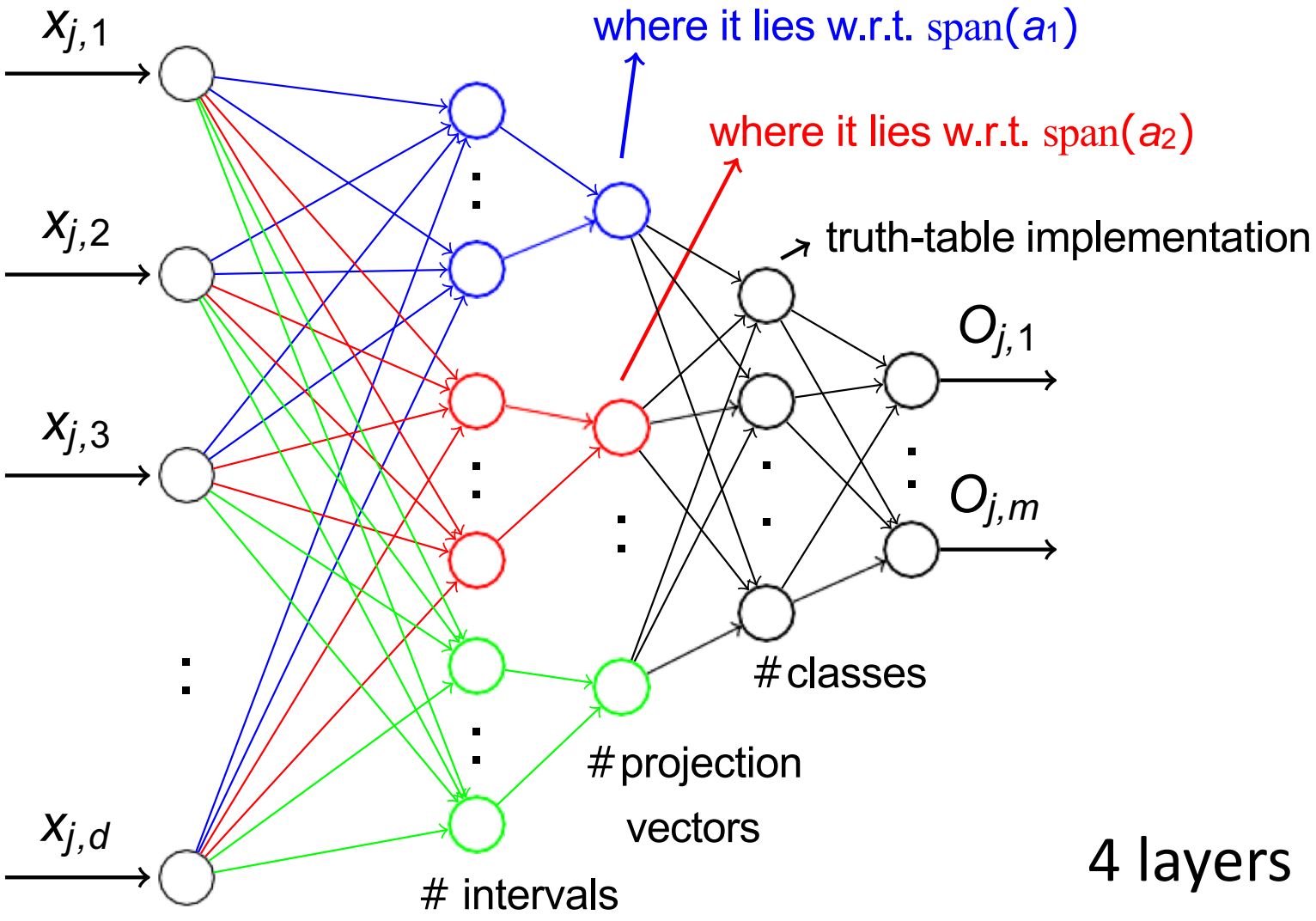


1) Construction for non-separable Data (Width)

Use multiple projection vectors:



1) Construction for non-separable Data (Width)

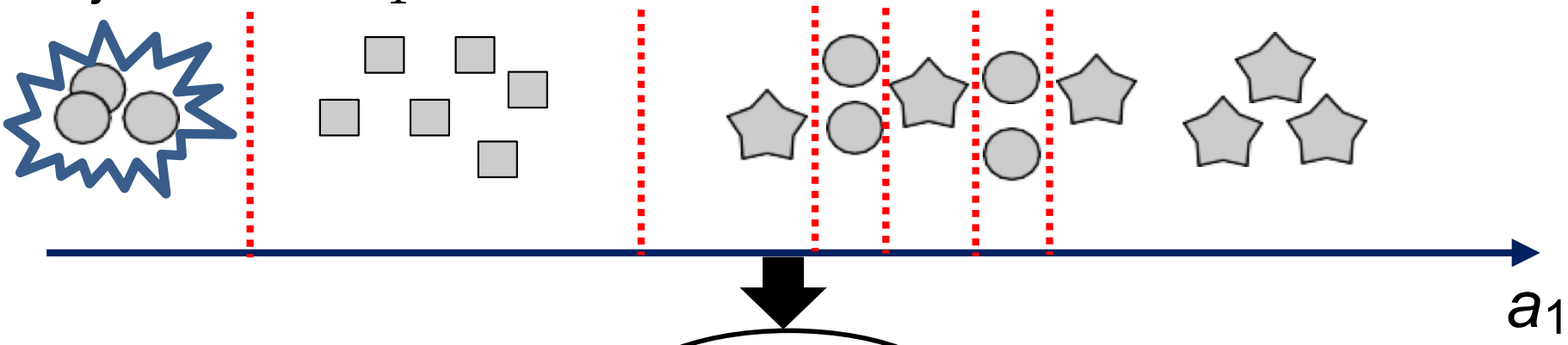


2) Construction for non-separable Data (Depth)

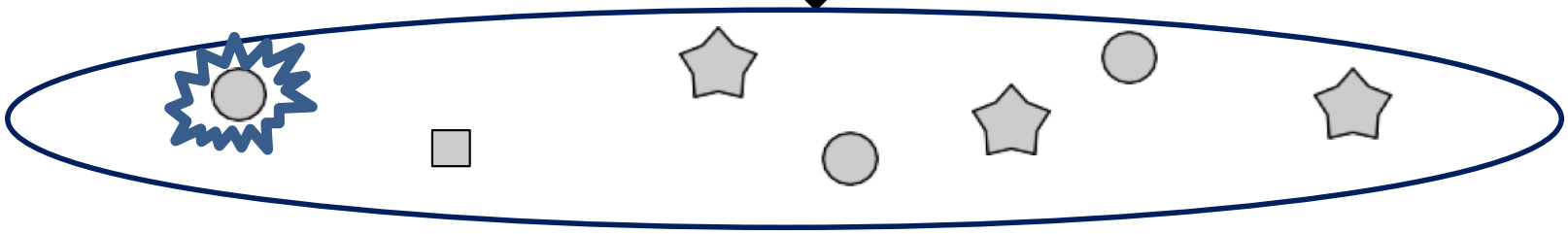
- To exploit depth in building an interpolator use a framework based on progressive refinement of the training data.
- N.N. is formed in ℓ-steps, corresponding to ℓ-layers
- Representation of the data is refined in the sequence of layers.

2) Construction for non-separable Data (Depth)

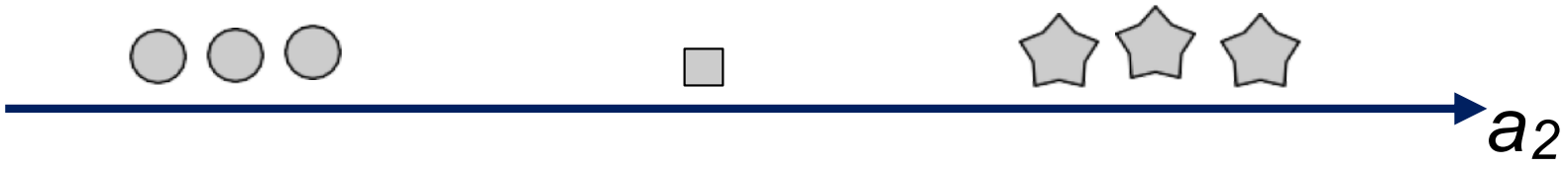
- Projection onto a_1



- New representation of data



- Projection onto a_2



Conclusion and next steps

- A framework for constructing an NN interpolator by projections and sequentially refining the data rep. in multiple layers.
- Gives a bound on the size (width, depth, # layers) of a NN interpolator
- In the worst case, ~101K parameters for CIFAR-10
- By comparison, many other popular neural networks are are also interpolators for CFAR10 but have many more parameters. (e.g. DenseNet ~ 25.6M)
- But generalization is one BIG open issue
- Now, looking into the generalization question.
- Aside: chip has been designed and fabricated will be reported at circuits conference this year [H. Valavi, et al., VLSI Symp. 2018].