Deep Learning: Theory and Practice

Deep Learning

15-03-2018

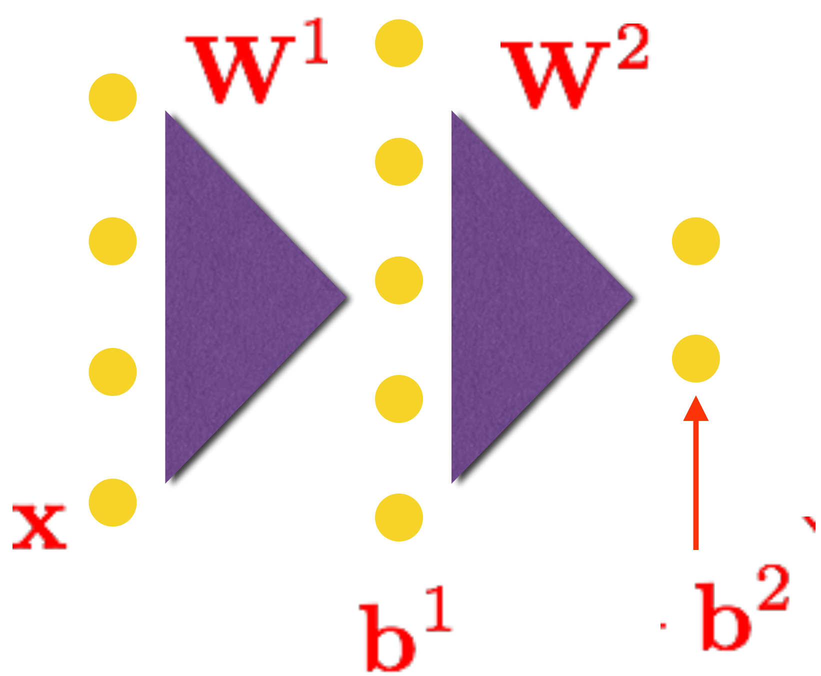
deeplearning.cce2018@gmail.com

Established
1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Neural Networks

Multi-layer Perceptron [Hopfield, 1982]



$$\mathbf{v}^2 = \psi\Big(\mathbf{W}^2\phi(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) + \mathbf{b}^2\Big)$$

ϕ non-linear function (*tanh*,*sigmoid*)

ψ thresholding function

- Useful for classifying non-linear data boundaries - non-linear class separation can be realized given enough data.

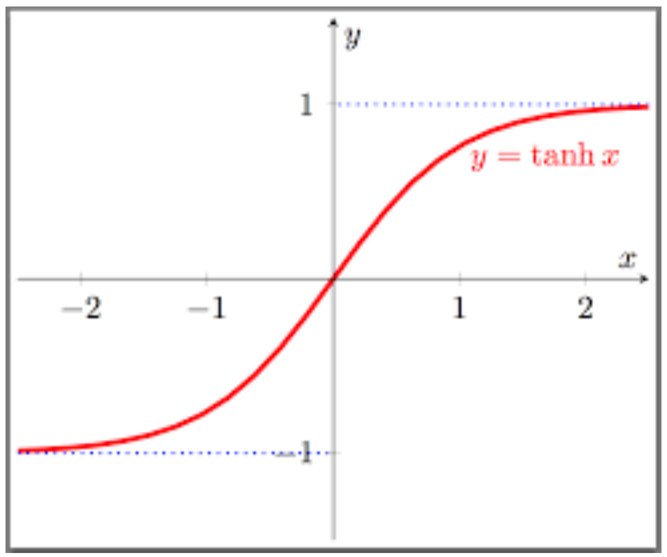
Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

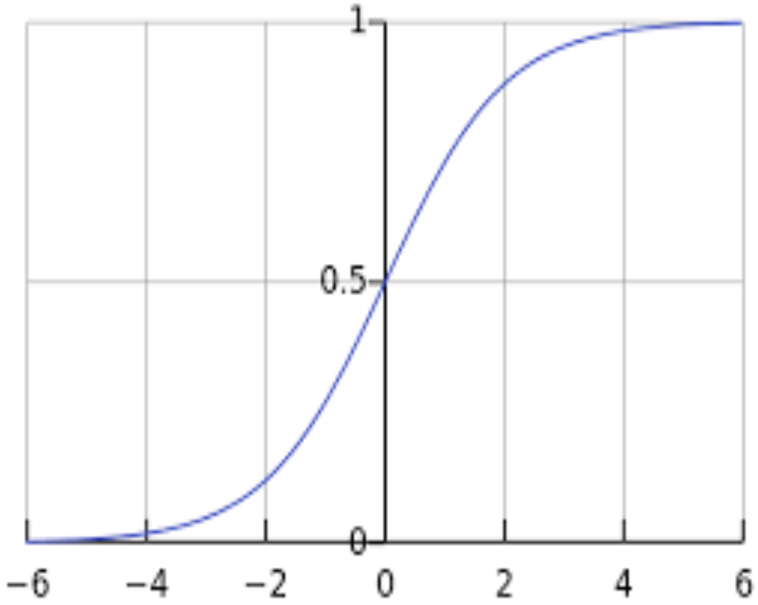
Neural Networks

Types of Non-linearities ϕ

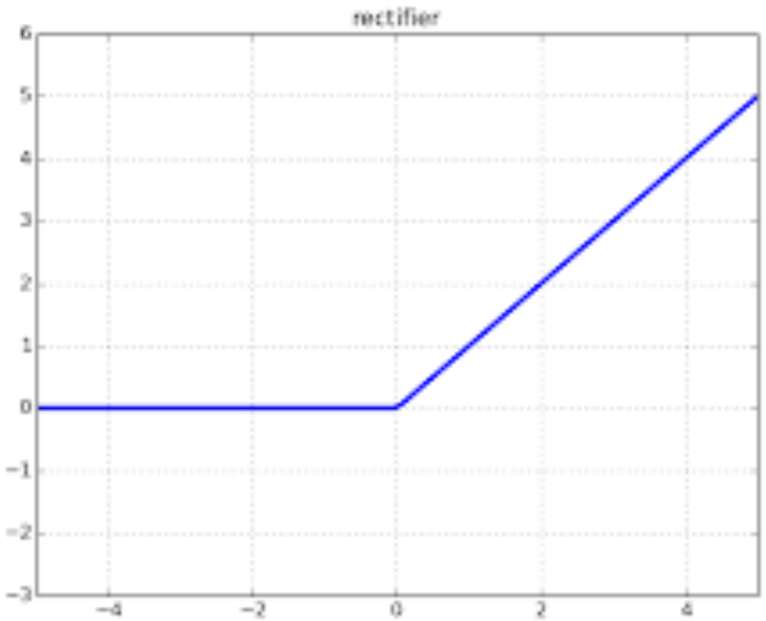
tanh



sigmoid



ReLu



Cost-Function

Mean Square Error

$$J_{MSE} = \sum_{i=1}^{M} ||\mathbf{v}_i - \mathbf{y}_i||^2$$

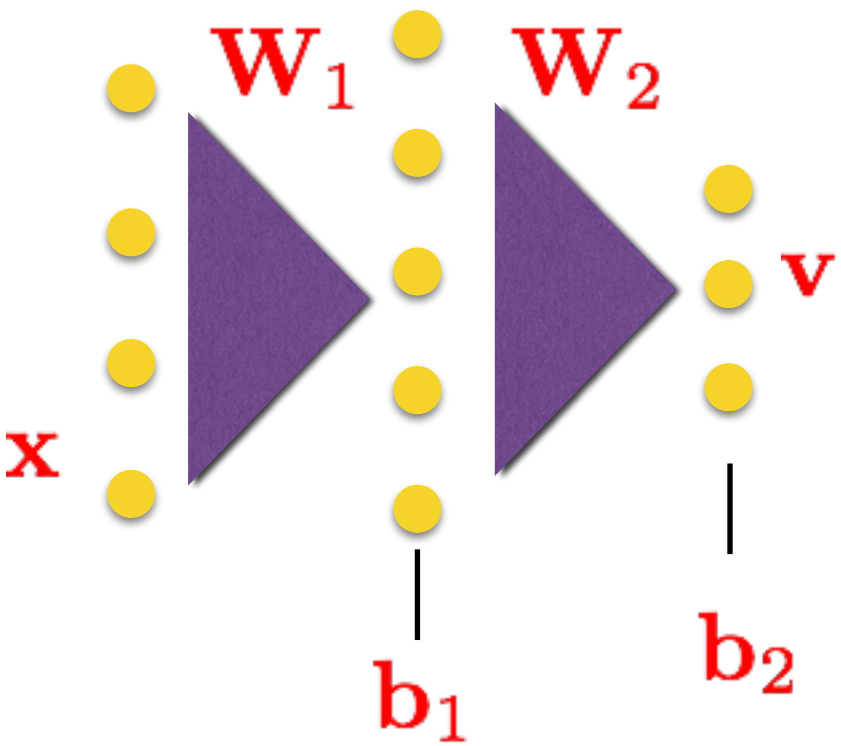
Cross Entropy

$$J_{CE} = -\sum_{i=1}^{M} \mathbf{y}_i^T \log(\mathbf{v}_i)$$

$\mathbf{y}_i$ are the desired outputs

Learning Posterior Probabilities with NNs

Neural networks predict posterior probabilities [Richard, 1991]



$$P(C_i|\mathbf{X}) = \frac{p(\mathbf{X}|C_i)p(C_i)}{p(\mathbf{X})}$$

When DNNs are trained with CE or MSE

$$\mathbf{v}(\mathbf{x}) = \mathcal{E}_{y|\mathbf{X}=\mathbf{x}}[\mathbf{y}]$$

Neural networks estimate conditional expectation of the desired targets given the input

When the targets are discrete classes $\mathbf{y} = [0\ 0\ ..0\ 1\ 0\ ..0]$
conditional expectation is the class posterior !

Richard, Michael D., and Richard P. Lippmann. "Neural network classifiers estimate Bayesian a posteriori probabilities." *Neural computation* 3.4 (1991): 461-483.

Learning Posterior Probabilities with NNs

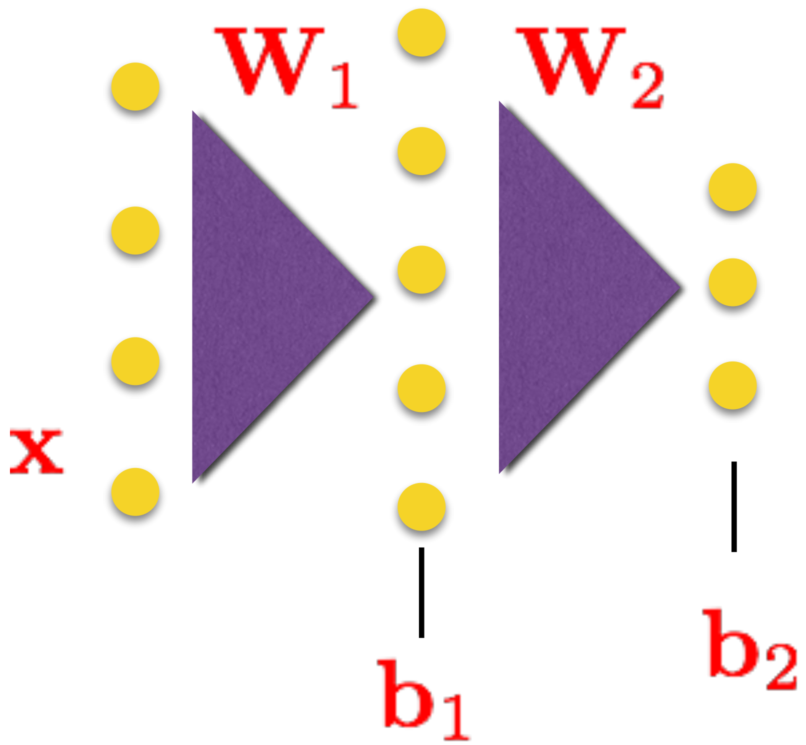
Choice of target function ψ

- Softmax function for classification

$$\psi(v_i) = \frac{e^{v_i}}{\sum_i e^{v_i}}$$

- Softmax produces positive values that sum to 1
- Allows the interpretation of outputs as posterior probabilities

Parameter Learning

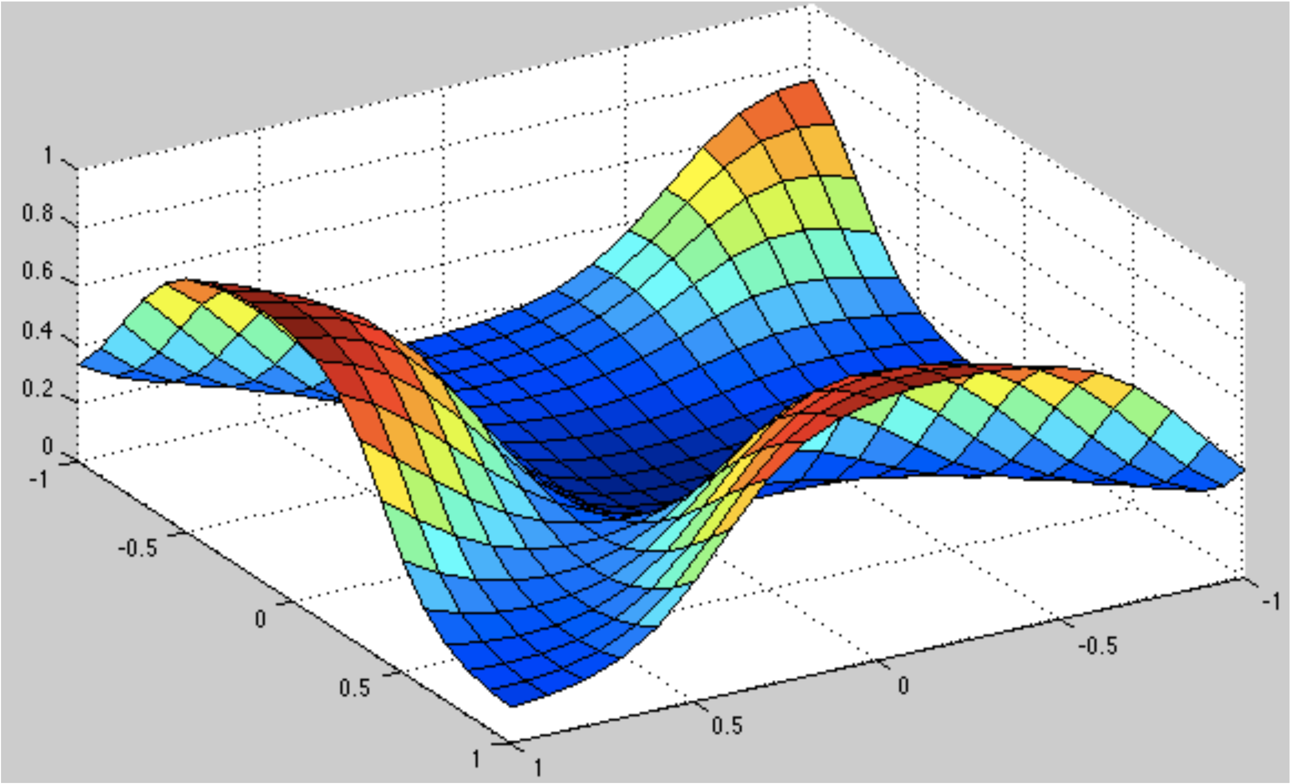


$$\mathbf{v}^2 = \psi\left(\mathbf{W}^2\phi(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) + \mathbf{b}^2\right)$$

Error function for entire data

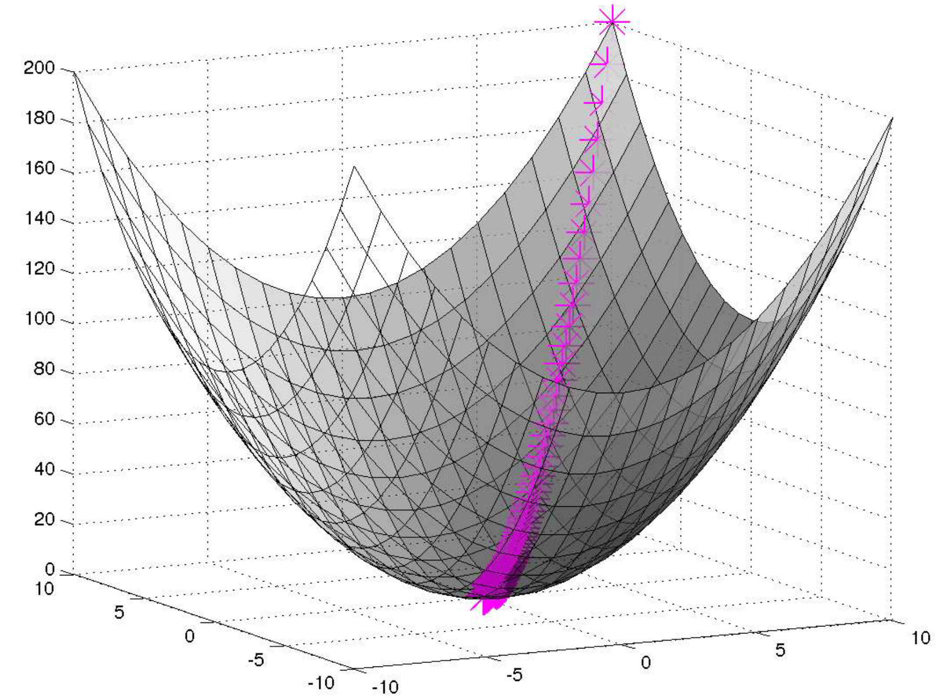
$$J_{MSE} = \sum_{i=1}^{M} ||\mathbf{v}_i - \mathbf{y}_i||^2$$

Typical Error Surface as a function of parameters (weights and biases)



Parameter Learning

Error surface close to a local optima



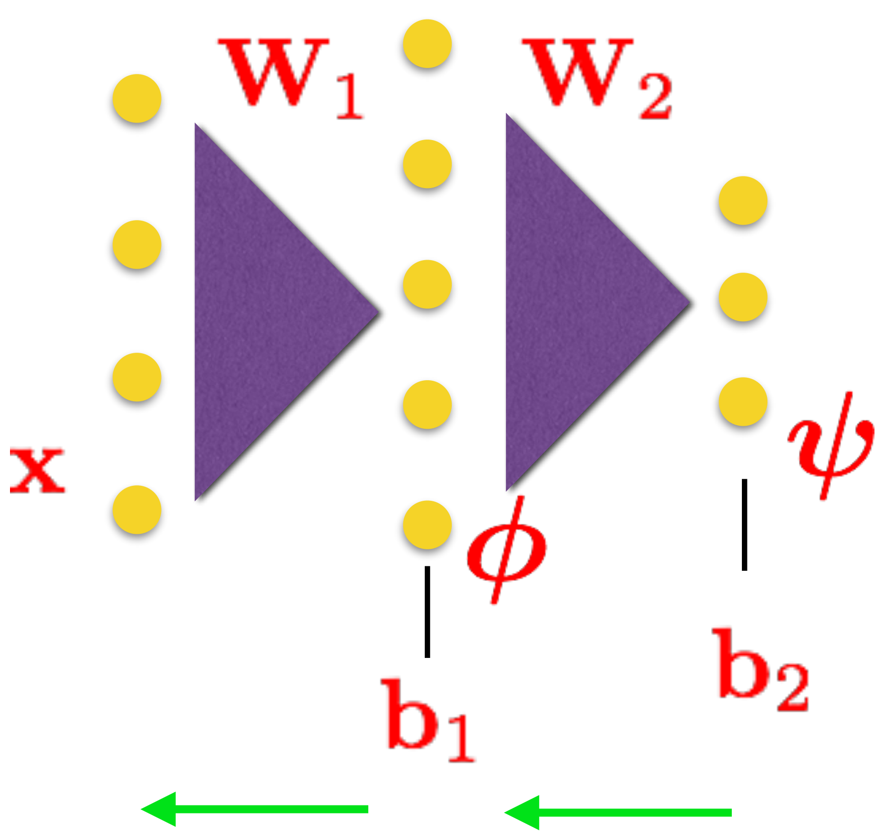
Non-linear nature of error function

- Move in the reverse direction of the gradient

$$\mathbf{W}_1^t = \mathbf{W}_1^{t-1} - \eta \frac{\partial J}{\partial \mathbf{W}_1}$$

Error back propagation

$$\frac{\partial J}{\partial \mathbf{W}_1} = \frac{\partial J}{\partial \psi} \times \frac{\partial \psi}{\partial \phi} \times \frac{\partial \phi}{\partial \mathbf{W}_1}$$



Summary so far…

- Neural networks as discriminative classifiers
- Need for hidden layer
- Choice of non-linearities and target functions
- Estimating posterior probabilities with NNs
- Parameter learning with back propagation.

Need For Deep Networks

Modeling complex real world data like speech, image, text

- Single hidden layer networks are too restrictive.
- Needs large number of units in the hidden layer and trained with large amounts of data.
- Not generalizable enough.

Networks with multiple hidden layers - deep networks (Open questions till 2005)

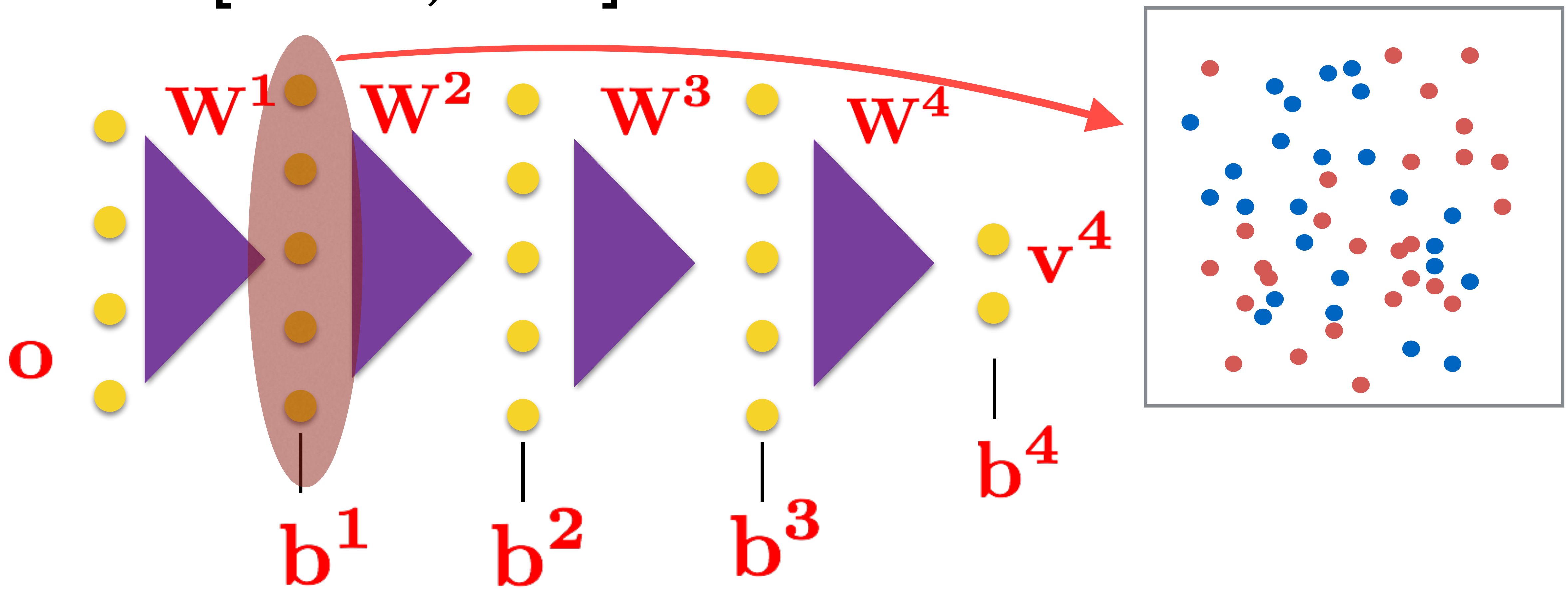
- Are these networks trainable ?
- How can we initialize such networks ?
- Will these generalize well or over train ?

Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

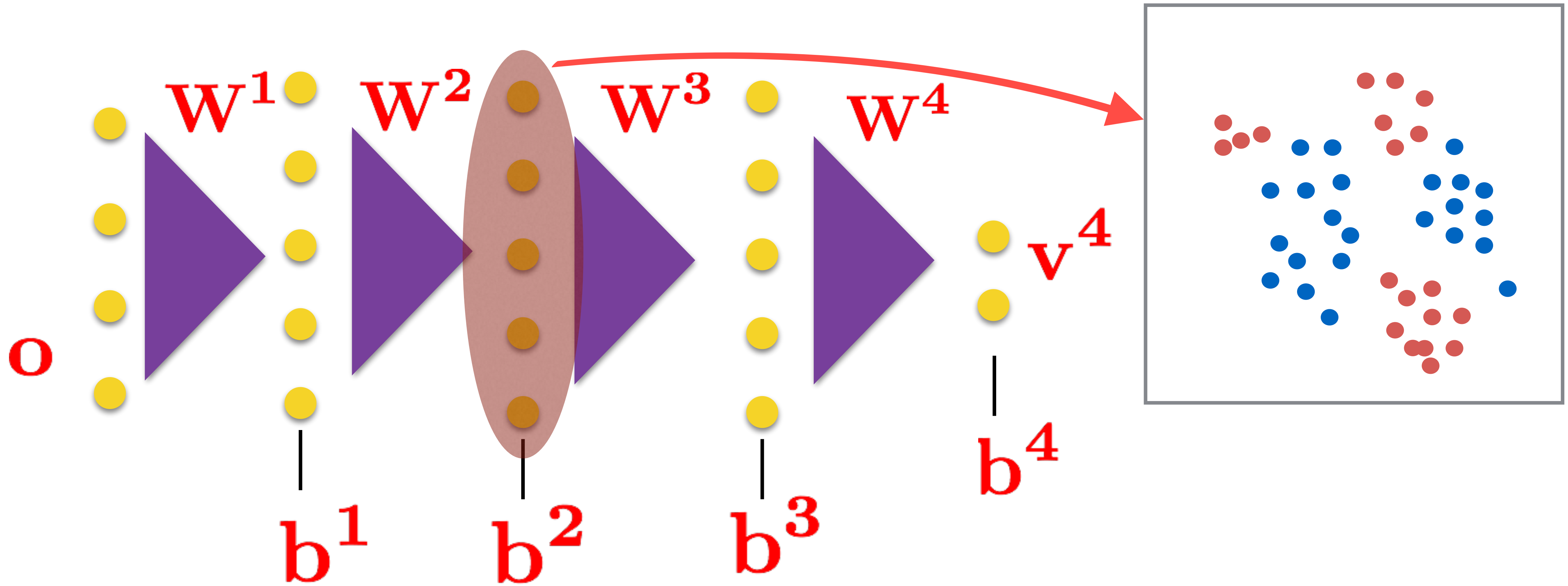
Deep Networks Intuition

Neural networks with multiple hidden layers - Deep networks [Hinton, 2006]



Deep Networks Intuition

Neural networks with multiple hidden layers - Deep networks

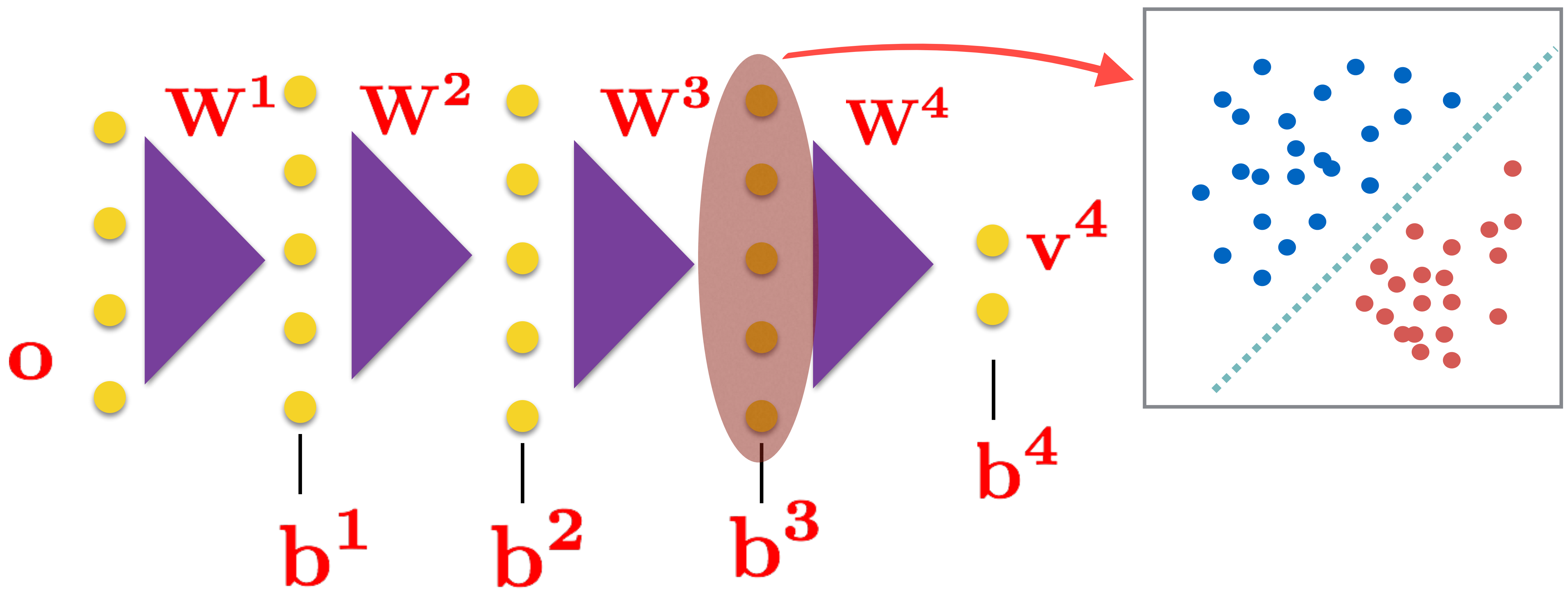


Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Deep Networks Intuition

Neural networks with multiple hidden layers - Deep networks



Deep networks perform hierarchical data abstractions which enable the non-linear separation of complex data samples.

Summary so far…

- Linear models to neural network.
- Deep Neural networks as extensions of NNs.
- Intuition behind behind multiple hidden layers

Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Deep Networks

- Will the networks generalize with deep networks
  - DNNs are quite data hungry and performance improves by increasing the data.
  - Generalization problem is tackled by providing training data from all possible conditions.
    - Many artificial data augmentation methods have been successfully deployed
  - Providing the state-of-art performance in several real world applications.

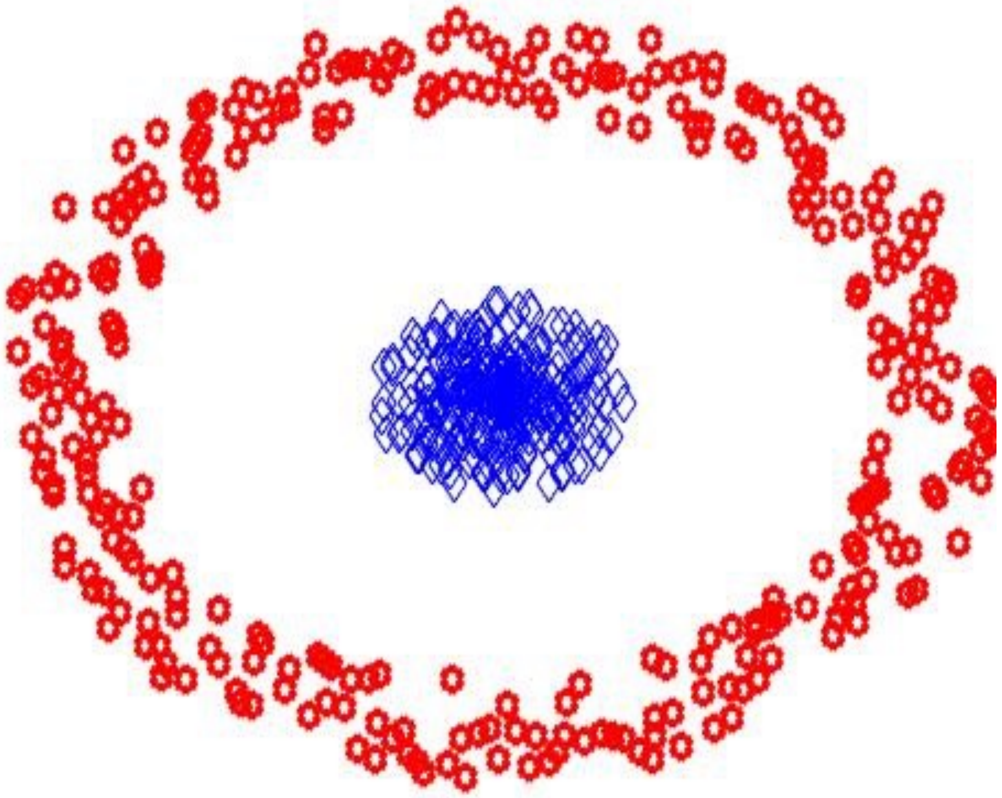
Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

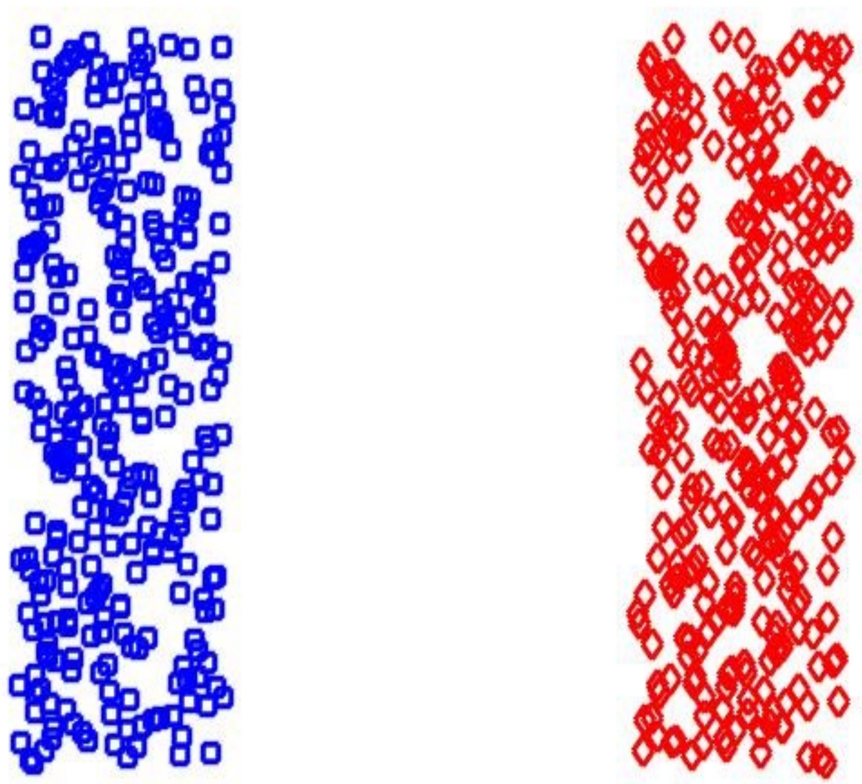
Representation Learning in Deep Networks

- The input data representation is one of most important components of any machine learning system.

Cartesian Coordinates



Polar Coordinates



Representation Learning in Deep Networks

- The input data representation is one of most important components of any machine learning system.
    - Extract factors that enable classification while suppressing factors which are susceptible to noise.

- Finding the right representation for real world applications - substantially challenging.
    - Deep learning solution - build complex representations from simpler representations.
    - The dependencies between these hierarchical representations are refined by the target.

Representation Learning in Deep Networks

[Zeiler, 2014]