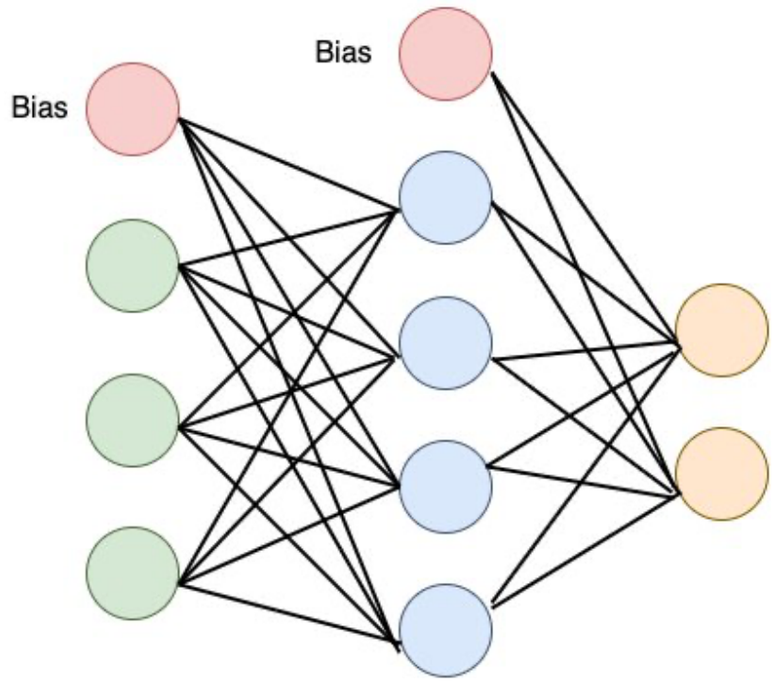
CS 4824/ECE 4424: Neural Networks I

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

Two-layer Feed-forward Network

- Architecture



- Hidden nodes: $z_j = h_1\left(\boldsymbol{w_j^{(1)T}}\, \boldsymbol{x}\right)$
- Output nodes: $y_k = h_2\left(\boldsymbol{w_k^{(2)T}}\, \boldsymbol{z}\right)$
- Overall: $y_k = h_2(\sum_j w_{kj}^{(2)}\, h_1(\sum_i w_{ji}^{(1)}\, x_i))$

Common Activation Functions h

- Identity $h(a) = a$
- Threshold $h(a) = \begin{cases} 1 & if\ a \geq 0 \\ 0 & if\ a < 0 \end{cases}$
- Sigmoid $h(a) = \sigma(a) = \frac{1}{1 + e^{-a}}$
- Gaussian $h(a) = e^{-\frac{1}{2}(\frac{a-\mu}{\sigma})^2}$
- Tanh $h(a) = tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$

Two-layer Feed-forward Network

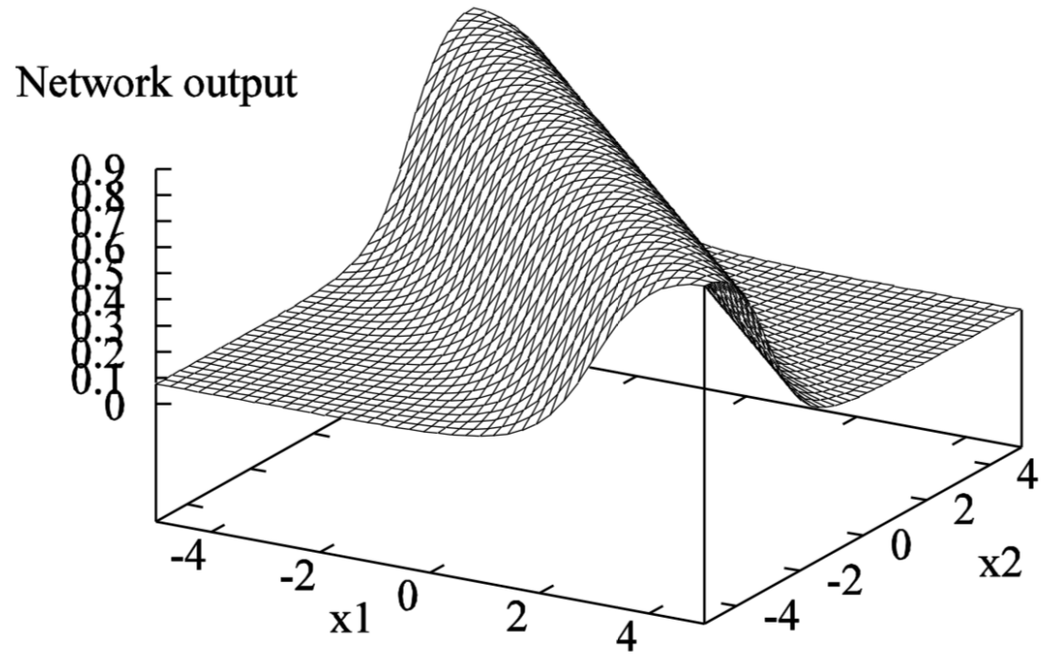
- Regression

- Classification

Combining Activation Functions

- Adding two sigmoid nodes with parallel but opposite "cliffs" produces a **ridge**

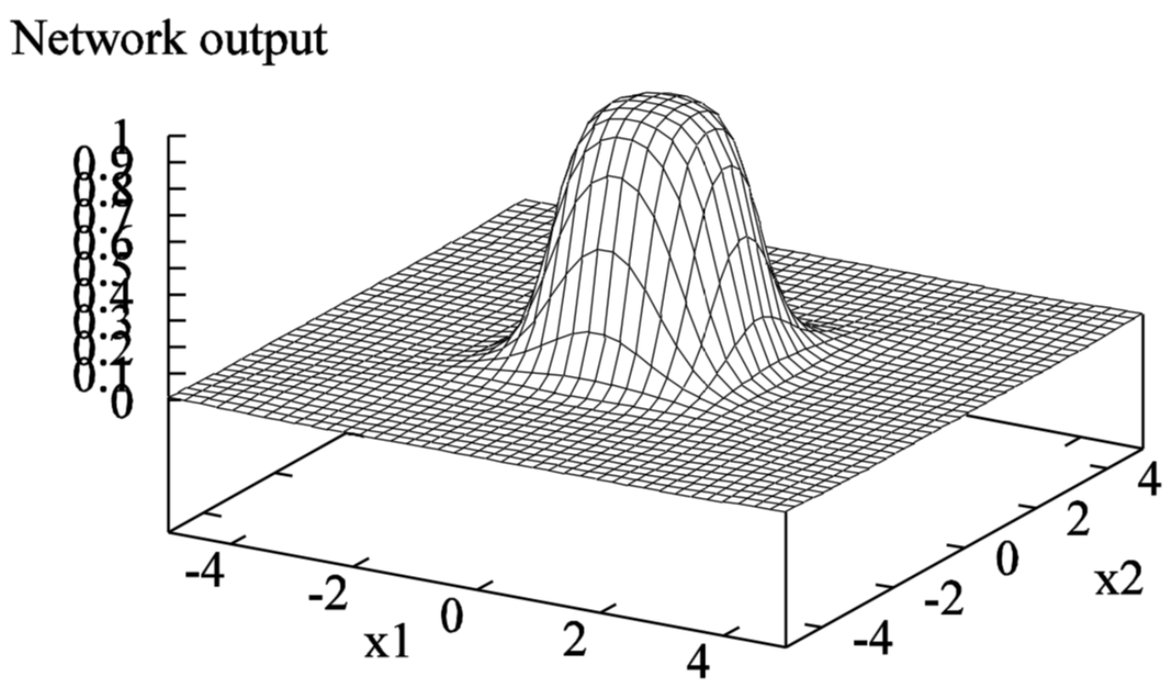
- Schematic



Combining Activation Functions

- Adding two intersecting ridges (and thresholding) produces a **bump**

- Schematic



Combining Activation Functions

- A bump can classify linearly non-separable data points

- By tiling bumps of various heights together, we can approximate any function

Combining Activation Functions

- Combining activation functions in a neural network enables us to approximate any function, hence millions of applications
  - Machine translation
  - Computer vision
  - Speech recognition
  - Word embedding
  - …