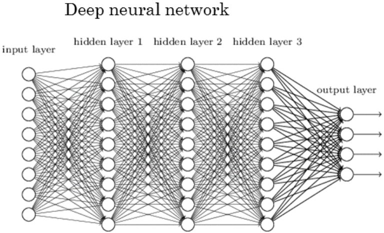
# Distributed Stochastic Gradient Descent

Kevin Yang and Michael Farrell

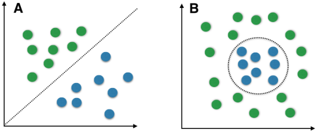
April 26, 2016

# Motivation - Deep Learning

- Deep-Learning
  - Objective: Learn a complicated, non-linear function that minimzes some loss function
- Why do we need deep models?
  - The class of linear functions is inadequate for many problems.

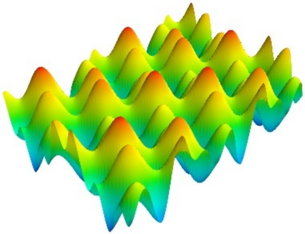


http://www.rsipvision.com/exploring-deep-learning/



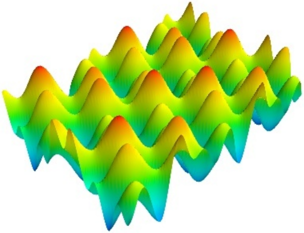
http://sebastianraschka.com/Articles/2014_naive_bayes_1.html

# Motivation - Deep Learning



- How do we learn these deep models?
  - Choose a random example
  - Run the neural network on the example
  - Adjust the parameters of the network such that our loss function is minimized more than it was before
  - Repeat
- Difficulties?
  - Local Minima
  - Non-convexity
  - Neural Networks can have millions or even billions of parameters

# Motivation - Deep Learning



- How do we learn these deep models?
  - Choose a random example
  - Run the neural network on the example
  - Adjust the parameters of the network such that our loss function is minimized more than it was before
  - Repeat
- Difficulties?
  - Local Minima
  - Non-convexity
  - Neural Networks can have millions or even billions of parameters

# Motivation - SGD

- How do we maximize our reward function?
  - One common technique is Stochastic Gradient Descent
  - **w** is the vector of parameters for the model
  - $\eta$ is the learning rate
  - $\mathbf{f}(\mathbf{w})$ is the loss function evaluated with the current parameters **w**
  - 
    ```
    $\mathbf{w} \leftarrow \mathbf{0}$
    while $\mathbf{f}(\mathbf{w})$ is not minimized do
        for $i = 1, n$ do
            $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla f(\mathbf{w})$
    ```
  - As the number of training examples, $n$, and the number of parameters, $|\mathbf{w}|$, increases, this algorithm quickly becomes very slow...

# Motivation - Distributed SGD

- Since some of these models take days/weeks/months to run, we would hope that we could use a distributed computing cluster parallelize this process.

# DistBelief

# TensorFlow

# gRPC

# Our Model

# Extensions