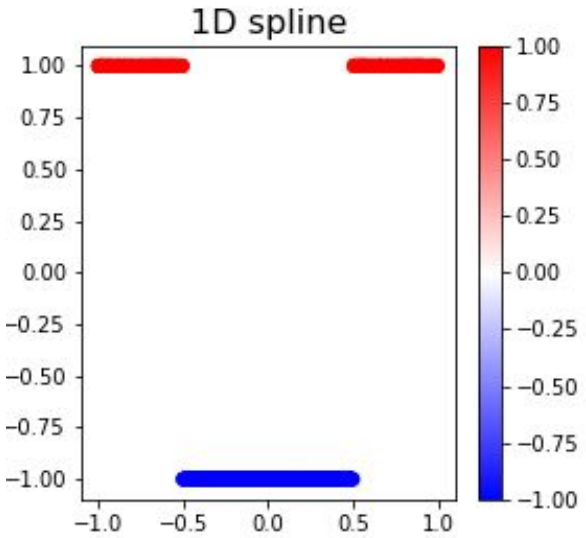
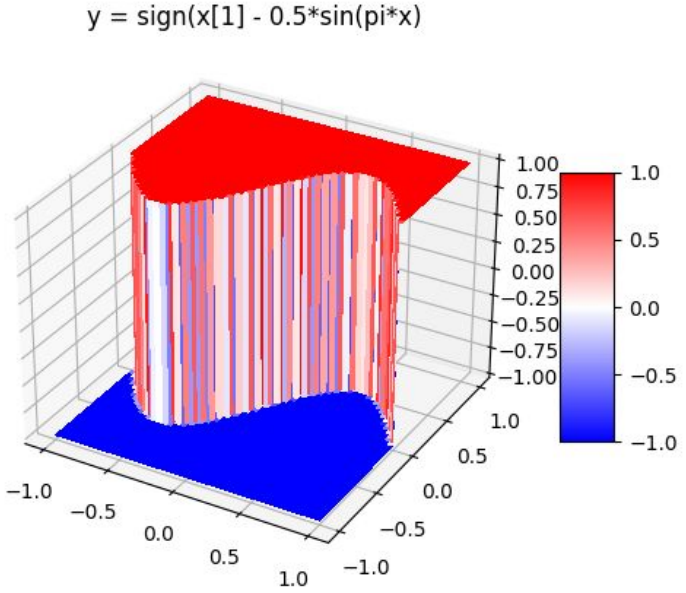
Neural Network Regularization

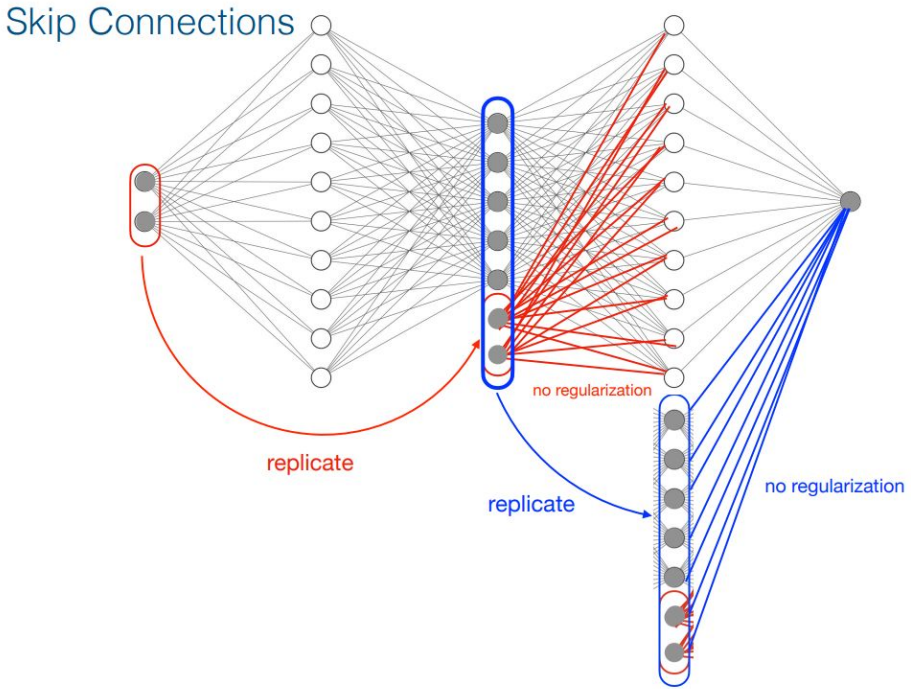
Dr. Robert Nowak, Rahul Parhi, Jack Wolf
University of Wisconsin-Madison

Data


1D spline
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0


y = sign(x[1] - 0.5*sin(pi*x)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Network Set-Up



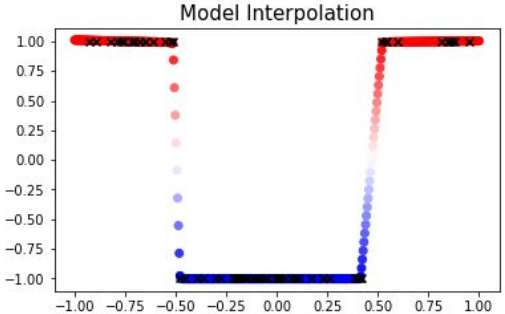
- 2 layer ReLU network of width k*n, where k is a large scalar and n is the number of training examples
- Single hidden linear layer of width closer to n
- Skip connections
- Trained with SGD and Adam
- Regularization applied to ReLU network (excluding skip connections)

Experiment

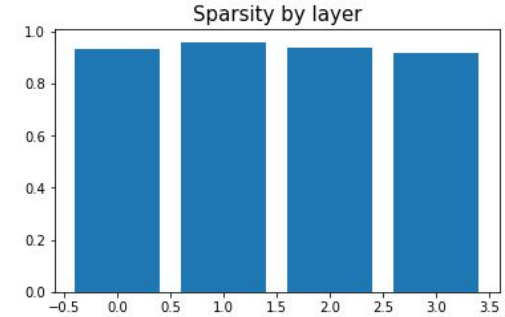
- Goal: use regularization techniques to learn sparse representation of data
- Definitions:
  - Active node: node whose weight has magnitude greater than a threshold value
  - Sparsity: percentage of non-active nodes in network
  - Threshold value defined per layer as 1e-3*max(abs(layer.weight))
- Theory:
  - Let R = width of ReLU layers in network and N = number of training examples
  - Model can learn training data with as little as (N/R)% active nodes

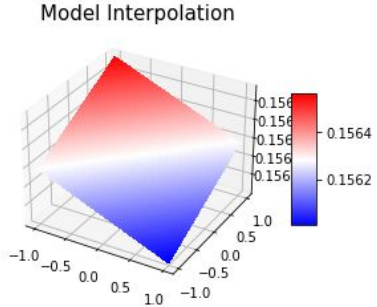
Experiment outcomes

- N = 64, R = 640 should lead to model layers that are 1-(64/640)=90% sparse
- Roughly achieved this result



Params:
-Training samples: 64
-ReLU dim: 640
-Linear dim: 10
-Training epochs: 100000
-Optimizer: Adam
-LR: 0.001
-WD: 1e-10 --> 1e-3
-Regularization scalar: 0.6
-Regularization method: 1
Results
-Prediction accuracy: 0.98
-Prediction mse: 0.05
-Sparsity: [0.93, 0.96, 0.94, 0.92]





Params:
-Training samples: 64
-ReLU dim: 640
-Linear dim: 10
-Training epochs: 100000
-Optimizer: Adam
-LR: 0.001
-WD: 1e-10 --> 1e-3
-Regularization scalar: 0.6
-Regularization method: 1
Results
-Prediction accuracy: 0.51
-Prediction mse: 1.02
-Sparsity: [0.88, 0.85, 0.89, 0.74]