A Theoretical Analysis on Feature Learning in Neural Networks: Emergence from Inputs and Advantage over Fixed Features

Zhenmei Shi*, Junyi Wei*, Yingyu Liang
University of Wisconsin-Madison

Motivation

• Hidden Layers: good representations of the inputs for prediction
• Neurons: correspond to interesting patterns in the inputs

Problem Setting

Dictionary Learning: input = sparse combination of base patterns

Network: 2-Layer, Hinge-loss, L2-regularizer, Gaussian init, Gradient descent
• Train a network: \( g(x) = \sum_{l=1}^{L} a_l \sigma(w_l x + b_l) \)
• Activation: truncated ReLU \( \sigma(x) = \min(1, \max(0, x)) \)

Network Learning Result

Theorem (informal)
For any \( \epsilon, \delta \in (0, 1), \) if
\[
k = \Omega\left(\log^{O(d)} \frac{d^2}{\epsilon} \right), \quad p_0 = \Omega \left( \frac{k}{d} \right), \quad m \geq \max\{D, \Omega \left( \frac{k^{1/3}}{\epsilon} \right) \}
\]
then with proper hyperparameters (e.g., step size), w.p. at least \( 1 - \delta \) we can get a network with error at most \( \epsilon. \)

Questions
• How features learned from inputs via gradient descent?
• Is learning features from inputs necessary for the superior performance?

Our results
• Propose a theoretical model of the data with input structure
• Prove network learning via gradient descent can succeed
• Prove fixed feature approaches fail
• Prove learning without input structure fails

Lower Bound for Fixed Feature Approach
• Fixed feature approach:
  1. Let \( \Psi(x) \in [-1, 1]^D \) be any data-independent \( D \)-dim feature mapping
  2. Linear model \( h(x) = \langle \Psi(x), \theta \rangle \) with bounded weight \( ||\theta||_2 \leq B \)

Theorem (informal)
There exist data distributions on which all such models \( h \) must have hinge-loss at least \( p_0 \left( 1 - \frac{\sqrt{2kB}}{2k} \right) \)

Lower Bound for Without Input Structure
• Without input structure: sample \( \Phi \) uniformly from \( (0,1)^D \)
• Statistical Query (SQ) algorithms:
  1. Ask statistical queries \( \langle Q, \ell \rangle \) about the data
  2. Receives an estimation of \( \Pr[Q(x,y) = \ell] \) within error \( \varepsilon \)

Theorem (informal)
For any SQ algorithm that can learn without the input structure to classification error less than \( \frac{1}{2} - \frac{1}{2} \left( \frac{D^{1/3}}{k} \right) \), either the number of queries or the number of classifiers \( 1 \) must be at least \( \frac{1}{2} \left( \frac{D^{1/3}}{k} \right) \).

Without input structure, all poly algo in the Statistical Query model (including networks and fixed features above) cannot achieve as small loss.

Feature Learning on Synthetic Data
• Visualization of the neuron weights (normalized to unit length)
• They clustered around \( \sum_{i \in A} M_i \) and \( -\sum_{i \in A} M_i \)

Feature Learning on MNIST(0/1)
• The neurons gradually form two clusters around ground-truth weights
• Show the emergence of the features in the neural networks
• However, in fixed feature approaches, there is no feature learning

Take Home Message

Input Structure ➔ Feature Learning ➔ Superior Performance