Sharper Guarantees for Learning Neural Network Classifiers with Gradient Methods

Hossein Taheri[1] Christos Thrampoulidis[2] Arya Mazumdar[1]

[1]UC San Diego [2]University of British Columbia

Introduction

Understanding generalization and convergence behavior of gradient methods for training neural networks (NN) is central to modern learning theory.

Two approaches for understanding NN:

Neural Tangent Kernel (NTK): The solution exists near the initialization as a result of large width and large initialization. Kernel regime requires large network width often increasing with data dimension.

Feature Learning(FL): The weights move a significant distance from initialization to learn the features. Initialization is usually small and much fewer neurons are required than the kernel regime.

Questions

1. How large can the **distance between the solution and initialization** be for a deep NN to operate in the kernel regime?
2. What is the **generalization rate** in the kernel regime?
3. How is the generalization guarantees improved in the feature learning regime?
4. Can networks of **constant width** (w.r.t. data dimension) provably obtain good generalization?

Main Theorems

Notation: The **L-layer network** model is $\Phi(w,x) := \frac{1}{\sqrt{m}} W_{L+1}^\top (\frac{1}{\sqrt{m}}\sigma(W_L^\top \cdots \frac{1}{\sqrt{m}}\sigma(W_1^\top x)\cdots)$. where σ is the activation function and each layer has **width m**. Concatenate all weight matrices W_i into the vector $w \in R^p$. Let the loss function to be logistic loss, denote the **sample size by n** and demote the train and test loss by $\widehat{F}(w), F(w)$.

Theorem 1 (Error rates and sufficient conditions for NTK)

Let all parameters of the network be initialized as i.i.d. standard Gaussian and assume the step-size η satisfies the condition of the descent lemma. Fix T and assume the target weights vector $w^\star \in \mathbb{R}^p$ that obtain small training loss such that $\|w^\star - w_0\|^2 \geq \eta T \widehat{F}(w^\star)$. Moreover, assume the width m is large enough such that it satisfies $\|w^\star - w_0\| \leq m^{O(1/L)}$. Then GD achieves with w.h.p. over initialization

$$\widehat{F}(w_T) \lesssim \frac{\|w^\star - w_0\|^2}{\eta T}, \quad \mathbb{E}_{\mathcal{S}}[F(w_T) - \widehat{F}(w_T)] \lesssim \frac{G_0^2\, \mathbb{E}_{\mathcal{S}}[\|w^\star - w_0\|^2]}{n}$$

.

- The condition $\|w^\star - w_0\|^2 \geq \eta T \widehat{F}(w^\star)$ ensures the existence of an interpolating solution $w^\star$ achieving small train loss.
- The other condition $\|w^\star - w_0\| \leq m^{O(1/L)}$ is a sufficient condition for learning an L layer network in the NTK regime. Previous works assumed $\|w^\star - w_0\| = O_m(1)$, whereas we showed that the **distance can grow with width**.
- Under these conditions, the test error after $T = n$ iterations is $\frac{\|w^\star - w_0\|^2}{n}$.

- To interpret the bound let's consider a stylized data model (namely **XOR dataset**)where $x_i \in \{\pm 1\}^d, y = x_1 \cdot x_2$.
- For the XOR data, Thm 1 implies the sample complexity $\mathbf{n} = \tilde{\mathbf{O}}(\mathbf{d^2})$ with $\mathbf{m} = \mathbf{\Omega}(\mathbf{poly}(\mathbf{d}))$ neurons.
- For this dataset, next theorem shows that with large step size, networks of **constant width** can achieve the **optimal sample complexity with a fast rate.**

Theorem 2 (Error rates and sufficient conditions for FL)

For 1-hidden layer nets of constant width(i.e., $m = \Omega_d(1)$) with quadratic activations, step size $\eta = m$, SGD with $n = \widetilde{O}(d)$ samples achieves w.h.p. perfect test accuracy on the XOR dataset in $\log(d)$ steps.

Experimental Results

MNIST and FashionMNIST: We test our generalization bound on the 2-hidden-layer net with Softplus activation and with logistic loss on the FashionMNIST and MNIST datasets.

XOR Classification: 1-hidden-layer network, quadratic activation, constant width $m = 16$, step-size $\eta = m$ shows fast convergence with only d samples as predicted by Theorem 2.

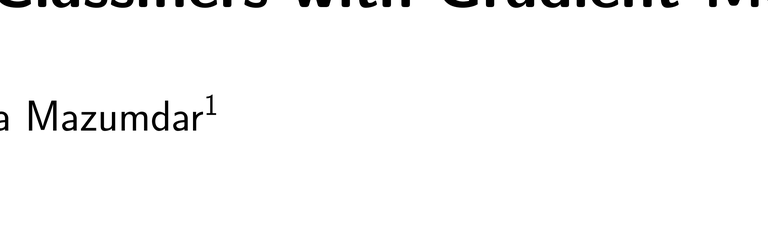

Figure: Theoretical and empirical generalization bounds in each iteration of GD for learning MNIST and FashionMNIST datasets with 2-hidden layer Neural Networks.

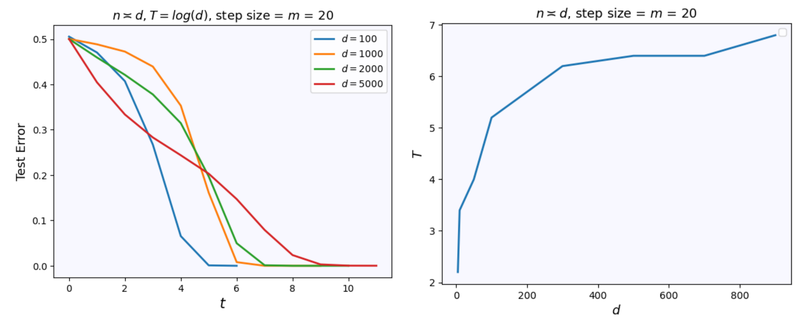

Figure: Left: Test accuracy vs. iterations on the XOR task with dimension d. Right: Total number of steps needed to reach near zero test accuracy vs data dimension for XOR dataset.