# Generalization and Convergence of Gradient Descent in Interpolating Neural Networks

Hossein Taheri

March 2025

# Overview

- Preliminaries and Motivation
- Training loss convergence
- Generalization error bounds
- Results under NTK
- Proof Sketch

[1] H.T, and C. Thrampoulidis "Generalization and Stability of Interpolating Neural Networks with Minimal Width." JMLR 2024.
[2] H.T, C. Thrampoulidis and A. Mazumdar "Sharper Guarantees for Learning Neural Network Classifiers with Gradient Methods" ICLR 2025.

# Background

- Over-parameterized neural networks are renowned for their ability to memorize datasets, often achieving near-zero training loss.
- Despite the complexity of wide nets, they also demonstrate good generalization performance.

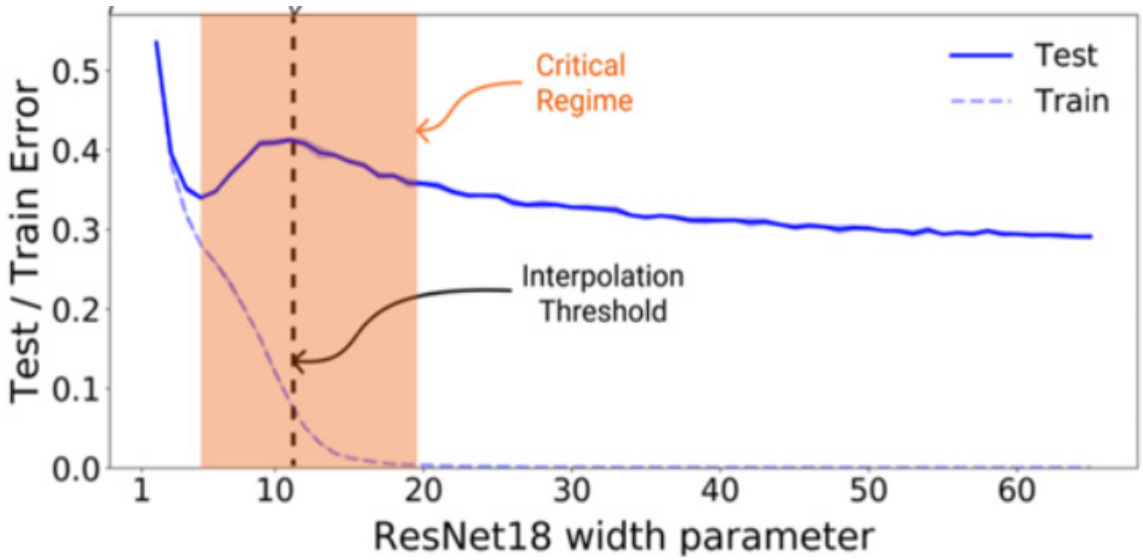



Figure: Train/Test error based on model size for CIFAR10 dataset [Nakirran et al 2021].

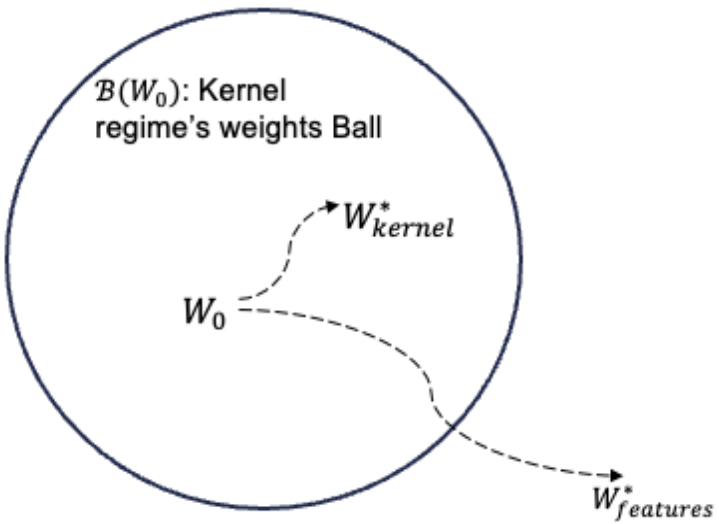
# Motivation

1. How many neurons are needed to ensure convergence of Gradient Descent (GD) in the interpolation regime?
2. What is the convergence behavior of GD under this over-parameterization condition?
3. How do the solutions found by GD generalize to unseen data?

# Neural networks' theory

Two theories for understanding NN optimization:

1. **Neural Tangent Kernel (NTK):**
   weights stay close to initialization, almost convex landscape
2. **Feature Learning:**
   weights move a significant distance to learn the ground truth, highly non-convex landscape



# Preliminaries

- One hidden-layer Neural Network:

$$\Phi(w, x) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma\left(\langle w_j, x \rangle\right),$$

  - $x$ : d-dimensional data
  - $w_j \in \mathbb{R}^d$ : First layer weights (trainable)
  - $a_j = \pm 1$ : Second layer weights (fixed)
  - $m$ : Network width
  - $\sigma$ : Smooth and Lipschitz non-linearity, e.g., GeLU, Softplus,...

# Preliminaries

- Unregularized Empirical Risk Minimization:

$$\min_{w \in \mathbb{R}^{md}} \left\{ \widehat{F}(w) = \frac{1}{n} \sum_{i=1}^{n} f\left(y_i \Phi\left(w, x_i\right)\right) \right\}.$$

- $y_i = \pm 1$ : labels
- $f : \mathbb{R} \to \mathbb{R}^+$ is a loss function with the following properties:
  - Monotonically decreasing
  - Convex, Lipschitz and Smooth
  - Self-bounded i.e., $|f'(t)| \leq c\, f(t)$

Examples: Logistic loss, Exponential loss, Polynomial loss ...

# First theorem: Training Loss

**Theorem (Training loss)**

*Fix the training horizon $T \geq 1$ and step-size $\eta$. Assume the target wights $w^* \in \mathbb{R}^{md}$ to satisfy $\|w^* - w_0\|^2 \geq \eta T \widehat{F}(w^*)$ and the hidden-layer width ($m$) to satisfy $m \gtrsim \|w^* - w_0\|^4$. Then, the training loss satisfies*

$$\widehat{F}(w_T) \lesssim \frac{\|w^* - w_0\|^2}{\eta T}.$$

- if there exists good target vector that achieves small loss while being close enough to initialization, then the convergence of GD is guaranteed.

# Generalization Gap

**Theorem (Generalization gap)**

*Fix the time horizon $T \geq 1$ and step size $\eta$. Let any $w^* \in \mathbb{R}^{md}$ such that $\|w^* - w_0\|^2 \geq \eta T \widehat{F}(w^*)$. Suppose hidden-layer width $m$ satisfies $m \gtrsim \|w^* - w_0\|^4$. Then, the generalization gap of GD at iteration $T$ is bounded as*

$$\mathbb{E}_{\mathcal{S}}\Big[F(w_T) - \widehat{F}(w_T)\Big] \lesssim \frac{1}{n}\sum_{t=1}^{T} \widehat{F}(w_t)$$
$$\lesssim \frac{1}{n}\,\mathbb{E}_{\mathcal{S}}\left[\|w^* - w_0\|^2\right].$$

- The conditions are similar to the previous theorem and shows that if $T \approx n$, the test loss decays with rate $\|w^* - w_0\|^2/n$.

# Remarks

- Despite non-convexity both training and generalization rates look remarkably same as convex case (linear models).
- This presumes a realizability condition that there exists good target vector that achieves small loss while being close enough to initialization.
- These conditions and the final results lead to an exponential improvement in the width condition.

- When can we ensure the realizability conditions of our theorems on a good target $w^\star$ are satisfied?
- These conditions subsume a Neural Tangent Kernel (NTK) separability condition.

# Results under NTK

**Assumption (Separability by NTK (Nitanda et al. 2019))**

*For almost surely all $n$ training samples from the data distribution there exists $w^\star \in \mathbb{R}^{md}$ and $\gamma > 0$ such that $\|w^\star\| = 1$ and for all $i \in [n]$, it holds $y_i \left\langle \nabla_w \Phi(w_0, x_i), w^\star \right\rangle \geq \gamma$.*

**Corollary (Results under NTK-separability)**

*Let the assumption above hold and assume logistic loss. Suppose $m \gtrsim \frac{\log^4(T)}{\gamma^4}$ for a fixed training horizon $T$. Then,*

$$\widehat{F}(w_T) \lesssim \frac{\log^2(T)}{\gamma^2 \eta T},$$

$$\mathbb{E}_{\mathcal{S}}\left[F(w_T) - \widehat{F}(w_T)\right] \lesssim \frac{\log^2(T)}{\gamma^2 n}.$$

# Example: XOR data

Consider d-dimensional XOR data:

$$x = (x^1, x^2, \cdots, x^d), \forall k : x^k \in \{\frac{1}{\sqrt{d}}, \frac{-1}{\sqrt{d}}\}$$

$$y = \mathrm{sign}(x^1 \cdot x^2).$$

**Lemma (XOR Margin)**

*Consider the noisy XOR data $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$. Assume Gaussian initialization $w_0 \in \mathbb{R}^{md}$ with entries iid $N(0,1)$. If $m \gtrsim d^3$, then with high probability, the NTK-separability Assumption is satisfied with margin $\gamma \gtrsim 1/d$.*

- As a result, for XOR data if $m \gtrsim d^4$ we have $\widehat{F}(w_T) = \tilde{O}(\frac{d^2}{T})$ and $\mathbb{E}_{\mathcal{S}}[F(w_T) - \widehat{F}(w_T)] = \tilde{O}(\frac{d^2}{n})$.

# Comparison to prior works

- For **quadratic loss** and smooth activations several works e.g. [Oymak and Soltanolkotabi 2020] derived polynomial conditions on the network width for convergence in the NTK regime where $m = \Omega(n^2)$.

- Through a stability-based proof [Hardt et al. 2016] derived a generalization bound $O(\frac{T^{\beta c/(\beta c+1)}}{n})$ for general $\beta$-smooth and non-convex objectives, but this required a **time-decaying step-size** $\eta_t \leq c/t$.

# Comparison to prior works

- For logistic loss [Richards and Rabbat 2020] utilized a stability argument for neural networks to derive the generalization rate of order $1/\sqrt{n}$ given the width condition $\mathbf{m} = \mathbf{\Omega}(\mathbf{n^2})$.

- We showed improved rate $1/n$ with **fewer neurons** $\mathbf{m} = \mathbf{\Omega}(\mathbf{log^4(n)})$.

# Comparison to prior works

- For logistic loss and ReLU activations, [Ji and Telgarsky 2020] showed that if $m = \tilde{\Omega}(1/\gamma^8)$ the test error is bounded by $\frac{1}{\gamma^2\sqrt{n}}$ where $\gamma$ is the NTK margin.
- Our results showed that by utilizing the Hessian information the width condition can be improved to $m = \tilde{\Omega}(1/\gamma^4)$.
- Utilizing the Hessian information also allowed using a new algorithmic stability argument which led to tighter generalization bounds.

# Proof sketch of convergence

## Self-bounded Weak Convexity

**Proposition:** The NN objective function is self-bounded weakly convex i.e., there exists constant $\kappa > 0$ such that for all $w$,

$$\lambda_{min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\kappa\, \widehat{F}(w)\,.$$

Let $w_{\alpha t}$ be some point in the line between $w_t$ and $w^*$. Then,

$$\begin{aligned}\widehat{F}(w^*) &\geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle + \frac{1}{2}\,\lambda_{min}(\nabla^2 \widehat{F}(w_{\alpha t}))\left\|w^* - w_t\right\|^2 \\ &\geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle - \frac{1}{2}\,\kappa\, \widehat{F}(w_{\alpha t})\left\|w^* - w_t\right\|^2.\end{aligned}$$

# Proof sketch of convergence

For two-layer NN it can be shown that $\kappa \propto 1/\sqrt{m}$.

$$\widehat{F}(w^*) \geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle - \frac{1}{2\sqrt{m}} \widehat{F}(w_{\alpha t}) \left\| w^* - w_t \right\|^2$$

Therefore,

$$\widehat{F}(w^*) \geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle - \frac{1}{2\sqrt{m}} \max_{w \in [w_t, w^*]} \widehat{F}(w) \left\| w^* - w_t \right\|^2$$

## Generalized Local Quasi-Convexity

Fix arbitrary $w_1, w_2 \in \mathbb{R}^{md}$, any constant $\lambda > 1$, and $m$ large enough such that $\sqrt{m} \gtrsim \lambda \|w_1 - w_2\|^2$. Then, the 2-layer NN objective satisfies

$$\max_{w \in [w_1, w_2]} \widehat{F}(w) \leq \frac{1}{1 - 1/\lambda} \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}.$$

# Proof sketch of generalization

- For the generalization error, it suffices to bound the leave-one-out error:

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq 2G\,\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|w_T - w_T^{\neg i}\|\Big].$$

- $w_T^{\neg i}$ is the $T$-th iteration of GD on the leave-one-out loss $\widehat{F}^{\neg i}(w) := \frac{1}{n}\sum_{j\neq i}\widehat{F}_j(w)$.

- G is the Lipschitz parameter of objective.

# Proof sketch of generalization

So we had,

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq 2G\,\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|w_T - w_T^{\neg i}\|\Big].$$

If $\sqrt{m} \gtrsim \|w_t - w_t^{\neg i}\|^2$, the leave-one-out error term is computed recursively through

$$\begin{aligned}
&\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| \\
&\leq \left\|\left(w_t - \eta\nabla\widehat{F}^{\neg i}(w_t)\right) - \left(w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i})\right)\right\| + \frac{\eta}{n}\widehat{F}_i(w_t)\,. \\
&\leq \left(1 + \frac{2\eta}{\sqrt{m}}\max\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i})\}\right)\|w_t - w_t^{\neg i}\| + \frac{\eta}{n}\widehat{F}_i(w_t)
\end{aligned}$$

# Proof sketch of generalization

So far we derived

$$\left\| w_{t+1} - w_{t+1}^{\neg i} \right\|$$
$$\leq \left( 1 + \frac{2\eta}{\sqrt{m}} \max\{ \widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i}) \} \right) \left\| w_t - w_t^{\neg i} \right\| + \frac{\eta}{n} \widehat{F}_i(w_t)$$

Recall that we are interested in the averaged leave-one-out error:

$$\mathbb{E}\Big[ F(w_T) - \widehat{F}(w_T) \Big] \lesssim \mathbb{E}\Big[ \frac{1}{n} \sum_{i=1}^{n} \| w_T - w_T^{\neg i} \| \Big].$$

By taking the average over $i \in [n]$ and replacing $\widehat{F}^{\neg i}(w_t^{\neg i})$ and $\widehat{F}^{\neg i}(w_t)$ with our optimization guarantees :

$$\frac{1}{n} \sum_{i=1}^{n} \left\| w_T - w_T^{\neg i} \right\| \leq \frac{\eta}{n} \sum_{t=0}^{T-1} \widehat{F}(w_t)\,,$$

# Proof sketch of generalization

Therefore, the generalization gap can be bounded based on the average training loss performance:

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \lesssim \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|w_T - w_T^{\neg i}\|\Big]$$
$$\leq \frac{\eta}{n}\sum_{t=0}^{T-1}\widehat{F}(w_t).$$

Further replacing $\widehat{F}(w_t)$ with the optimization guarantees, concludes the proof.

# Extensions to Deep Networks

**Theorem (Test loss for deep nets under NTK)**

*Assume separability by the model's neural tangent by margin $\gamma$. Let the width be large enough such that $m \gtrsim (\log(n)/\gamma)^{6L+4}$, where $L$ is the number of hidden layers. Then after $T = O(n)$ iterations, the test loss is bounded as,*

$$\mathbb{E}[F(w_T)] = \widetilde{O}\left(\frac{e^{O(L)}}{\gamma^2 n} + \frac{1}{\gamma^2 T}\right).$$

- The width condition is still poly-logarithmic in $n$, although there is an exponential dependence on depth.
- Prior work by [Chen et al 2021] derived the rate $\widetilde{O}(e^{O(L)}\sqrt{\frac{m}{\gamma^2 n}} + \frac{1}{\gamma^2 T})$ under a similar width condition.

# Conclusions

- We provided improved bounds for the test loss and width condition of interpolating neural nets.
- We demonstrated an exponential improvement on the width condition for the stability of NN.
- The resulting generalization bounds also improve upon well-established methods such as uniform convergence.