# Thesis Proposal

# Training and Generalization of Overparametrized Neural Networks

Shreyas Kalvankar

October 3, 2025

## 1 Introduction

Overparametrized neural networks (models with more parameters than training samples) are capable of perfectly interpolating data, yet they often generalize surprisingly well. The *Neural Tangent Kernel* (NTK) framework (Jacot et al., 2020; Lee et al., 2020) provides a rigorous description of infinitely wide networks: in this regime, training is equivalent to kernel regression with a fixed kernel, so optimization and generalization can be analyzed with classical kernel methods. However, this “lazy training” limit rules out feature learning, since the kernel does not evolve during training.

Recent results show that in realistic finite-width and finite-depth networks, the NTK is random at initialization and evolves non-trivially during training (Hanin and Nica, 2019). This suggests the existence of a “weak feature learning regime,” where kernel evolution remains small but non-negligible, interpolating between lazy training and full representation learning.

This thesis will develop an analysis of the training dynamics under an evolving NTK, quantify how kernel drift scales with the depth($d$)/width($n$) ratio $\beta = d/n$, and interpret the induced sequence of RKHSs $(\mathcal{H}_t)_{t\geq 0}$ to understand the regimes where weak feature learning arises and how it affects generalization.

## 2 Research Question

### Main research question:

> How can we characterize the evolution of the NTK in finite-width and finite-depth neural networks, and what does this imply for effective function spaces and generalization?

### Sub-questions

1. How can the NTK evolution be written as $K_t = K_0 + \Delta K_t$, and how does this perturbation affect the gradient flow dynamics relative to $K_0$?

2. How does the size of $\Delta K_t$ scale with $\beta = d/n$? Can we identify thresholds for when $\Delta K_t$ significantly alters training dynamics?

3. How can we describe the evolving RKHSs $(\mathcal{H}_t)$ induced by $K_t$? Do these spaces enlarge compared to $\mathcal{H}_0$ and allow approximation of functions outside $\mathcal{H}_0$?

4. How does movement between RKHSs affect the inductive bias, e.g. via eigenvalue decay of $K_t$ and the minimal-norm interpolant in $\mathcal{H}_t$?

# 3 Background

## 3.1 Overparametrized Neural Networks

Overparametrized models often interpolate the training data yet generalize well. This paradoxical behavior connects to the implicit bias of gradient-based optimization, whereby gradient descent converges to specific low-complexity solutions among many possible interpolants (Soudry et al., 2024; Gunasekar et al., 2017). Phenomena such as double descent and benign overfitting underscore the need for new analytical frameworks.

## 3.2 Neural Tangent Kernel

The NTK provides one such framework: by linearizing a network at initialization, one defines a kernel that describes parameter coupling during training (Jacot et al., 2020). In the infinite-width limit, this kernel is deterministic and fixed, reducing training to kernel regression (Lee et al., 2020). While elegant, this "lazy training" regime excludes feature learning, since the feature map is frozen at initialization.

## 3.3 Finite Depth/Width Corrections

Hanin and Nica (2019) show that when depth and width co-scale, the NTK is stochastic and evolves over training. They identify $\beta = d/n$ as a key scaling parameter: the NTK variance grows like $\exp(c\beta)$, and even the first SGD update changes the kernel. This suggests a weak feature learning regime $0 < \beta \ll 1$ where kernel drift is non-negligible but stable.

## 3.4 Function Spaces and RKHS Perspective

Every positive definite kernel $K$ defines a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, consisting of functions with an inner product structure induced by $K$. In the NTK framework, the infinite-width limit corresponds to kernel regression in the RKHS $\mathcal{H}_0$ induced by the initialization kernel (Jacot et al., 2020; Bartolucci et al., 2023). This space encodes the inductive bias of lazy training: solutions are minimum-norm interpolants in $\mathcal{H}_0$. When the kernel evolves during training, the effective hypothesis space can be viewed as a sequence of RKHSs $(\mathcal{H}_t)_{t\geq 0}$. Studying how these spaces change provides an interpretation of weak feature learning.

# 4 Research Gap

Current theory establishes that the NTK is fixed at infinite width and evolves when depth and width co-scale. However, several gaps remain:

- No formal perturbation analysis connecting kernel drift to deviations in training dynamics.
- Limited characterization of the boundary between lazy and weak feature learning regimes in terms of $\beta$.
- Lack of interpretation of evolving NTKs as a path of RKHSs.

# 5 Scope

The work will be analytical in nature, focusing on:

- Perturbation expansions of the training dynamics under evolving $K_t$.
- Bounds on kernel drift as a function of depth/width scaling.
- Analysis of the RKHS sequence $(\mathcal{H}_t)_{t\geq 0}$ induced by $K_t$.

# 6 Methodology

## 6.1 Perturbation Analysis of Kernel Dynamics

Write $K_t = K_0 + \Delta K_t$, with $K_0$ the lazy NTK. Substitute into the gradient flow ODE and develop a perturbation expansion. Analyze conditions on $\Delta K_t$ (e.g. scaling in $\beta$) under which corrections are small or large.

## 6.2 Scaling Regimes

The aim here is to understand how the depth/width ratio $\beta = d/n$ marks transitions between lazy, weak feature learning, and unstable regimes. This involves considering how measures of kernel drift depend on $\beta$, in light of theoretical results such as Hanin–Nica's exponential scaling.

## 6.3 Functional Analysis Perspective

We view training as generating a family of RKHSs $\{\mathcal{H}_t\}_{t\geq 0}$, where each $\mathcal{H}_t$ is induced by the kernel $K_t$. This provides a function-space lens on kernel evolution where changes in $K_t$ correspond to changes in the associated norm and hypothesis space. The question is then can we use this to understand generalization?

# 7 Expected Outcomes

- Perturbation-theoretic description of NTK evolution and its effect on training dynamics.
- Theoretical characterization of scaling regimes in terms of $\beta$ (lazy, weak feature learning, unstable).
- Functional analysis interpretation of evolving NTKs as a family of RKHSs.

# 8 Tentative Timeline (partial)

- **Weeks 1–2**
  Refine understanding of NTK theory and finite-width corrections. Implement a small-scale JAX setup to reproduce key diagnostics from Hanin and Nica (2019)

  1. Verify the explosion of the normalized second moment of $K_N(x, x)$ at initialization, scaling like $\exp(5\beta)$ with $\beta = d/n$.
  2. Measure the expected first-step kernel update $\Delta K$, confirming its growth with $\beta$ as predicted by their theory.

- **Weeks 3–6**
  Develop the perturbative description $K_t = K_0 + \Delta K_t$. Derive preliminary bounds or conditions on $\Delta K_t$. Set up some experiments to test the bounds.

- **Weeks 7–10**
  Analyze how $\Delta K_t$ depends on $\beta = d/n$ using existing results (e.g., Hanin–Nica). Try to characterize regime boundaries (lazy, weak feature learning, unstable).

- **Weeks 11–14**
  Frame $K_t$ as inducing a family of RKHSs $(\mathcal{H}_t)_{t\geq 0}$. Explore preliminary results on how $\mathcal{H}_t$ compares to $\mathcal{H}_0$ in terms of approximation or inductive bias.

## References

Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, 2023. doi: 10.1016/j.acha.2022.08.006. URL `https://www.sciencedirect.com/science/article/pii/S1063520322000768`.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization, 2017. URL `https://arxiv.org/abs/1705.09280`.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel, 2019. URL `https://arxiv.org/abs/1909.05989`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL `https://arxiv.org/abs/1806.07572`.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent $*$. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc62b. URL `http://dx.doi.org/10.1088/1742-5468/abc62b`.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data, 2024. URL `https://arxiv.org/abs/1710.10345`.