Training and generalization in overparameterized neural networks

Stage 1 review meeting for literature organization

Shreyas Kalvankar

December 16, 2025

TU Delft

1. General Overview of Draft

Introduction & Difficulties for establishing theory

- Neural network training involves high-dimensional, non-convex objectives.
- Loss landscapes are characterized by saddle points and complex critical structures.
- **Goal:** To bridge the gap between empirical success and theoretical understanding via simplified models and asymptotic limits.

Linear Networks

- *One-layer:* Gradient descent is a linear dynamical system governed by the data covariance spectrum. $f_w(x) = x^\top w, \quad x, w \in \mathbb{R}^d$
- Parameter dynamics known, can be characterized and analysed:

$$w_t = \left(I - \left(I - \eta X X^\top\right)^t\right) X^+ y$$

1. General Overview of Draft

Linear Networks

- *Deep Linear:* Introducing depth creates complex dynamics (coupled ODEs of order three) with no known general analytic solutions. $f_{v,W}(x) := v^\top Wx \qquad W \in \mathbb{R}^{m\times d}$, $v \in \mathbb{R}^m$, m is width of network.

$$\dot{v}_t = -\nabla_v \mathcal{L}(v_t, W_t) = W_t X\big(y - X^\top W_t^\top v_t\big),$$

$$\dot{W}_t = -\nabla_W \mathcal{L}(v_t, W_t) = v_t\big(y - X^\top W_t^\top v_t\big)^\top X^\top.$$

Non-Linear Networks

- Activation functions mix matrix products with elementwise operations, we don't know how to handle these kinds of expressions.

2a. Related Work: The NTK Regime

Linearization at Infinite Width

- **Motivation:** To analyze training dynamics by linearizing the network around initialization.
- **Formalization:**
  - As width $m \to \infty$, the empirical kernel Θ_t converges to a deterministic, static limit $\bar{\Theta}$.

$$\bar{\Theta}(x, x') = \mathbb{E}_w\big[\varphi(w^\top x)\,\varphi(w^\top x')\big] + \sigma_v^2\, x^\top x'\, \mathbb{E}_w\big[\varphi'(w^\top x)\,\varphi'(w^\top x')\big]$$

  - The network function f_t evolves linearly with respect to this frozen kernel.
- **Training Dynamics:**
  - Equivalent to Kernel Ridge Regression with the NTK.
  - Convergence is guaranteed if the limit kernel is positive definite.

2b. Related Work: The Mean-Field View

Feature Learning at Infinite Width

- **Motivation:** To capture feature learning, which is absent in the "lazy" NTK regime.
- **Formalization:**
  - Weights are treated as particles drawn from a distribution μ.
  - Output is scaled by $1/n$ (vs $1/\sqrt{n}$ in NTK).

$$f_{v,w}(x) = \frac{1}{m}\sum_{\alpha=1}^{m} v_\alpha \varphi(w_\alpha^\top x) = \int_\Omega v\, \varphi(w^\top x)\, d\mu(v,w)$$

- **Training Dynamics:**
  - Modeled as a Wasserstein gradient flow of the probability measure μ_t.
  - Allows the kernel (and features) to evolve during training.

2c. Related Work: Spectral Bias

Frequency-Dependent Convergence

- **Phenomenon:** Neural networks fit low-frequency components of the target function faster than high-frequency noise.
- **Theoretical Basis:**
  - Convergence along eigen-directions is determined by the eigenvalues λ_k of the NTK integral operator.
  - High frequencies correspond to small eigenvalues $\rightarrow$ slow convergence.
- **Implications:** Provides a theoretical basis for early stopping and generalization in overparameterized models.

3. Preliminary Experiments

1. **Function Space Convergence**
2. **Kernel Drift & Regime Transitions**
3. **Spectral Analysis of Empirical NTK**

Summary & Discussion

Structure of Introduction & Literature Review:

1. **Foundations:** From linear regression to deep linear dynamics.
2. **Infinite Limits:** NTK limit (static kernel), Mean-Field limit (changing kernel).
3. **Spectral Properties:** How eigenvalues dictate learnability (Spectral Bias).

Question for Supervisors: Does this structure logically support the move to finite-width deviations in later chapters?