Neural Tangent Kernel: High-Dimensional Probability View

Hooman Zolfaghari[1], Abdollah Zohrabi[1], and Amirreza Velae[1]

Sharif University of Technology, Tehran, Iran

Abstract. The Neural Tangent Kernel (NTK) has emerged as a pivotal theoretical concept for understanding the training dynamics and generalization properties of Neural Networks. This proposal outlines a research agenda aimed at extending NTK theory to encompass advanced neural network architectures, broader data distributions. By addressing current limitations and exploring novel applications, this research seeks to deepen the theoretical foundations of deep learning and enhance the practical utility of NTK in guiding the design and optimization of modern neural networks.

Keywords: Neural Tangent Kernel · Deep Learning Thoery · High-Dimensional Probability

1 Introduction and Definitions

The rapid advancement of deep learning has revolutionized various domains, like computer vision and natural language processing. Central to understanding the efficacy of Deep Neural Networks (DNN) is the **Neural Tangent Kernel (NTK)** is a theoretical framework which captures the behavior of fully-connected deep nets in the infinite-width limit trained by gradient descent. [5]

Introduced by Jacot et al. (2018), NTK connects infinite-width neural networks with kernel methods. In this regime, the network's training dynamics under gradient descent become equivalent to those of a kernel machine with the NTK as its kernel function. Consequently, NTK offers profound insights into NNs convergence properties and generalization capabilities. By connecting neural networks with kernel methods, NTK leverages the extensive study and well-established theory of kernels, thereby making the analysis of neural networks more accessible and grounded in a rich mathematical foundation.

As deep learning models become more complex and widely used, it is essential to expand NTK's theory to cover practical and different network structures. This proposal outlines our research goals to analys NTK theory using High-Dimensional Probability concepts, address its current limitations, and explore its possible extensions, improvements and applications in modern neural network designs.

1.1 Preliminaries

Consider a fully connected neural network $f(x;\theta)$ with parameters θ, trained using gradient descent. For a dataset $\{(x_i, y_i)\}_{i=1}^n$, the Neural Tangent Kernel (NTK) is defined as:

$$\Theta(x, x') = \nabla_\theta f(x;\theta)^\top \nabla_\theta f(x';\theta),$$

where $\nabla_\theta f(x;\theta)$ is the gradient of the network's output with respect to its parameters. The NTK measures how changes in the parameters affect the output similarity between inputs x and x'.

In the infinite-width limit, the training dynamics of the neural network under gradient descent can be described as:

$$\frac{d}{dt} f(x;\theta(t)) = -\eta \sum_{i=1}^{n} \Theta(x, x_i)\big(f(x_i;\theta(t)) - y_i\big),$$

where η is the learning rate.

In this regime, the NTK remains nearly constant during training, leading to a simplified, linearized dynamics for $f(x;\theta)$.

2 Related Works

Pointwise Convergence and CNTK: In [6] Aurora et al, give the first non-asymptotic proof showing that a fully-trained sufficiently wide net is indeed equivalent to the kernel regression predictor using NTK.

Theorem (Convergence to the NTK at initialization). Fix $\epsilon > 0$ and $\delta \in (0, 1)$. Suppose

$$\sigma(z) = \max(0, z) \quad \text{and} \quad \min_{h \in [L]} d_h \geq \Omega\left(\frac{L^6}{\epsilon^4} \log\left(\frac{L}{\delta}\right)\right).$$

Then for any inputs $x, x' \in \mathbb{R}^{d_0}$ such that $\|x\| \leq 1, \|x'\| \leq 1$, with probability at least $1 - \delta$, we have:

$$\left| \left\langle \frac{\partial f(\theta, x)}{\partial \theta}, \frac{\partial f(\theta, x')}{\partial \theta} \right\rangle - \Theta^{(L)}(x, x') \right| \leq (L + 1)\epsilon.$$

They also give an efficient algorithm for exact computation of this kernel on Convolutional NNs, the CNTK.

Generalization Bound: In [4], they provide a generalization bound using NTK. Assuming: Fix failure probability $\delta \in (0,1)$ data $S = \{(x_i, y_i)\}_{i=1}^n$ Distribution D is $(\lambda_0, \delta/3, n)$-non-degenerate. $\kappa = \mathcal{O}\left(\frac{\lambda_0 \delta}{n}\right)$. Width $m \geq \kappa^{-2}\text{poly}(n, \lambda_0^{-1}, \delta^{-1})$. Loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,1]$ is 1-Lipschitz in the first argument such that $\ell(y,y) = 0$. Gradient descent runs for $k \geq \Omega\left(\frac{1}{\eta\lambda_0} \log \frac{n}{\delta}\right)$ iterations.

Then with probability at least $1-\delta$ over the random initialization and the training samples,The two-layer neural network $f_{\mathbf{W}(k),\mathbf{a}}$ has population loss $L_D(f_{\mathbf{W}(k),\mathbf{a}}) = \mathbb{E}_{(\mathbf{x},y)\sim D}[\ell(f_{\mathbf{W}(k),\mathbf{a}}(\mathbf{x}), y)]$, bounded as:

$$L_D(f_{\mathbf{W}(k),\mathbf{a}}) \leq \sqrt{\frac{2\mathbf{y}^\top (\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + O\left(\sqrt{\frac{\log \frac{n}{\lambda_0 \delta}}{n}}\right).$$

Finite Width and Spectral Bias: In [7] They provide quantitative bounds measuring the L_2 difference in function space between the trajectory of a finite-width network trained on finitely many samples from the idealized kernel dynamics of infinite width and infinite data. They apply the result and find that eigenfunctions of the NTK integral operator are learned at rates corresponding to their eigenvalues. They demonstrate that the network will inherit the bias of the kernel at the beginning of training even when the width only grows linearly with the number of samples

Uniform Convergence of Neural Tangent Kernel and Streaming Data: In [1], they create a uniform convergence bound of the Neural Tangent Kernel (NTK) and analyze the convergence of stochastic gradient descent (SGD) in the streaming data setting.

- With random initialization, the NTK converges to a deterministic function **uniformly** over the input space for all layers as the number of neurons tends to infinity.
- Using this uniform convergence, it was further proven that the prediction error of multilayer neural networks under SGD converges in expectation in the streaming data setting.

Under Gaussian initialization, for $m \geq Cd^2 \exp(L^2)$ (for some constant C), there exist constants C_1, C_2, and C_3 such that, with probability at least $1 - \exp(-C_1 m^{1/3})$,

$$\left\| H^{(\ell)} - \Phi^{(\ell)} \right\|_\infty \leq C_2 \left(\frac{C_3^L}{m^{1/6}} + \sqrt{\frac{dL \log m}{m}} \right), \quad \forall 1 \leq \ell \leq L.$$

Here, $H^{(\ell)}$ represents the empirical NTK matrix at layer ℓ, and $\Phi^{(\ell)}$ is its deterministic counterpart. This result establishes a high-probability bound on the uniform convergence of the NTK, with the error decaying as the number of neurons m increases.

Sensitivity Analysis: The sensitivity of a model f_θ is defined as: $\mathcal{S}_{f_\theta}(z) = \mathbb{E}_{z,\hat{z},\theta}\left|\nabla_z f(z,\theta)^\top (z - \hat{z})\right|$..

For the NTK, the sensitivity is given by: [2]

$$\mathcal{S}_{\text{NTK}}(z) = \|z\|_2 \left\| \nabla_z \Phi_{\text{NTK}}(z)^\top \Phi_{\text{NTK}}^\top K_{\text{NTK}}^{-1} Y \right\|_2.$$

This simplifies to: $\mathcal{S}_{\text{NTK}}(z) = \mathcal{O}\left(\log k \sqrt{\frac{ND}{p}}\right) \sim \mathcal{O}(1)$. The results have certain assumptions whose necessity remains an open problem.

Spurious Data Memorization: Spurious data refers to irrelevant features memorized by the model. For data samples. Spurious memorization impacts generalization. Denoting *feature data alignment* as $\mathcal{F}_\varphi(z, z_1) = \frac{\varphi(z)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\| P_{\Phi_{-1}}^\perp \varphi(z_1) \right\|_2^2}$. and *model stability* as $S_{z_1}(z) = \mathcal{F}_\varphi(z, z_1) S_{z_1}(z_1)$., the difference in NTK feature alignment for spurious data satisfies: [3]

$$\left| \mathcal{F}_{\text{NTK}}(z_1^s, z_1) - \gamma_{\text{NTK}} \right| = o(1),$$

where:

$$\gamma_{\text{NTK}} = \alpha \frac{\sum_{l=1}^\infty \mu_l'^2 \alpha^l}{\sum_{l=1}^\infty \mu_l'^2}, \quad 0 < \gamma_{\text{NTK}} < 1.$$

Here, μ_l' denotes the l-th Hermite coefficient of the derivative of the activation function φ'. This result is applied to generalization bounds on the NTK model.

3 Our Goal

Problem Definition: While NTK provides foundational insights into neural network training dynamics in the infinite-width limit, several challenges persist. Current NTK results are limited to Gaussian or bounded data distributions and specific activation functions, restricting their broader applicability. Additionally, NTK theory primarily addresses fully-connected networks, necessitating extensions to modern architectures like Transformers and CNNs. Furthermore, the influence of NTK on Sample Complexity, Feature Selection, and settings other than Supervised Learning remains inadequately understood.

Building on existing NTK research, our study aims to investigate these topics, leveraging High-Dimensional Probability as the foundational framework to achieve our objectives:

- **Assess Sample Complexity and Model Capacity:**
  - Study explicit sample complexity, VC-Dimension, and Rademacher Complexity in NTK models.
- **Broaden NTK to Diverse Data Distributions:**
  - Extend current results depending on Gaussian/bounded to sub-Gaussian/sub-Exponential distributions.
  - Examine how data distribution affects NTK convergence and generalization.
- **Investigate Feature Selection Properties of NTK:**
  - Analyze NTK's spectral distribution for feature learning and bias.
  - Examine how NTK models memorize spurious data and factors influencing this behavior.
  - Develop methods to reduce overfitting to irrelevant data within the NTK framework.
  - Conduct sensitivity analyses to assess model robustness to input changes.
- **Expand NTK to Advanced Architectures:** For example Transformers and Kolmogorov-Arnold Networks
- Extend NTK to Unsupervised/Self-Supervised/Transfer Learning settings.

References

1. Arora, S., Du, S.S., Hu, W., Li, Z., Wang, R.: Overparametrized multi-layer neural networks: Uniform concentration of neural tangent kernel and convergence of stochastic gradient descent. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019), https://arxiv.org/abs/1904.11955
2. Bombari, S., Kiyani, S., Mondelli, M.: Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 2738–2776. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/bombari23a.html
3. Bombari, S., Mondelli, M.: How spurious features are memorized: Precise analysis for random and NTK features. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=o6N1Bqay0k
4. Cao, Y., Gu, Q.: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML) **97**, 1047–1055 (2019), https://arxiv.org/abs/1905.11661
5. Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. Advances in Neural Information Processing Systems (NeurIPS) **31** (2018), https://arxiv.org/abs/1806.07572
6. Lee, J., Xiao, L., Schoenholz, S.S., Bahri, Y., Novak, R., Sohl-Dickstein, J., Pennington, J.: On exact computation with an infinitely wide neural net. Advances in Neural Information Processing Systems (NeurIPS) **31** (2018), https://arxiv.org/abs/1902.04760
7. Rahaman, N., Arpit, D., Draxler, F., Lin, M., Hamprecht, F.A., Bengio, Y., Courville, A.: Spectral bias outside the training set for deep networks in the kernel regime. International Conference on Learning Representations (ICLR) (2019), https://arxiv.org/abs/1906.00719