## Literature Review

#### Kailong Wang

#### 1 Parameters

•  $p_i$ : Number of Parameters

• m: Number of Samples

### 2 Naming

• HDP: High Dimensional Probability

• NTK: Neural Tangent Kernel

• MLP: Multi-Layer Perceptron, a.k.a. Fully Connected (FC) Neural Network

• ResNet: Residual Neural Network, MLP with skip connections

#### 3 Convergence and Generalization of Wide Neural Network

Paper	Key Result	Condition Setup
[1]	Convergence in Polynomial Time with Polynomial Size of Data	Two or Three Layer MLP with
		SGD
[2]	ReLU networks is not Lipschitz but holds weaker Hölder smooth-	CNN or MLP with ReLU activa-
	ness	tion
[3]	Tighter and less conditioned bound.	2 layer MLP under less over-
		parameterized condition, ReLU,
		SGD
[4]	Analyze CNN with NTK and an algorithm designed for CNN	MLP, CNN, ReLU, SGD
	inspired by NTK.	

## 4 Convergence and Generalization of Deep Neural Network

Paper	Key Result	Condition Setup
[Hayoua, 5,	NTK does not work under finite width infinity depth NN archi-	MLP, ResNet, ReLU, SGD
6, 7]	tecture. However, NTK does work under ResNet architecture.	
	Initialization matters more for deep network.	
[8]	An ordinary differential equations point of view	CNN, ReLU, SGD
[9]		MLP, CNN, ReLU, SGD

# 5 Convergence and Generalization of Wide Deep Neural Network

Paper	Key Result	Condition Setup
[10]	Triple Descent Phenomenon	Single Layer MLP, $p_1 \ll m \ll$
		$p_2 \ll m^2 \ll p_3$
[11, 12]	When both depth and width are finite, the convergence and	
	generalization is determined by the ratio of $p/m$	

## 6 Spectral Analysis of NTK

Paper	Key Result	Condition Setup
[13]	The Spectrum of ResNTK shows stable frequencies while FC-NTK	The eigenfunctions of ResNTK
	has spike frequencies	are (scaled) spherical harmonics
		and that its eigenvalues decay
		with frequency $k$ at the rate of
		$k^{-p}$ .
[14]	NTK fail to learn high spectral modes unless the sample size $p$ is	
	sufficiently large	
[15]	NTK fail to filter out high frequency noise and thus fail to re-	MLP, CNN, ReLU, SGD
	construct the signal, while CNN shows robust performance under	
	such a case.	
[16, 17]	RFF, NTK and Neural Value Approximation	one line code to help RFF and
		NTK capture high frequency fea-
		ture
[18]	FIM, eigenspace, etc	

A common agreement is that the NTK fail to capture high frequency feature at the initial setup.

# 7 Researches inspired by NTK

Paper	Key Result	Condition Setup
[19, 20]	A second order optimization method to train NN inspired by NTK	
[21]	In the transfer learning setup, the model convergence does not	
	depends on input dimension but the feature dimension.	
[22]	Using NTK analyzes the generalization bound of robust optimiza-	
	tion	
[23, 24, 25,	NTK enables Gaussian Process and Bayesian Inference in design-	contextual bandit algorithm, re-
26]	ing Neural Network based bandit algorithm.	inforcement learning
[27]	How the intermediate parameter looks like after train the Neural	
	Network with NTK.	
[28]	Explains the feature representation capability with NTK.	MLP, ReLU, SGD
[29]	Explain the superior performance of ResNet over MLP by analyz-	MLP, ResNet, ReLU, SGD
	ing NTK.	
[30]	NTK under $L_2$ regularizer.	MLP, ReLU, SGD
[31]	NTK for Domain Adaptation.	
[32]	Reinforcement with NTK and Gaussian Process	
[33, 34]	NTK is similar to the Laplace Kernel	

# 8 Other Kernel

Paper	Key Result	Condition Setup
[35]	A new type of random features as a kernel function	
[36]	Composition Kernel	
[37]	Another sampling based kernel function	

# 9 Proofs of NTK

Paper	Key Result	Condition Setup
[38]	The first proof, a function space perspective	
[39]	Reverse Proof of NTK	
[40]	Parameter Space Perspective	
[41]	Random Kernel Function converges to NTK in expectation with	MLP, ReLU, SGD
	high Probability.	
[42, 43, 44]	Under infinity width architecture, NN's parameters almost re-	
	maining unchanging during training. This is called lazy training.	

#### References

- [1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. "Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers". In: ().
- [2] Alberto Bietti and Julien Mairal. "On the Inductive Bias of Neural Tangent Kernels". In: ().
- [3] Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. "Gradient Descent Can Learn Less Over-parameterized Two-layer Neural Networks on Classification Problems". In: ().
- [4] Sanjeev Arora et al. On Exact Computation with an Infinitely Wide Neural Net. Nov. 4, 2019. arXiv: 1904.11955 [cs, stat]. URL: http://arxiv.org/abs/1904.11955 (visited on 08/14/2023). preprint.
- [5] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. "The Curse of Depth in Kernel Regime". In: ().
- [6] Etai Littwin, Tomer Galanti, and Lior Wolf. On Random Kernels of Residual Architectures. June 17, 2020. arXiv: 2001.10460 [cs, stat]. URL: http://arxiv.org/abs/2001.10460 (visited on 08/17/2023). preprint.
- [7] Kaixuan Huang et al. Why Do Deep Residual Networks Generalize Better than Deep Feedforward Networks? A Neural Tangent Kernel Perspective. Dec. 22, 2020. arXiv: 2002.06262 [cs, stat]. URL: http://arxiv.org/abs/2002.06262 (visited on 08/17/2023). preprint.
- [8] Jiaoyang Huang and Horng-Tzer Yau. "Dynamics of Deep Neural Networks and Neural Tangent Hierarchy". In: ().
- [9] Jongmin Lee et al. "Neural Tangent Kernel Analysis of Deep Narrow Neural Networks". In: ().
- [10] Ben Adlam and Jeffrey Pennington. "The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization". In: ().
- [11] Boris Hanin and Mihai Nica. "FINITE DEPTH AND WIDTH CORRECTIONS TO THE NEURAL TANGENT KERNEL". In: (2020).
- [12] Yuan Cao and Quanquan Gu. "Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks". In: ().
- [13] Yuval Belfer et al. "Spectral Analysis of the Neural Tangent Kernel for Deep Residual Networks". In: ().
- [14] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. "Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks". In: ().
- [15] Stefani Karp et al. "Local Signal Adaptivity: Provable Feature Learning in Neural Networks Beyond Kernels". In: ().
- [16] Matthew Tancik et al. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. June 18, 2020. arXiv: 2006.10739 [cs]. URL: http://arxiv.org/abs/2006.10739 (visited on 11/11/2022). preprint.
- [17] Ling Yang et al. Diffusion Models: A Comprehensive Survey of Methods and Applications. Oct. 23, 2022. arXiv: 2209.00796 [cs]. URL: http://arxiv.org/abs/2209.00796 (visited on 11/11/2022). preprint.
- [18] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks. Sept. 27, 2020. arXiv: 1910.05992 [cond-mat, stat]. URL: http://arxiv.org/abs/1910.05992 (visited on 07/08/2023). preprint.
- [19] Tianle Cai et al. "Gram-Gauss-Newton Method: Learning Overparameterized Neural Networks for Regression Problems". In: ().
- [20] Arthur Jacot, Franck Gabriel, and Clement Hongler. "The Asymptotic Spectrum of the Hessian of DNN throughout Training". In: (2020).
- [21] Alex Damian, Jason D Lee, and Mahdi Soltanolkotabi. "Neural Networks Can Learn Representations with Gradient Descent". In: ().
- [22] Zhun Deng, Kenji Kawaguchi, and Jiaoyang Huang. "How Shrinking Gradient Noise Helps the Performance of Neural Networks". In: ().

- [23] Michal Lisicki, Arash Afkanpour, and Graham W Taylor. "EMPIRICAL ANALYSIS OF REPRESENTATION LEARNING AND EXPLORATION IN NEURAL KERNEL BANDITS". In: ().
- [24] Dongruo Zhou, Lihong Li, and Quanquan Gu. "Neural Contextual Bandits with UCB-based Exploration". In: ().
- [25] Parnian Kassraie and Andreas Krause. Neural Contextual Bandits without Regret. Feb. 28, 2022. arXiv: 2107.03144 [cs, stat]. URL: http://arxiv.org/abs/2107.03144 (visited on 08/17/2023). preprint.
- [26] Yiling Jia et al. Learning Neural Contextual Bandits Through Perturbed Rewards. Mar. 18, 2022. arXiv: 2201.09910 [cs]. URL: http://arxiv.org/abs/2201.09910 (visited on 08/17/2023). preprint.
- [27] Philip M Long. "Properties of the After Kernel". In: ().
- [28] Yizhang Lou et al. "Feature Learning and Signal Propagation in Deep Neural Networks". In: ().
- [29] Tom Tirer. "Kernel-Based Smoothness Analysis of Residual Networks". In: ().
- [30] Colin Wei et al. "Regularization Matters: Generalization and Optimization of Neural Nets v.s. Their Induced Kernel". In: ().
- [31] Jun Wu et al. "Distribution-Informed Neural Networks for Domain Adaptation Regression". In: ().
- [32] Imene R. Goumiri, Benjamin W. Priest, and Michael D. Schneider. "Reinforcement Learning via Gaussian Processes with Neural Network Dual Kernels". In: 2020 IEEE Conference on Games (CoG). 2020 IEEE Conference on Games (CoG). Osaka, Japan: IEEE, Aug. 2020, pp. 1–8. ISBN: 978-1-72814-533-4. DOI: 10.1109/CoG47356.2020.9231744. URL: https://ieeexplore.ieee.org/document/9231744/ (visited on 08/17/2023).
- [33] Lin Chen and Sheng Xu. "DEEP NEURAL TANGENT KERNEL AND LAPLACE KERNEL HAVE THE SAME RKHS". In: (2021).
- [34] Amnon Geifman et al. On the Similarity between the Laplace and Neural Tangent Kernels. Nov. 14, 2020. arXiv: 2007.01580 [cs, stat]. URL: http://arxiv.org/abs/2007.01580 (visited on 08/17/2023). preprint.
- [35] Insu Han, Amir Zandieh, and Haim Avron. "Random Gegenbauer Features for Scalable Kernel Methods". In: ().
- [36] Vaishaal Shankar et al. "Neural Kernels Without Tangents". In: ().
- [37] David P Woodruff and Amir Zandieh. "Leverage Score Sampling for Tensor Product Matrices in Input Sparsity Time". In: ().
- [38] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. Feb. 10, 2020. arXiv: 1806.07572 [cs, math, stat]. URL: http://arxiv.org/abs/1806.07572 (visited on 08/13/2023). preprint.
- [39] James B Simon, Sajant Anand, and Michael R DeWeese. "Reverse Engineering the Neural Tangent Kernel". In: ().
- [40] Jaehoon Lee et al. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: Journal of Statistical Mechanics: Theory and Experiment 2020.12 (Dec. 1, 2020), p. 124002. ISSN: 1742-5468. DOI: 10.1088/1742-5468/abc62b. arXiv: 1902.06720 [cs, stat]. URL: http://arxiv.org/abs/1902.06720 (visited on 08/13/2023).
- [41] Jiaming Xu and Hanjing Zhu. "One-Pass Stochastic Gradient Descent in Overparametrized Two-layer Neural Networks". In: ().
- [42] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. Jan. 7, 2020. arXiv: 1812.07956 [cs, math]. URL: http://arxiv.org/abs/1812.07956 (visited on 08/17/2023). preprint.
- [43] Mario Geiger et al. "Disentangling Feature and Lazy Training in Deep Neural Networks". In: Journal of Statistical Mechanics: Theory and Experiment 2020.11 (Nov. 1, 2020), p. 113301. ISSN: 1742-5468. DOI: 10.1088/1742-5468/abc4de. URL: https://iopscience.iop.org/article/10.1088/1742-5468/abc4de (visited on 08/17/2023).
- [44] Behrooz Ghorbani et al. "Limitations of Lazy Training of Two-layers Neural Networks". In: ().