

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**Understanding deep learning through ultra-wide
neural networks**

by

Wei Huang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2021

Certificate of Authorship/Originality

I, Wei Huang, declare that this thesis, is submitted in fulfilment of the requirements for the award of PhD, in the Faculty of Engineering and IT at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 09/09/2021

Acknowledgements

In the first place, I want express my sincere gratitude to my supervisor, professor Richard Xu, for his unwavering support and invaluable instructions. Prof. Xu has brought me into the topics of non-parametric Bayesian and deep learning. He taught me a lot of machine learning knowledge by insisting on teaching every week. By revising the paper along with the discussion, he taught me how to write scientific papers. His profound knowledge and rigorous scientific research spirit deeply influenced the research path I would not have been able to carry out research projects without his valuable suggestions.

I would like to extend my sincere thanks to my co-supervisor Prof. Massimo Piccardi for helping with my candidature assessment, and I am greatly inspired by his warm support.

I am grateful to my collaborator Mr. Weitao Du who has been providing valuable support on mathematical techniques and proofs. We have been best friends since undergraduate students. We have spent plenty of time on discussion and collaborating. I have improved a lot in mathematical proofs. I sincerely appreciate my collaborator Mr. Yayong Li for his considerable discussion on the graph neural network. I am thankful for what I have learned from him during our collaborations.

I would like to offer my special thanks to all other members of our machine learning lab, especially, Mr. Chunrui Liu, Mr. Yunce Zhao, Mr. Ziyue Zhang, Mr. Xuan Liang, Mr. Chen Deng, Dr. Caoyuan Li, Miss Ying Li, Miss Sophie Yuan, Mr. Chris Markos, Dr. Jason Traish, Miss Ningkai Xiao, Miss Erica Huang, Mr. Haodong Chang, and Dr. Shawn Jiang.

My sincere thanks go to my friends for a happy time with them during my Ph.D. study. Those thanks especially go to Dr. Haihan Sun, Dr. Ye Shi, Miss Xuan Wang,

Miss Ye Su, Mr. Feng Shan, Mr. Haiming Qian, Mr. Leijie Zhang, Dr. Hongwen Yu, Dr. Tianyu Yang, Dr. Wentian Zhang, Miss Maral Ansari, Mr. Li Wang and Miss Chen Wang.

Finally, I would like to express my forever thanks to my parents for their selfless love and eternal support. My parents will always be my strong backing.

ABSTRACT

Understanding deep learning through ultra-wide neural networks

by

Wei Huang

Deep learning has been responsible for a step-change in performance across machine learning, setting new benchmarks in a large number of applications. However, the existing accounts fail to resolve why deep learning can achieve such great success. There is an urgent need to address the deep learning theory caused by the demand of understanding the principles of deep learning. One promising theoretical tool is the infinitely-wide neural network. This thesis focuses on the expressive power and optimization property of deep neural networks through investigating ultra-wide networks with four main contributions.

We first use the mean-field theory to study the expressivity of deep dropout networks. The traditional mean-field analysis adopts the gradient independence assumption that weights used during feed-forward are drawn independently from the ones used in backpropagation, which is not how neural networks are trained in a real setting. By breaking the independence assumption in the mean-field theory, we perform theoretical computation on linear dropout networks and a series of experiments on dropout networks. Furthermore, we investigate the maximum trainable length for deep dropout networks through a series of experiments and provide a more precise empirical formula that describes the trainable length than the original work.

Secondly, we study the dynamics of fully-connected, wide, and nonlinear networks with orthogonal initialization via neural tangent kernel (NTK). Through a series of propositions and lemmas, we prove that two NTKs, one corresponding to Gaussian weights and one to orthogonal weights, are equal when the network width

is infinite. This suggests that the orthogonal initialization cannot speed up training in the NTK regime. Last, with a thorough empirical investigation, we find that orthogonal initialization increases learning speeds in scenarios with a large learning rate or large depth.

The third contribution is characterizing the implicit bias effect of deep linear networks for binary classification using the logistic loss with a large learning rate. We claim that depending on the separation conditions of data, the loss will find a flatter minimum with a large learning rate. We rigorously prove this claim under the assumption of degenerate data by overcoming the difficulty of the non-constant Hessian of logistic loss and further characterize the behavior of loss and Hessian for non-separable data.

Finally, we demonstrate the trainability of deep Graph Convolutional Networks (GCNs) by studying the Gaussian Process Kernel (GPK) and Graph Neural Tangent Kernel (GNTK) of an infinitely-wide GCN, corresponding to the analysis on expressivity and trainability, respectively. We formulate the asymptotic behaviors of GNTK in the large depth, which enables us to reveal the dropping trainability of wide and deep GCNs at an exponential rate.

List of Publications

Journal Papers

- J-1. **Wei Huang** and Richard Yi Da Xu, "Gaussian Process Latent Variable Model Factorization for Context-aware Recommender Systems," 2019. Submitted to *Pattern Recognition Letters*.

Conference Papers

- C-1. **Wei Huang**, Richard Yi Da Xu, Weitao Du, Yutian Zeng, and Yunce Zhao, "Mean field theory for deep dropout networks: digging up gradient backpropagation deeply," *ECAI 2020, 24th European Conference on Artificial Intelligence*.
- C-2. **Wei Huang**, Weitao Du, and Richard Yi Da Xu, "On the neural tangent kernel of deep networks with orthogonal initialization," 2020. Submitted to IJCAI 2021.
- C-3. **Wei Huang**, Weitao Du, Richard Yi Da Xu, and Chunrui Liu, "Implicit bias of deep linear networks in the large learning rate phase," 2020. Submitted to IJCAI 2021.
- C-4. **Wei Huang**, Yayong Li, Weitao Du, Richard Yi Da Xu, Jie Yin, and Ling Chen, "Wide Graph Neural Networks: Aggregation Provably Leads to Exponentially Trainability Loss," 2021. Submitted to ICML 2021.

Contents

Certificate	ii
Acknowledgments	iii
Abstract	v
List of Publications	vii
List of Figures	xii
Abbreviation	xx
1 Introduction	1
1.1 Background	1
1.1.1 Machine learning	1
1.1.2 Deep learning	2
1.2 Motivation and Contribution	5
1.3 Thesis Organization	9
2 Literature Review	11
2.1 Mean-Field Theory	11
2.2 Neural Tangent Kernel	13
2.3 Implicit Bias in Deep Learning	14
3 Mean Field Theory for Deep Dropout Networks	17
3.1 Introduction	17
3.2 Background	18

3.2.1	Feed Forward	19
3.2.2	Back Propagation	22
3.3	Gradient Backpropagation	23
3.3.1	Breaking the gradient independence assumption	24
3.3.2	Emergence of universality	27
3.4	Experiments	29
3.4.1	Training speed	30
3.4.2	Trainable length	31
3.5	Discussion	33
3.6	Proof	34
3.6.1	Derivation of \tilde{q}_{aa}^l on linear dropout networks with a single input	34
3.6.2	Derivation of \tilde{q}_{ab}^l on linear dropout networks with a pair of inputs	39
4	Orthogonally-Initialized Networks and the Neural Tan- gent Kernel	45
4.1	Introduction	45
4.2	Preliminaries	47
4.2.1	Networks and Parameterization	47
4.2.2	Dynamical Isometry and Orthogonal Initialization	48
4.2.3	Neural Tangent Kernel	49
4.3	Theoretical results	51
4.3.1	An Orthogonally Initialized Network at Initialization	51
4.3.2	The limit of the NTK at initialization	52

4.3.3	Neural Tangent Kernel during training	55
4.4	Numerical experiments	57
4.5	Conclusion	61
4.6	Proof	61
4.6.1	NNGP at Initialization	62
4.6.2	NTK at Initialization	70
4.6.3	NTK during Training	76
5	Implicit bias of deep linear networks in the large learning rate phase	79
5.1	Introduction	79
5.2	Background	81
5.2.1	Setup	82
5.2.2	Separation Conditions of Dataset	82
5.3	Theoretical results	84
5.3.1	Convex Optimization	84
5.3.2	Non-convex Optimization	86
5.4	Experiments	91
5.5	Discussion	93
5.6	Proof	94
6	Wide graph neural networks: aggregation provably leads to exponentially trainability loss	108
6.1	Introduction	108
6.2	Background	111
6.2.1	Mean Field Theory and Expressivity	111

6.2.2	Neural Tangent Kernel and Trainability	112
6.2.3	GCNs	113
6.3	Theoretical Results	114
6.3.1	Expressivity of Infinitely Wide GCNs	114
6.3.2	Trainability of Infinitely Wide GCNs	117
6.3.3	Analysis on techniques to deepen GCNs	119
6.4	Experiments	123
6.4.1	Setup	123
6.4.2	Convergence Rate of GPKs and GNTKs	123
6.4.3	Trainability of Ultra-Wide GCNs	124
6.5	Conclusion	125
6.6	Proof	125
7	Conclusions and Future Work	139
7.1	Conclusions	139
7.2	Future Work	140
	Bibliography	142

List of Figures

1.1	A typical fully-connected neural network.	3
1.2	Double descent phenomenon in deep learning versus U-shaped curve in traditional machine learning [1].	5
3.1	The iterative squared length mapping of Equation (3.2) and Equation (3.4) with different activations and dropout rates. (a) q_{aa}^l in linear network at $\sigma_w = 0.5$ and $\sigma_b = 1.5$. Theoretical results match well with the simulations within a standard error (shadow). Different color correspond to different dropout rates: $\rho = 1$ is red, $\rho = 0.7$ is green, and $\rho = 0.4$ is blue. (b) The iterative length map of q_{aa}^l in Tanh network at $\sigma_w = 2.5$ and $\sigma_b = 0.5$. (c) The iterative length map of c_{ab}^l in ReLU network at $\sigma_w = 0.9$ and $\sigma_b = 0.5$. Only intersection of network at $\rho = 1$ (red) is $c_{ab}^* = 1$, the others are $c_{ab}^* < 1$. (d) The iterative length map of c_{ab}^l in Erf network at $\sigma_w = 0.9$ and $\sigma_b = 0.5$. Again, $c_{ab}^* = 1$ only holds at $\rho = 1$	19

- 3.2 Theoretical calculations versus network simulations for metric of gradient. (a) g_{aa}^l as a function of layer l , for a 200 layers random linear network with $\sigma_w^2 = 0.5$ and $\sigma_b^2 = 0.1$. (b) g_{ab}^l as a function of layer l . Theoretical calculations (solid lines) fail to predict empirical simulations (dashed lines). (c) g_{ab}^l as a function of layer l in the range of length $l = 170 - 200$. Theoretical calculations (solid lines) can predict empirical simulations (dashed lines) in the few last layers. (d) g_{ab}^l as a function of layer l . The solid lines are $g_{ab}^l \propto \chi_1^{L-l}$ for different ρ . Theoretical calculations failed to predict empirical simulations (dashed lines). 23
- 3.3 The metric of gradient with one and two different inputs, g_{aa}^l (solid lines), \tilde{g}_{ab}^l (dashed lines), and $g^l \propto \chi_1^{L-l}$ (dotted lines) as a function of layer l with different activation. (a) ReLU network with $\sigma_w^2 = 1.0$ and $\sigma_b^2 = 0.1$. (b) Tanh network with $\sigma_w^2 = 1.4$ and $\sigma_b^2 = 0.1$. (c) Hard Tanh network with $\sigma_w^2 = 1.4$ and $\sigma_b^2 = 0.1$ 23
- 3.4 Universal relationship between variance and mean of g_{aa}^l, g_{ab}^l , and \tilde{g}_{ab}^l , on the 200 layers and width $N = 500$ random dropout networks. Different color represents a different dropout rate. The black line is the function of $V \propto m^2$. (a) V_{aa}^l as a function m_{aa}^l . (b) V_{ab}^l as a function of m_{ab}^l . (c) \tilde{V}_{ab}^l as a function of \tilde{m}_{ab}^l . All the curves regarding different activations collapse to a line, and the power coefficient of all curves is consistent with 2. 25
- 3.5 Universal relationship between variance and mean of g_{aa}^l, g_{ab}^l , and \tilde{g}_{ab}^l , on the 200 layers, Tanh random dropout networks with $\rho = 0.9$. All the curves regarding different width collapse to a line. Different color represents a different network width. (a) V_{aa}^l as a function m_{aa}^l . (b) V_{ab}^l as a function of m_{ab}^l . (c) \tilde{V}_{ab}^l as a function of \tilde{m}_{ab}^l 26

- 3.6 The relation between steps and the learning rate η . (a) Network without dropout, colors reflect different network depth L from 50 (black) to 400 (green). (b) Network with dropout $\rho = 0.99$, colors reflect different network depth L from 20 (black) to 120 (green), additional $L = 300$ is colored blue for comparison. Curves with $L \leq 120$ collapse to a universal curve without any re-scale. (c) Network with dropout $\rho = 0.98$, colors reflect different network depth L from 10 (black) to 55 (green), additional $L = 200$ is colored blue for comparison. Curves with $L \leq 55$ collapse to a universal curve without any re-scale. 28
- 3.7 The training accuracy for neural networks as a function of the depth L and initial weight variance σ_w^2 from a high accuracy (bright yellow) to low accuracy (black). Comparison is made by plotting $12\xi_1$ (white solid line), $6\xi_2$ (green dashed line), and $12\xi_2$ (white dashed line). (a) 2000 training steps of $\rho = 1$ network with Gaussian weights on the MNIST using SGD. (b) 1000 training steps of $\rho = 1$ network with Gaussian weights on the MNIST using RMSProp. (c) 2000 training steps of $\rho = 1$ network with Orthogonal weights on the MNIST. (d) 3000 training steps of $\rho = 1$ network with Orthogonal weights on CIFAR10. (e) 3000 training steps of $\rho = 0.99$ network with Orthogonal weights on the MNIST. (f) 3000 training steps of $\rho = 0.98$ network with Orthogonal weights on the MNIST using SGD. (g) 10000 training steps of $\rho = 0.98$ network with Gaussian weights on the MNIST. (h) 3000 training steps of $\rho = 0.95$ network with Orthogonal weights on the MNIST using SGD. 30
- 4.1 (a) Gaussian initialized network of NNGP. (b) Orthogonally initialized network of NNGP. (c) Gaussian initialized network of NTK. (d) Orthogonally initialized network of NTK. All the kernels are consistent with convergence rate of $O(n^{-\frac{1}{2}})$ 52

4.2 Changes of weights, empirical NTK on a three hidden layer Erf Network. Solid lines correspond to empirical simulation and dotted lines are theoretical predictions, i.e. black dotted lines are $1/\sqrt{n}$ while red dotted lines are $1/n$. (a) weight changes on Gaussian initialized network. (b) weight changes on the orthogonal initialized network. (c) NTK changes on both Gaussian and orthogonal networks. 53

4.3 Dynamics of full batch gradient descent on Gaussian and orthogonal initialized networks of $T = 10^4$ steps. Orthogonal networks behaves similarly to dynamics on the corresponding Gaussian networks, for loss and accuracy functions. The dataset is selected from full CIFAR10 with $D = 256$, while MSE loss and tanh fully-connected networks are adopted for the classification task. (a)(b) Network with depth $L = 3$ and width of $n = 400$, with $\sigma_w^2 = 1.5$, and $\sigma_b^2 = 0.01$. (c)(d) Network with depth $L = 7$ and width of $n = 800$, with $\sigma_w^2 = 1.5$, and $\sigma_b^2 = 0.1$. While the solid lines stand for Gaussian weights, dotted lines represent orthogonal initialization. 56

4.4 Orthogonally initialized networks behave similarly to the networks with Gaussian initialization in the NTK regime. (a)(b) We adopt the network architecture of depth of $L = 5$, width of $n = 800$, activation of tanh function, with $\sigma_w^2 = 2.0$, and $\sigma_b^2 = 0.1$. The networks are trained by SGD with a learning rate $\eta = 10^{-3}$ with $T = 10^5$. (c)(d) The hyper-parameters are: depth of $L = 9$, width of $n = 1600$, activation of ReLU function, with $\sigma_w^2 = 2.0$, and $\sigma_b^2 = 0.1$. The networks are trained by PMSProp with a learning rate $\eta = 10^{-5}$ with $T = 1.2 \times 10^4$ steps with a batch size of 10^3 on MSE loss on MNIST. While the solid lines stand for Gaussian weights, dotted lines represent orthogonal initialization. 57

- 4.5 Learning dynamics measured by the optimization and generalization accuracy on train set and test set. The depth is $L = 100$ and width is $n = 400$. Black curves are the results of orthogonal initialization, and red curves are performances of Gaussian initialization. (a) The training speed of an orthogonally initialized network is faster than that of a Gaussian initialized network. (b) On the test set, the orthogonally initialized network not only trains with a higher speed but also ultimately converges to a better generalization performance. 58
- 4.6 The steps τ as a function of learning rate η of two lines of networks on both train and test dataset. The results of orthogonal networks are marked by dotted lines while those of Gaussian initialization are plotted by solid lines. Networks with varying width, i.e. $n = 400, 800,$ and 1600 , on (a) train set and (b) test set; Networks with varying depth, i.e. $L = 50, 100,$ and 200 , on (c) train set and (d) test set. Different colors represent the corresponding width and depth. While curves of orthogonal initialization are lower than those of Gaussian initialization with a small learning step, the differences become more significant when we increase the learning rate. Besides, the greater the depth of the network, the more significant the difference in performance between orthogonal and Gaussian initialization. 59

- 5.1 Dependence of dynamics of training loss on the learning rate for linear predictor, with (a,b) exponential loss and (c,d) logistic loss on Example 5.1 and 5.2. (a,c) The experimental learning curves are consistent with the theoretical prediction, and the critical learning rates are $\eta_{\text{critical}} = 1.66843$ and $\eta_{\text{critical}} = 8.485$ respectively. (b,d) For non-separable data, the dynamics of training loss regarding the learning rate for non-separable data are similar to those of a degenerate case. Hence the critical learning rates can be approximated by $\eta_{\text{critical}} = 0.895$ and $\eta_{\text{critical}} = 4.65$ respectively. . . . 86
- 5.2 Dependence of dynamics of training loss and maximum eigenvalue of NTK on the learning rate, with (a,b,e,f) exponential loss and (c,d,g,h) logistic loss on Example 5.3 and 5.4. (a,b,c,d) The loss increases at the beginning and converges to a global minimum. (e,f,g,h) The maximum eigenvalue of NTK converges to a value which is lower than its initial position. 88
- 5.3 Top eigenvalue of NTK (λ_0) and Hessian (h_0) measured at $t = 100$ as a function of the learning rate, with (a,b) exponential loss and (c,d) logistic loss on Example 5.3 and 5.4. The green dashed line $\eta = \eta_0$ represents the boundary between the lazy phase and catapult phase, while black dashed line $\eta = \eta_1$ separates the other two phases. The setting are: $\sigma_w^2 = 0.5$ and $m = 100$ for exponential loss, and the setting for logistic loss is $\sigma_w^2 = 0.5$ and $m = 200$. (a,c) The curves of maximum eigenvalue of NTK and Hessian coincide as predicted by the corollary 5.1. (b,d) For the non-separable data, the trend of the two eigenvalue curves is consistent with the change of learning rate. . . 90

5.4 Test performance of deep linear networks with respect to different learning rate phases. The data size is of $n_{\text{train}} = 2048$ and $n_{\text{test}} = 512$. (a,b) A two-layer linear network without bias of $\sigma_w^2 = 0.5$ and $m = 500$. (c,d) A three-layer linear network with a bias of $\sigma_w^2 = 0.5$, $\sigma_b^2 = 0.01$, and $m = 500$. (a,c) The test accuracy is measured at the time step $t = 500$ and $t = 300$ respectively. (b,d) The test accuracy is measured at the physical time step (red curve), after which it continues to evolve for a period of time at a small learning rate (purple): $t_{\text{phy}} = 50/\eta$ and extra time $t = 500$ at $\eta = 0.01$ for the decay case. Although the results in the catapult phase do not perform as well as the lazy phase when there is no decay, the best performance can be found in the catapult phase when adopting learning rate annealing. 92

5.5 Graph of $\phi(x)$ for the two losses. (a) Exponential loss with learning rate $\eta = 10$. (b) Logistic loss with learning rate $\eta = 10$ 97

5.6 Different colors represent different $\lambda(\text{NTK})$ values. (a) graph of $\phi_\lambda(x)$ equipped with the exponential loss. (b) graph of the derivative of $\phi_\lambda(x)$ equipped with the exponential loss. (c) graph of $\phi_\lambda(x)$ equipped with the logistic loss. (d) graph of the derivative of $\phi_\lambda(x)$ equipped with the logistic loss. Notice that the critical point of the exponential loss moves to the right as λ decreases. 100

6.1 Overview of the information propagation in a general GCN 114

6.2 Convergence rate for the GPK and the GNKT. (a) Value changes of the GPK elements as the depth grows; although their initial values are different, they all tend to the same value as depth increases. (b) The distance changes between GPK elements and their limiting value as the depth grows; the converge rate can be bounded by a exponential function $y = \exp(-0.15x)$. (c) Value changes of re-normalized GNTK elements as the depth grows. (d) The distance changes between re-normalized GNTK elements and a random element from GNTK as the depth grows. The converge rate can be bounded by a exponential function $y = \exp(-0.15x)$ 121

6.3 Train and test accuracy depending on the depth on different datasets. Solid lines are train accuracy and dashed lines are test accuracy. 123

Abbreviation

ML - Machine Learning

DL - Deep Learning

FCN - Fully-Connected Network

MLP - Multi-Layer Perceptro

GP - Gaussian Process

GPK - Gaussian Process Kernel

GNTK - Graph Neural Tangent Kernel

GD - Gradient Descent

Resnet - Residual Network

ReLU - Rectified Linear Unit

Erf - Error Function

FIM - Fisher Information Matrix

CLT - Central Limit Theorem

MSE - Mean Squared Error

SVM - Support Vector Machine

GCN - Graph Convolutional Network