Why Do Deep Neural Networks Generalize Well?

From the topological and geometrical perspectives

Implicit Biases of Deep Neural Networks

Classical statistical learning theory posits a trade-off between bias and variance, implying that highly overparameterized models should generalize poorly to unseen data. However, modern deep neural networks (DNNs) defy this expectation: they achieve near-zero training error yet generalize remarkably well [1, 2]. This phenomenon, known as *benign overfitting*, has also been observed in non-deep models such as trees, boosting methods, and linear regression [3, 4, 5, 6]. A related observation is the *double descent* curve of generalization error with respect to model complexity, which deviates from the classical U-shaped pattern by exhibiting a second descent in the overparameterized regime [3, 7, 8]. These phenomena have sparked significant interest in understanding the underlying mechanisms driving generalization in DNNs [7, 9, 10, 11, 12, 13]. Yet, existing explanations are fragmented: most studies focus on a single aspect and are primarily validated empirically in small-sample settings, lacking a unified theoretical framework.

Broadly, the literature attempts to explain this puzzle along two directions. The first develops new model complexity measures that are often significantly smaller than the number of parameters, arguing that overparameterized models may not be as complex as they appear [14, 7]. A representative work in this group is [7], which argues that the double descent pattern arises from projecting a fundamentally two-dimensional phenomenon—referring to a two-dimensional complexity measure—onto a one-dimensional axis. When the full two-dimensional representation is used, the classical U-shaped generalization curve is recovered. However, this analysis is limited to non-deep models, and its applicability to DNNs remains uncertain. An alternative complexity measure is the *real log canonical threshold* (RLCT), a central concept from *singular learning theory*, which we will explore further later.

The second direction hypothesizes that DNNs exhibit *implicit bias* towards low-complexity functions that generalize well [15, 16]. The origin of this bias is debated. One line attributes it to the structure of the loss landscape or the architecture itself (i.e., the parameter-function map) [15, 9, 17], while another focuses on the properties of gradient-based optimizers like stochastic gradient descent (SGD) [18, 19, 20, 21, 13]. Disentangling these sources is difficult, as current empirical evidence is often fragmented or contradictory.

This proposal aims to contribute a theoretically grounded and unified view of generalization in DNNs by leveraging tools from geometry and topology. While much of the existing literature focuses on isolated explanations—either model complexity or optimization dynamics—our approach seeks to integrate these perspectives through geometric and topological analysis. By doing so, we hope to clarify the sources of implicit bias and provide principled complexity measures that are both theoretically meaningful and practically applicable to overparameterized models like DNNs.

Topology and Geometry in Neural Networks

Topology and geometry offer powerful yet underutilized lenses for understanding the generalization properties of DNNs. In this section, I review key insights from these perspectives and outline existing limitations.

Recent work has used *persistent homology*, a tool from algebraic topology, to analyze how the topological complexity of data evolves through the layers of trained deep neural networks [22, 23, 24]. These studies provide *empirical* insights: (1) nonsmooth activations like ReLU outperform smooth ones because they more effectively reduce topological complexity, and (2) deep networks outperform shallow ones by gradually simplifying topological features layer by layer. However, both shallow and deep networks ultimately reduce topological complexity, suggesting that topological simplification alone does not fully account for the generalization advantage of DNNs. Moreover, these works focus on *trained networks* and do not address why overparameterized networks generalize in the first place. Building on this line of topological analysis, more recent studies have turned to the topology

of the training process itself. For instance, [25, 26, 27] examine the topology of training trajectories and demonstrate that generalization error can be bounded using a topological descriptor called the *persistent homology dimension* (PHD) [28]. Empirical results suggest that PHD predicts generalization well, indicating that training trajectory topology plays a role. However, these works do not distinguish whether the observed implicit biases stem from the optimizer (e.g., SGD) or the loss landscape.

From a geometric perspective, *singular learning theory* (SLT) provides a rigorous framework for analyzing models like DNNs, whose parameter-function maps are non-injective and whose Fisher information matrices may be degenerate at some parameters [14, 29, 30]. These *singularities* in the parameter space challenge classical learning theory. SLT introduces the *real log canonical threshold* (RLCT), which quantifies the effective model complexity via resolution of singularities. It refines the classical bias-variance trade-off: models with lower RLCT (i.e., more singular) are preferred when sample sizes are small, while models with higher RLCT (i.e., less singular but more accurate) are preferred with larger samples. Importantly, SLT offers insights into why DNNs generalize well in the overparameterized regime—it suggests their loss sublevel sets contain richer singular structures (i.e., lower RLCT) than those of other models. However, SLT is developed in a Bayesian setting, and its relevance to models trained via SGD—outside the Bayesian framework—remains an open question.

Research Plan

This project will develop a unified theoretical understanding of generalization in overparameterized DNNs by pursuing both major explanatory directions outlined above.

(1) Model Complexity via RLCT. We will explore RLCT as a generalization-aware complexity measure for DNNs. Unlike the complexity measure proposed in [7], which is tailored to non-deep learning models like trees and linear regressors, the RLCT from singular learning theory is grounded in algebraic geometry and extends more naturally to deep neural networks. Moreover, similar to the approach in [7], there remains flexibility in how RLCT is estimated—such as using test data rather than training data—making it adaptable to the deep learning context. Two challenges must be addressed:

- *Efficient Estimation.* While estimating RLCTs is generally intractable in large models, recent advances suggest that approximate estimation may be feasible in practice. In particular, [31] proposes a method that leverages stochastic gradient Langevin dynamics (SGLD) to estimate local learning coefficients—closely related to RLCTs—by sampling from the posterior distribution near singularities, offering a scalable and geometry-aware approximation framework.
- *Applicability to SGD-trained Models.* SLT connects RLCT to generalization via KL divergence under Bayesian inference, whereas deep learning typically uses empirical generalization gaps. Establishing whether RLCT remains predictive for SGD-trained models is a key goal.

(2) Disentangling Implicit Bias. We aim to identify whether implicit bias arises from the loss landscape or from the optimization algorithm:

- *Loss Landscape.* We will analyze and compare the number and RLCT of singularities in sublevel sets of DNNs versus other models. Exploring the topology of these singularities—beyond RLCT's quantitative measure—offers a promising new direction for *qualitative* analysis.
- *Optimization Algorithm.* We will compare the behaviour of SGD and Bayesian samplers in navigating the loss landscape. Prior work [17] shows *empirical* alignment between SGD outputs and Bayesian posteriors. We aim to investigate how each interacts with singularities in the loss landscape.

References

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[2] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849--15854, 2019.

[4] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063--30070, 2020.

[5] Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1--76, 2023.

[6] Shange Tang, Jiayun Wu, Jianqing Fan, and Chi Jin. Benign overfitting in out-of-distribution generalization of linear models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A u-turn on double descent: rethinking parameter counting in statistical learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[8] Ouns El Harzli, Bernardo Cuenca Grau, Guillermo Valle-Pérez, and Ard A. Louis. Double-descent curves in neural networks: A new perspective using gaussian processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11856--11864, Mar. 2024.

[9] Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2022.

[10] Kedar Karhadkar, Erin George, Michael Murray, Guido F Montufar, and Deanna Needell. Benign overfitting in leaky relu networks with moderate input dimension. *Advances in Neural Information Processing Systems*, 37:36634--36682, 2024.

[11] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

[12] Niladri S. Chatterji, Philip M. Long, and Peter L. Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *Journal of Machine Learning Research*, 23(263):1--48, 2022.

[13] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12978--12991. Curran Associates, Inc., 2021.

[14] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, USA, 2009.

[15] Giacomo De Palma, Bobak Kiani, and Seth Lloyd. Random deep neural networks are biased towards simple functions. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] Chris Mingard, Henry Rees, Guillermo Valle-Pérez, and Ard A Louis. Deep neural networks have an inbuilt occam's razor. *Nature Communications*, 16(1):220, 2025.

[17] Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is sgd a bayesian sampler? well, almost. *J. Mach. Learn. Res.*, 22(1), January 2021.

[18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1--57, 2018.

[19] Hui Jin and Guido Montúfar. Implicit bias of gradient descent for mean squared error regression with two-layer wide neural networks. *J. Mach. Learn. Res.*, 24(1), January 2023.

[20] Amit Peleg and Matthias Hein. Bias of stochastic gradient descent or the architecture: disentangling the effects of overparameterization of neural networks. *arXiv preprint arXiv:2407.03848*, 2024.

[21] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? --a mathematical framework. In *International Conference on Learning Representations*, 2022.

[22] Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1--40, 2020.

[23] German Magai and Anton Ayzenberg. Topology and geometry of data manifold in deep learning. *arXiv preprint arXiv:2204.08624*, 2022.

[24] Satoru Watanabe and Hayato Yamana. Topological measurement of deep neural networks using persistent homology. *Annals of Mathematics and Artificial Intelligence*, 90(1):75--92, 2022.

[25] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776--6789, 2021.

[26] Benjamin Dupuis, George Deligiannidis, and Umut Şimşekli. Generalization bounds using data-dependent fractal dimensions. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[27] Rayna Andreeva, Benjamin Dupuis, Rik Sarkar, Tolga Birdal, and Umut Simsekli. Topological generalization bounds for discrete-time stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 37:4765--4818, 2024.

[28] Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Joshua Mirth, Rachel Neville, Chris Peterson, and Clayton Shonkwiler. A fractal dimension for measures via persistent homology. In Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule, editors, *Topological Data Analysis*, pages 1--31, Cham, 2020. Springer International Publishing.

[29] Sumio Watanabe and Manfred Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

[30] Sumio Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867--897, 2013.

[31] Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure, 2024.