# Abstract

A decade of technological advancement driven by the success of deep learning in various tasks [1–3] is not yet supported by a theoretical framework able to capture the features of architectures, loss functions and dynamics that make the learning possible and in fact very fruitful [4–7]. This challenge has raised the attention of the theoretician community in multiple areas of science. However, despite notable effort to analytically study deep learning [8–15], there are some fundamental questions that are yet to be addressed, for example (i) can we predict practically relevant scores like train and generalization error of deep networks in realistic regimes? (ii) how does information contained in real-world datasets is exploited by the network to extract useful representations (features)?

A great part of the classic theoretical results that we have in machine learning make use of some simple assumption on the training data distribution [16–20]. The physical reason to make such assumption has its roots in the classical statistical mechanics description of disordered systems, mainly spin-glass theory. A fruitful line of research in machine learning, that inherits from disordered systems, indeed aims to compute a partition function that is a *quenched* or *annealed* average over the training data distribution, which is the source of the disorder, allowing to describe the performances of the typical solution independently of the specific dataset used to train the network

[21–23]. This analysis, although providing results that hold in full generality, suffers some limitations, mainly (i) the assumption of simple data distributions, which is essential to analytically compute the averaged partition function, is unrealistic when applied to practical settings in machine learning, for example in computer vision, where the spatial information contained in the dataset is crucial to achieve almost-optimal generalization performance. (ii) averaging over data is very hard when dealing with deep networks. The architectures that are amenable to this kind of study have at most one layer of trainable parameters (i.e. the perceptron, the random features model, and the committee machine).

In my PhD work, as is suggested by the title, I explored a complementary approach to the one discussed above, which does not make any assumption on the structure/distribution of the training data. In this framework, I will show how to derive explicit formulas for the training and generalization error of trained fully-connected deep networks, shallow convolutional and locally-connected networks, in a regime of learning, called *proportional regime*, that assumes the size of the dataset $P$ to be comparable in magnitude to the width of the hidden layers in the model $N_\ell$ ($\ell = 1, \dots, L$, $L$ being the (finite) depth of the network). The observables that I will show how to compute in this scenario retain an explicit dependence on the training data, since this is never averaged out. Remarkably, it is indeed this dependence that helps to conjecture how the network can operatively exploit the information contained in the trainset to make informed prediction on unseen data, and how this capability is linked to the topology of the network connections. The present work is organized as follows.

In Chapter 1, I will introduce kernel methods, the state-of the art algorithms for object recognition before deep learning, explaining why they still retain a theoretical interest as limiting dynamics of neural networks in a certain regime (the *infinite-width* limit). A crucial link between kernel methods, wide networks and Gaussian Proccesses will also be explored in this chapter, in view of the forthcoming discussion.

Chapter 2 will be dedicated to a class of results on the so-called *infinite-width limit* of neural networks that leverage the same data-agnostic spirit [24–32]. The infinite-width limit is informally defined as the regime where the size of each hidden layer $N_\ell$ is much larger than the size of the training set $P$. Here, one shows that the stochastic process that describes information flow in the deep neural network is a familiar Gaussian process (GP), which is completely determined by a non-linear kernel $K_L$. A fundamental consequence of this finding is that learning in the infinite-width limit is equivalent to kernel learning [8, 20, 33, 34] with a static kernel $K_L$ that does not evolve during the training dynamics and is completely fixed once the network's weights are initialized. Notably, given the incredibly general nature of these results, GP limits can be derived for virtually any feedforward architecture [29, 35, 36].

In Chapter 3, I will discuss the critical topic of *feature learning* [37–41], i.e. the capability of deep networks to automatically detect useful representations from raw data. This is a fundamental aspect that any minimal theory of deep learning should be able to quantitatively address, and constitutes a limitation of the infinite-width regime, where it is essentially absent [42]. On the contrary, I will show evidence that feature learning occurs and is in fact essential in finite-width convolutional networks, but is almost absent in finite-width standard scaled 1HL fully-connected networks in the proportional regime.

In Chapter 4, I will show how this data-agnostic approach can be extended, using the tools of physics, beyond the infinite-width limit, in particular in the proportional regime introduced above. I will show how a statistical mechanics description is possible in this scenario both for FC networks of arbitrary finite depth, and for shallow networks with local connections, with and without weight sharing.

In Chapter 5, I will try and rationalize the observation made in Chapter 3, through the lense of the framework introduced in Chapter 4. I will show how, thanks the mechanism of *local kernel renormalization*, one can effectively quantify what it means to be "far" from the kernel regime, providing a possible mathematical description of what it means to learn features in neural networks. Inspection of the effective action for a simple architecture with one convolutional HL in the proportional regime, shows a striking difference with respect to the fully-connected case: whereas the FC kernel is just globally renormalized by a scalar parameter, the CNN kernel undergoes a local renormalization, meaning that many more free parameters are allowed to be fine-tuned during training. This finding can be employed to highlight a simple mechanism for feature learning that can take place in finite-width shallow CNNs, but neither in shallow FC architectures nor in LCNs without weight sharing.

# Bibliography

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. → [p1]

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. → [p]

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. → [p1]

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. → [p1]

[5] Lenka Zdeborová. Understanding deep learning is also a job for physicists. *Nature Physics*, 16(6):602–604, 2020. → [p]

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017. → [p]

[7] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 2021. → [p1]

[8] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021. → [p1], [p3]

[9] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, Sep 2021. → [p]

[10] Albert J. Wakhloo, Tamara J. Sussman, and SueYeon Chung. Linear classification of neural manifolds with correlated variability. *Phys. Rev. Lett.*, 131:027301, Jul 2023. → [p]

[11] Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, 02 2023. → [p]

[12] Carlo Baldassi, Clarissa Lauditi, Enrico M. Malatesta, Rosalba Pacelli, Gabriele Perugini, and Riccardo Zecchina. Learning through atypical phase transitions in overparameterized neural networks. *Phys. Rev. E*, 106:014116, Jul 2022. → [p]

[13] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1):501–528, 2020. → [p]

[14] Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, 119(40):e2201854119, 2022. → [p]

[15] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, Dec 2020. → [p1]

[16] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. → [p1]

[17] Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965. → [p]

[18] E Gardner. Maximum storage capacity in neural networks. *Europhysics Letters (EPL)*, 4(4):481–485, aug 1987. → [p]

[19] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271–284, jan 1988. → [p]

[20] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82:2975–2978, Apr 1999. → [p1], [p3]

[21] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124013, dec 2021. → [p2]

[22] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. 2nd Mathematical and Scientific Machine Learning Conference, 2021. Online, August 16-19, 2021. → [p]

[23] Hugo Cui, Florent Krzakala, and Lenka Zdeborová. Optimal learning of deep random networks of extensive-width. *arXiv preprint arXiv:2302.00375*, 2023. → [p2]

[24] Radford M. Neal. *Priors for Infinite Networks*, pages 29–53. Springer New York, New York, NY, 1996. → [p3]

[25] Christopher Williams. Computing with infinite networks. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. → [p]

[26] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. → [p]

[27] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. → [p]

[28] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019. → [p]

[29] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019. → [p3]

[30] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. → [p]

[31] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. → [p]

[32] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. → [p3]

[33] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. → [p3]

[34] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020. → [p3]

[35] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4376–4386. PMLR, 13–18 Jul 2020. → [p3]

[36] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. → [p3]

[37] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. → [p3]

[38] Yu Han Liu. Feature extraction and image recognition with convolutional neural networks. *Journal of Physics: Conference Series*, 1087(6):062032, sep 2018. → [p]

[39] Jianchang Mao and A.K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2):296–317, 1995. → [p]

[40] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. → [p]

[41] Dong Yu, Michael L Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide. Feature learning in deep neural networks-studies on speech recognition tasks. *arXiv preprint arXiv:1301.3605*, 2013. → [p3]

[42] Nikhil Vyas, Yamini Bansal, and Nakkiran Preetum. Limitations of the ntk for understanding generalization in deep learning. *arXiv preprint arXiv:2206.10012*, 2022. → [p3]

•