How to scale-up?
Foundations for Science of Scaling in Deep Learning

Lorenzo Noci

Motivation

Recent trends in deep learning have shown how the joint scaling of model size and number of training samples produces ever-increasing performances that follow a power law predictably [1–3]. These scaling laws have been extended to study how dataset size and model parameters should be traded under a fixed computational budget [2], and their predictions are arguably one of the core ingredients behind the design choices and success of modern large language models (LLMs) [4, 2]. With regard to model size, there is evidence that the benefits of scaling up the architecture apply transversally across several machine learning applications beyond LLMs [5–8], including computer vision [9, 10], continual learning [11], and scientific applications [12–14].

New Challenges at Scale. Despite the remarkable achievements, with greater deep learning systems come new and greater challenges. First, there are observed training instabilities at scale, such as unexpected loss spikes that require resuming the model's training from an earlier checkpoint [7], the emergence of large outliers features that prevents low-precision quantization [15, 16], and the entropy collapse of Transformers'[17] attention heads [18]. Second, establishing the optimal hyperparameters. In complex deep learning systems, the number of hyperparameters is large, including the learning rate, momentum, learning rate schedules. The search space is too large to allow for grid search, and it becomes economically prohibitive with increasing model sizes and training time. In fact, a survey performed at NeurIPS 2022 showed that more than half of the researchers uses less than 25 tuning trials [19]. At large scale, tuning is largely not performed, with the hope that the hyperparameters will transfer from small to large models [2, 20, 21, 10]. Finally, a third challenge comes with choosing *how* to increase model size. In deep learning, this choice consists in increasing either the width or the depth of the model. However, there remains a limited understanding of how to find depth and width trade-offs to maximize training speed [10, 22].

My research objective is to build the foundations for scaling up deep learning, with the aim of improving architecture design and solving these new challenges that come with larger scale. To achieve that, I intend to contribute to the theory of scaling limits for neural networks, where model's size (width, depth) and the sample size are taken to infinity.

Theoretical Foundations for Science of Scaling

Why scaling limits? Studying the asymptotic limit has a twofold potential. Firstly, the limiting behaviour of the system is often mathematically more tractable than at finite size. This has the potential to study training and generalization of neural networks through its simplified limit [23–26], and understand complex state-of-the-art architectures [27–30]. Secondly, through scaling limits we might ensure the existence of a limit with desirable properties (e.g. non vanishing gradients) [31]. Satisfying these desiderata might lead to architectural modifications, or to the prescription of how optimization components (e.g. the learning rate) should scale with the scaling quantities [32–35].

Literature Review. Existing theories for neural network's scaling limits mainly focus on the infinite width limit, either in the lazy regime [36–43] or rich/feature learning regime [44–47]. This class of limits relies on the central limit theorem and the law of large numbers to establish concentration to deterministic quantities of the kernels involved [46]. However, they generally treat depth as constant [48], thus they have limitations in modeling deep neural networks [49, 50, 34]. A theoretical characterization of the infinite depth-and-width limit has been mainly performed at initialization [49, 51–53], and in special cases during training [54, 33]. Depth scaling has also been studied in signal propagation in random networks [55–57], highlighting the roles of the activation function [57–60, 31] and the initialization [61] in mitigating signal degeneracies. However, the joint scaling of dataset size, width and depth has been less studied [62]. Finally, the joint dataset size and width limit in the realm of kernels or linear models has been explored in a number of foundational works [63, 64, 3, 65–68], and in few cases in the feature learning regime, after one step of training [69, 70].

Research achievements

Theory: Infinite Width-and-Depth limits. In Noci et al. [49] we use hypergeometric functions [71] to characterize the distribution at initialization of a ReLU network's output at *any finite* width and depth. We also devise its *joint* width-and-depth limit, and show how it retains the complex characteristics of finite deep networks that would be lost in the infinite width limit. In the joint limit, the depth-to-width ratio becomes central, controlling the amount of deviation from Gaussianity due to large depth [51–53]. In Noci et al. [35] we devise the first infinite width-and-depth model for the Transformer architecture using the neural covariance SDE framework [34], where the infinitesimal modification of the representations in depth can be described with stochastic differential equations in the joint limit. Finally, there is an alternative parametrization that introduces residual branches adequately scaled with the depth of the network [32], where the depth and width limits in fact commute [72]. In Bordelon et al. [54] (similarly, in current work [33]) we show that the initialization and training dynamics of depth-scaled residual networks can be characterized in the infinite width-and-depth limit.

Practical Consequences for Architecture Design. The existence of a well-defined scaling limit for the architecture often provides guidance on architectural and optimization design. In Noci et al. [73] we show how at large depth Transformers may enter a pathological regime, namely the rank collapse of the representations [60], where the representations of different inputs perfectly align at large depth. This hinders trainability by causing vanishing gradients [73]. In Noci et al. [35] we illustrate how an infinite width-and-depth transformer without these pathologies can be designed by making the attention function close to linear in a precise width/depth-dependent way. The Transformer's attention mechanism is modified by centering the Softmax output at identity, and scaling the Softmax logits by a width-dependent temperature parameter. We also provide experimental evidence showing how a Transformer with these modifications can be trained at large depth with comparable performances to widely adopted Transformers without presenting vanishing gradients, or the well-known instability of entropy collapse [74, 18].

Hyperparameter Transfer. Another practical consequence of the theory of scaling limits is presented in Bordelon et al. [54]: we show that when combined with the feature learning parametrization of μP [21, 33], depth-scaled residual networks exhibit the empirical phenomenon of hyperparameter transfer (parametrization that we call $\sqrt{\text{depth}}$-μP). There, we prescribe how certain

optimal hyperparameters, such as the learning rate, should be scaled from small to large models, thus avoiding expensive hyperparameter tuning. The transfer property holds for various hyperparameters, including momentum, regularization strength and cosine learning rate schedule, and various architectures, including ResNets with normalization layers and Vision Transformers. Beyond hyperparameter transfer, our parametrization gives a consistent improvement in performance with increasing width and depth, which is not the case for other unstable or kernel parametrizations. Finally, in Noci et al. [75], the phenomenon of hyperparameter transfer is examined by analyzing the loss landscape during training as model's width and depth increase. The study highlights how certain properties, such as sharpness (the largest eigenvalue of the loss Hessian), remain consistent across scales. Under the appropriate parametrization, the sharpness dynamics are shown to be largely independent of width, approaching the edge of stability threshold [76]. This provides strong evidence for an optimization-based explanation of why learning rate transfer works effectively.

Other Directions I have also contributed in the field of Bayesian Neural networks [77, 78], ensembling [79], Transformer's inference and the pheonomenon of outlier features [80, 81]. The full list of publications can be found in the CV or Google Scholar.

Future Directions

I aim to advance the foundations of large-scale neural networks, focusing on training dynamics and generalization in the joint limit of dataset size, width, and depth. My goal is to develop a unified theory that incorporates widely used architectural components, such as attention mechanisms and residual connections, while also exploring empirical applications that emerge during this research. To achieve that, I outline some intermediate research goals and possible applications:

Feature Learning in the Joint Limit. So far the depth-and-width limit has been characterized in the commutative setting where the limits can be taken sequentially, i.e. in neural networks with depth-scaled residual connections [72, 33, 54] ($\sqrt{\text{depth}}$-μP). I plan to study training dynamics and learning rate parametrization in the joint (non-commutative) limit, where also the activation function needs to be parametrized in terms of width and depth [34, 35]. These results would be valuable on their own; for instance, this scaling might enable hyperparameter transfer across width and depth, and would thus provide an alternative scaling as opposed to $\sqrt{\text{depth}}$-μP.

Which parametrization is better? In the presence of several different limits with different properties, it is unclear which would make more efficient use of parameters and data. Thus, I plan to classify the various scaling limits (e.g. $\sqrt{\text{depth}}$-μP versus joint non-commutative limit) using scaling laws, that allow to make predictions on the model's performances. This classification would give a precise prescription as to which scaling strategy should be adopted.

Joint Dataset Size, Depth and Width limit. Scaling laws predict an optimal trade-off between the number of parameters (depth and width) and data set size. When the width is much larger than the sample size, the model converges to the limit at a given rate independent of it [3]. However, the interesting power-law scaling of the loss happens in the scaling regime where dataset size and width are of the same order [1, 3, 82]. Explaining and improving the sample efficiency of deep neural networks thus requires understanding the joint dataset size, depth and width scaling limit. Establishing this limit in the feature learning regime would open several possibilities, such as (1) studying asymptotic generalization curves in a model that faithfully represents finite neural networks, and (2) providing a prescription for hyperparameter transfer when the data is also a

scaling quantity. To this end, there are tools that could leveraged, such as approximate message passing to study multiple gradient steps [64, 83] and tools from random matrix theory [84], such as Dyson Brownian motion [85] and spiked covariance models [86], to understand the large dimensional kernels that are involved in the dynamics.

Scaling Laws and Architecture Design It is plausible that scaling laws of performances are more reliable under a parametrization that has a well-defined feature learning limit. I intend to test how reliable the scaling exponents with current architectures are. To achieve this, I plan to examine the empirical scaling behavior of current architectures, under various parametrizations, including those derived from the proposed research plan, either in the fixed budget or compute optimal setting [2]. Additionally, I aim to integrate theoretical insights from the joint limit framework to propose modifications to activations functions [35], and study optimal depth-width trade-offs. The overall goal here is to converge on a parametrization for the deep learning system (i.e. architecture and optimizer) that can be used for optimal pretraining, both for language and multimodal models.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

[2] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556* (2022).

[3] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. "Explaining neural scaling laws". In: *arXiv preprint arXiv:2102.06701* (2021).

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[6] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. "Scaling language models: Methods, analysis & insights from training gopher". In: *arXiv preprint arXiv:2112.11446* (2021).

[7] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. "Palm: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113.

[8] N. Dey, G. Gosal, H. Khachane, W. Marshall, R. Pathria, M. Tom, J. Hestness, et al. "Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster". In: *arXiv preprint arXiv:2304.03208* (2023).

[9] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. "Scaling vision transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 12104–12113.

[10] I. M. Alabdulmohsin, X. Zhai, A. Kolesnikov, and L. Beyer. "Getting vit in shape: Scaling laws for compute-optimal model design". In: *Advances in Neural Information Processing Systems* 36 (2024).

[11] V. V. Ramasesh, A. Lewkowycz, and E. Dyer. "Effect of scale on catastrophic forgetting in neural networks". In: *International Conference on Learning Representations*. 2021.

[12] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[13] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, et al. "Open catalyst 2020 (OC20) dataset and community challenges". In: *Acs Catalysis* 11.10 (2021), pp. 6059–6072.

[14] N. C. Frey, R. Soklaski, S. Axelrod, S. Samsi, R. Gomez-Bombarelli, C. W. Coley, and V. Gadepally. "Neural scaling of deep chemical models". In: *Nature Machine Intelligence* 5.11 (2023), pp. 1297–1305.

[15] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. "Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 30318–30332.

[16] X. Wei, Y. Zhang, X. Zhang, R. Gong, S. Zhang, Q. Zhang, F. Yu, and X. Liu. "Outlier suppression: Pushing the limit of low-bit transformer language models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17402–17414.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[18] S. Zhai, T. Likhomanenko, E. Littwin, D. Busbridge, J. Ramapuram, Y. Zhang, J. Gu, and J. M. Susskind. "Stabilizing transformer training by preventing attention entropy collapse". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 40770–40803.

[19] F. Schneider, Z. Nado, N. Agarwal, G. E. Dahl, and P. Hennig. *HITY workshop poll, NeurIPS 2022*. https://github.com/fsschneider/HITYWorkshopPoll. 2022.

[20] A. J. Fetterman, E. Kitanidis, J. Albrecht, Z. Polizzi, B. Fogelman, M. Knutins, B. Wróblewski, J. B. Simon, and K. Qiu. "Tune As You Scale: Hyperparameter Optimization For Compute Efficient Training". In: *arXiv preprint arXiv:2306.08055* (2023).

[21] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao. "Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer". In: *arXiv preprint arXiv:2203.03466* (2022).

[22] L. Chizat and P. Netrapalli. "Steering Deep Feature Learning with Backward Aligned Feature Updates". In: *arXiv preprint arXiv:2311.18718* (2023).

[23] A. Bietti, L. Venturi, and J. Bruna. "On the sample complexity of learning under geometric stability". In: *Advances in neural information processing systems* 34 (2021), pp. 18673–18684.

[24] L. Xiao, J. Pennington, and S. Schoenholz. "Disentangling trainability and generalization in deep neural networks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10462–10472.

[25] W. Hu, L. Xiao, B. Adlam, and J. Pennington. "The surprising simplicity of the early-time learning dynamics of neural networks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17116–17128.

[26] N. Tsilivis and J. Kempe. "What can the neural tangent kernel tell us about adversarial robustness?" In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 18116–18130.

[27] R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. "Bayesian deep convolutional networks with many channels are gaussian processes". In: *arXiv preprint arXiv:1810.05148* (2018).

[28] J. Hron, Y. Bahri, J. Sohl-Dickstein, and R. Novak. "Infinite attention: NNGP and NTK for deep attention networks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4376–4386.

[29] H. Fu, T. Guo, Y. Bai, and S. Mei. "What can a single attention layer learn? a study through the random features lens". In: *Advances in Neural Information Processing Systems* 36 (2024).

[30] I. Lavie, G. Gur-Ari, and Z. Ringel. "Towards Understanding Inductive Bias in Transformers: A View From Infinity". In: *arXiv preprint arXiv:2402.05173* (2024).

[31] L. Noci, S. Anagnostidis, L. Biggio, A. Orvieto, S. P. Singh, and A. Lucchi. "Signal propagation in transformers: Theoretical perspectives and the role of rank collapse". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27198–27211.

[32] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau. "Stable resnet". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1324–1332.

[33] G. Yang, D. Yu, C. Zhu, and S. Hayou. "Tensor Programs VI: Feature Learning in Infinite-Depth Neural Networks". In: *arXiv preprint arXiv:2310.02244* (2023).

[34] M. Li, M. Nica, and D. Roy. "The neural covariance SDE: Shaped infinite depth-and-width networks at initialization". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 10795–10808.

[35] L. Noci, C. Li, M. Li, B. He, T. Hofmann, C. J. Maddison, and D. Roy. "The shaped transformer: Attention models in the infinite depth-and-width limit". In: *Advances in Neural Information Processing Systems* 36 (2024).

[36] R. M. Neal. "Bayesian Learning for Neural Networks". PhD thesis. University of Toronto, 1995.

[37] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. "Deep neural networks as gaussian processes". In: *arXiv preprint arXiv:1711.00165* (2017).

[38] A. Jacot, F. Gabriel, and C. Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018).

[39] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. "Wide neural networks of any depth evolve as linear models under gradient descent". In: *Advances in neural information processing systems* 32 (2019).

[40] L. Chizat, E. Oyallon, and F. Bach. "On lazy training in differentiable programming". In: *Advances in neural information processing systems* 32 (2019).

[41] G. Yang. "Tensor programs ii: Neural tangent kernel for any architecture". In: *arXiv preprint arXiv:2006.14548* (2020).

[42] B. Hanin. "Random neural networks in the infinite width limit as Gaussian processes". In: *The Annals of Applied Probability* 33.6A (2023), pp. 4798–4819.

[43] B. He. "On kernel and feature learning in neural networks". PhD thesis. University of Oxford, 2022.

[44] S. Mei, T. Misiakiewicz, and A. Montanari. "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit". In: *Conference on Learning Theory*. PMLR. 2019, pp. 2388–2464.

[45] G. Yang and E. J. Hu. "Tensor programs iv: Feature learning in infinite-width neural networks". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 11727–11737.

[46] B. Bordelon and C. Pehlevan. "Self-consistent dynamical field theory of kernel evolution in wide neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 32240–32256.

[47] L. Chizat, M. Colombo, X. Fernández-Real, and A. Figalli. "Infinite-width limit of deep linear neural networks". In: *Communications on Pure and Applied Mathematics* 77.10 (2024), pp. 3958–4007.

[48] A. Bietti and F. Bach. "Deep equals shallow for ReLU networks in kernel regimes". In: *arXiv preprint arXiv:2009.14397* (2020).

[49] L. Noci, G. Bachmann, K. Roth, S. Nowozin, and T. Hofmann. "Precise characterization of the prior predictive distribution of deep ReLU networks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20851–20862.

[50] S. Hayou, A. Doucet, and J. Rousseau. "The curse of depth in kernel regime". In: *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*. PMLR. 2022, pp. 41–47.

[51] B. Hanin and M. Nica. "Finite depth and width corrections to the neural tangent kernel". In: *arXiv preprint arXiv:1909.05989* (2019).

[52] B. Hanin and M. Nica. "Products of many large random matrices and gradients in deep neural networks". In: *Communications in Mathematical Physics* 376.1 (2020), pp. 287–322.

[53] D. A. Roberts, S. Yaida, and B. Hanin. *The principles of deep learning theory*. Vol. 46. Cambridge University Press Cambridge, MA, USA, 2022.

[54] B. Bordelon, L. Noci, M. B. Li, B. Hanin, and C. Pehlevan. "Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit". In: *arXiv preprint arXiv:2309.16620* (2023).

[55] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. "Deep information propagation". In: *arXiv preprint arXiv:1611.01232* (2016).

[56] B. Hanin. "Which neural net architectures give rise to exploding and vanishing gradients?" In: *Advances in neural information processing systems* 31 (2018).

[57] J. Martens, A. Ballard, G. Desjardins, G. Swirszcz, V. Dalibard, J. Sohl-Dickstein, and S. S. Schoenholz. "Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping". In: *arXiv preprint arXiv:2110.01765* (2021).

[58] S. Hayou, A. Doucet, and J. Rousseau. "On the impact of the activation function on deep neural networks training". In: *International conference on machine learning*. PMLR. 2019, pp. 2672–2680.

[59] S. Hayou, A. Doucet, and J. Rousseau. "On the selection of initialization and activation function for deep neural networks". In: *arXiv preprint arXiv:1805.08266* (2018).

[60] Y. Dong, J.-B. Cordonnier, and A. Loukas. "Attention is not all you need: Pure attention loses rank doubly exponentially with depth". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2793–2803.

[61] K. He, X. Zhang, S. Ren, and J. Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[62] B. Hanin and A. Zlokapa. "Bayesian interpolation with deep linear networks". In: *Proceedings of the National Academy of Sciences* 120.23 (2023), e2301345120.

[63] B. Adlam and J. Pennington. "The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 74–84.

[64] M. Celentano, C. Cheng, and A. Montanari. "The high-dimensional asymptotics of first order methods with random data". In: *arXiv preprint arXiv:2112.07572* (2021).

[65] A. Montanari and Y. Zhong. "The interpolation phase transition in neural networks: Memorization and generalization under lazy training". In: *The Annals of Statistics* 50.5 (2022), pp. 2816–2847.

[66] S. Mei and A. Montanari. "The generalization error of random features regression: Precise asymptotics and the double descent curve". In: *Communications on Pure and Applied Mathematics* 75.4 (2022), pp. 667–766.

[67] L. Xiao, H. Hu, T. Misiakiewicz, Y. Lu, and J. Pennington. "Precise learning curves and higher-order scalings for dot-product kernel regression". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 4558–4570.

[68] H. Cui, F. Krzakala, and L. Zdeborová. "Bayes-optimal learning of deep random networks of extensive-width". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 6468–6521.

[69] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. "High-dimensional asymptotics of feature learning: How one gradient step improves the representation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 37932–37946.

[70] H. Cui, L. Pesce, Y. Dandi, F. Krzakala, Y. M. Lu, L. Zdeborová, and B. Loureiro. "Asymptotics of feature learning in two-layer networks after one gradient-step". In: *arXiv preprint arXiv:2402.04980* (2024).

[71] C. Meijer. "Über whittakersche bzw. besselsche funktionen und deren produkte". In: *Nieuw Archief voor Wiskunde* 18.2 (1936), pp. 10–29.

[72] S. Hayou and G. Yang. "Width and Depth Limits Commute in Residual Networks". In: *arXiv preprint arXiv:2302.00453* (2023).

[73] L. Noci, S. Anagnostidis, L. Biggio, A. Orvieto, S. P. Singh, and A. Lucchi. "Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse". In: *Advances in Neural Information Processing Systems*. 2022.

[74] M. Wortsman, P. J. Liu, L. Xiao, K. Everett, A. Alemi, B. Adlam, J. D. Co-Reyes, I. Gur, A. Kumar, R. Novak, et al. "Small-scale proxies for large-scale transformer training instabilities". In: *arXiv preprint arXiv:2309.14322* (2023).

[75] L. Noci, A. Meterez, T. Hofmann, and A. Orvieto. "Why do Learning Rates Transfer? Reconciling Optimization and Scaling Limits for Deep Learning". In: *arXiv preprint arXiv:2402.17457* (2024).

[76] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability". In: *arXiv preprint arXiv:2103.00065* (2021).

[77] L. Noci, K. Roth, G. Bachmann, S. Nowozin, and T. Hofmann. "Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12738–12748.

[78] G. Bachmann, L. Noci, and T. Hofmann. "How Tempering Fixes Data Augmentation in Bayesian Neural Networks". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 1244–1260.

[79] K. Lion, L. Noci, T. Hofmann, and G. Bachmann. "How Good is a Single Basin?" In: *arXiv preprint arXiv:2402.03187* (2024).

[80] S. Anagnostidis, D. Pavllo, L. Biggio, L. Noci, A. Lucchi, and T. Hofmann. "Dynamic context pruning for efficient and interpretable autoregressive transformers". In: *Advances in Neural Information Processing Systems* 36 (2024).

[81] B. He, L. Noci, D. Paliotta, I. Schlag, and T. Hofmann. "Understanding and Minimising Outlier Features in Transformer Training". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[82] L. Xiao and J. Pennington. "Synergy and symmetry in deep learning: Interactions between the data, model, and inference algorithm". In: *arXiv preprint arXiv:2207.04612* (2022).

[83] C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborova. "Rigorous dynamical mean-field theory for stochastic gradient descent methods". In: *SIAM Journal on Mathematics of Data Science* 6.2 (2024), pp. 400–427.

[84] T. Tao. *Topics in random matrix theory*. Vol. 132. American Mathematical Soc., 2012.

[85] F. J. Dyson. "A Brownian-motion model for the eigenvalues of a random matrix". In: *Journal of Mathematical Physics* 3.6 (1962), pp. 1191–1198.

[86] A. Guionnet, J. Ko, F. Krzakala, P. Mergny, and L. Zdeborová. "Spectral phase transitions in non-linear wigner spiked models". In: *arXiv preprint arXiv:2310.14055* (2023).