Contrasting random and learned features in deep Bayesian linear regression

Jacob A. Zavatone-Veth[1,2], **William L. Tong**[3], **Cengiz Pehlevan**[3,2]
[1]Department of Physics, [2]Center for Brain Science,
[3]John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, MA 02138
jzavatoneveth@g.harvard.edu, cpehlevan@seas.harvard.edu

Abstract

Understanding how feature learning affects generalization is among the foremost goals of modern deep learning theory. Here, we use the replica method from the statistical mechanics of disordered systems to study how the ability to learn representations affects the generalization performance of a simple class of models: deep Bayesian linear neural networks trained on unstructured Gaussian data. By comparing deep random feature models to deep networks in which all layers are trained, we provide a detailed characterization of the interplay between width, depth, data density, and prior mismatch. Random feature models can have particular widths that are optimal for generalization at a given data density, while making neural networks as wide or as narrow as possible is always optimal. Moreover, we show that the leading-order correction to the kernel-limit learning curve cannot distinguish between random feature models and deep networks in which all layers are trained. Taken together, our findings begin to elucidate how architectural details affect generalization performance in this simple class of deep regression models.

1 Introduction

Deep neural networks (NNs) display a rich and often-perplexing spectrum of generalization behaviors. Highly overparameterized NNs may possess the expressivity to fit random noise, yet in practice can still generalize well to unseen data [1, 2]. The ability of NNs to flexibly learn features from data is widely believed to be a critical contributor to their practical success [1–4], but the precise contributions of feature learning to their generalization behavior remain incompletely understood [1–10].

In recent years, intensive theoretical work has begun to elucidate the properties of deep networks in the limit of infinite hidden layer width, where inference in deep networks is equivalent to kernel regression or classification [6, 11–13, 13–16]. This correspondence has enabled detailed characterizations of inference at infinite width in both maximum-likelihood and fully Bayesian settings, providing new insights into the inductive biases that allow deep networks to overfit benignly [17–27].

Yet, understanding inference in the kernel limit is not sufficient, because kernel descriptions cannot capture feature learning [3, 7–9, 28]. As a result, a growing number of recent works have aimed to perturbatively study the behavior of networks near the kernel limit, with the hope that leading-order corrections to the large-width behavior might elucidate how width and depth affect inference [4, 29–36].

However, previous studies of Bayesian neural network generalization near the kernel limit have not clearly differentiated the effect of width on feature learning from its other potential effects on inference. Concretely, it is not clear whether potential improvements in generalization afforded by

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

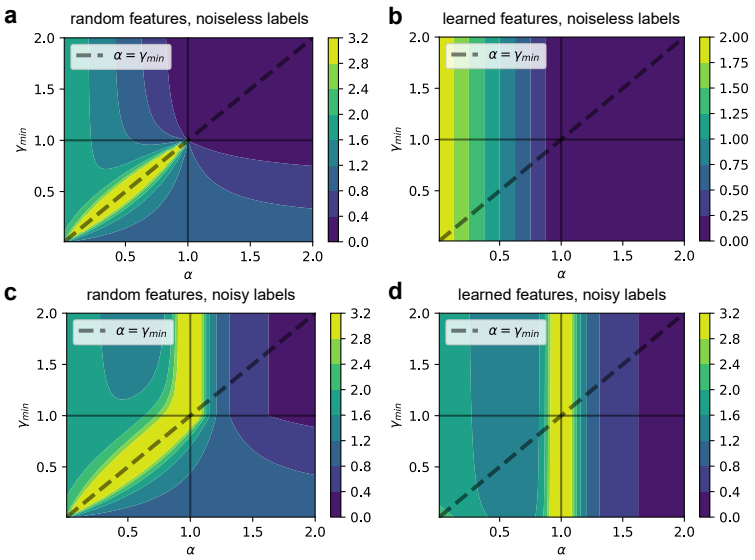

Figure 1: Generalization in deep Bayesian linear random feature models and neural networks. Here, we illustrate how representational flexibility eliminates the model-wise double descent that is present in random feature models with fewer features than inputs. **a**. Generalization error as a function of data density $\alpha = p/d$ and relative hidden layer width $\gamma = n/d$ in a random feature model of input dimension d and hidden layer width n trained on a dataset of p examples. When α and γ are less than one, the generalization error diverges at $\alpha = \gamma$. **b**. As in **a**, but for a network with all layers trained. Representational flexibility eliminates the divergence at $\alpha = \gamma$. **c**. As in **a**, but with noise-corrupted training labels. Now, an additional divergence is visible when $\alpha = 1$ and $\gamma > 1$. **d**. As in **b**, but with noise-corrupted training labels. The generalization error diverges when $\alpha = 1$.

the leading finite-width correction reflect the benefits of feature learning, or if a similar gain would be observed in random feature models, where only the readout layer is trained. Here, we explore how random and learned features affect generalization in the simplest class of Bayesian NNs—deep linear models—when trained on unstructured, noisy data. By developing a detailed understanding of this simple setting, one might hope to gain intuition that may prove useful in studying more complex networks [30, 37–42].

We have recently studied the asymptotic generalization performance of deep linear Bayesian regression for data generated with an isotropic Gaussian covariate model [43]. Using the replica trick [44–46], we compute learning curves for simple linear regression, deep linear Gaussian random feature (RF) models, and deep linear NNs. Using alternative replica-free methods and numerical simulation, we show that the predictions obtained under a replica-symmetric (RS) *Ansatz* are accurate for all three model classes.

In the presence of label noise, both RF and NN models display sample-wise non-monotonicity, which refer to as “double-descent," in their learning curves. As we work in a high-dimensional limit, this non-monontonicity is of a particularly extreme form: the generalization error diverges at a particular data density. If one introduces a bottleneck layer that is narrower than the input dimension, an RF model will display model-wise double-descent behavior at fixed data density—or equivalently sample-wise double-descent at fixed width—even in the absence of label noise, while an NN model will not show this divergence. This distinct small-width behavior shows one advantage afforded by the flexibility to learn features. We further analyze models of arbitrary depth perturbatively in the limit in which the network depth and dataset size are small relative to the hidden layer widths. We find that the leading order correction to the large-width behavior of RF and NN models is identical, hence first-order perturbation theory for the generalization error cannot distinguish between random and learned features. In total, our results provide new insight into how the generalization behavior of deep Bayesian linear regression in high dimensions depends on architectural details. Moreover, they shed light onto which qualitative features of generalization behavior can or cannot be captured by low-order perturbative corrections [30].

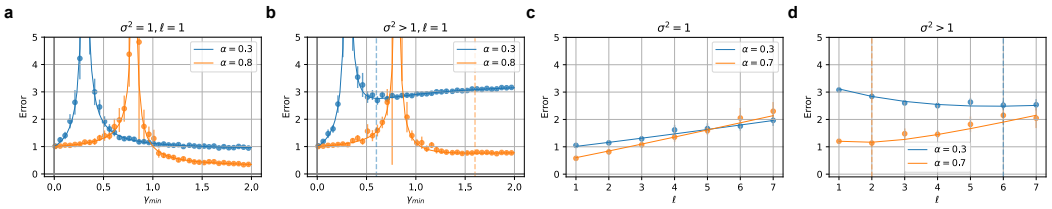

Figure 2: Optimal RF model architecture depends on target-prior mismatch. **(a).** Error across widths for a single-hidden-layer RF model with prior variance $\sigma^2 = 1$. **(b).** As in (a)., but for a single-hidden-layer RF model with higher prior variance ($\sigma^2 = 4$). Theoretical predictions for optimal width are marked with dashed vertical lines for each α. **(c).** Error across depths for prior variance $\sigma^2 = 1$ and fixed width $\gamma = 1.5$ **(d).** Error across depths for prior variance $\sigma^2 = 4$ and fixed width $\gamma = 1.5$. Theoretical predictions for optimal depth are marked with dashed vertical lines for each α.

2 Results

We consider deep linear models $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}/\sqrt{d}$ for an end-to-end weight vector $\mathbf{w} = \sigma \mathbf{U}_1 \cdots \mathbf{U}_\ell \mathbf{v}/\sqrt{n_1 \cdots n_\ell}$, where $\mathbf{U}_l \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\sigma > 0$ sets the predictor scale. The objective of our work is to compare RF models, in which the matrices $\mathbf{U}_l$ are fixed and random and only the readout $\mathbf{v}$ is trained, with NNs, in which all parameters are trained. We fix isotropic Gaussian priors $(U_l)_{ij} \sim \mathcal{N}(0,1)$ and $v_j \sim \mathcal{N}(0,1)$. We train these models on a dataset $\{(\mathbf{x}_\mu, y_\mu)\}_{\mu=1}^p$ generated by a noisy Gaussian covariate model, with $\mathbf{x}_\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y_\mu = \mathbf{w}_*^\top \mathbf{x}_\mu/\sqrt{d} + \xi_\mu$, where $\|\mathbf{w}_*\|^2 = d$ and $\xi_\mu \sim \mathcal{N}(0, \eta^2)$. We introduce an isotropic Gaussian likelihood of variance $1/\beta$, and denote expectations with respect to the resulting Bayes posterior by $\langle \cdot \rangle$. Using the non-rigorous replica method from statistical physics [44–46], we compute the asymptotic zero-temperature average generalization error $\epsilon = \lim_{\beta \to \infty} \lim_{d,p,n_1,\ldots,n_\ell \to \infty} \mathbb{E}_{\mathcal{D}} \langle \|\mathbf{w} - \mathbf{w}_*\|^2 \rangle / d$ in the proportional limit with $p/d \to \alpha$ and $n_l/d \to \gamma_l$, where $\mathbb{E}_{\mathcal{D}}$ denotes expectation over the random data and, for the RF model, the random weights. The details of this calculation are lengthy and technical, and are presented in [43].

For RF models, we obtain a closed-form expression for the learning curve at any depth. Letting $\gamma_{\min} = \min\{\gamma_1, \ldots, \gamma_\ell\}$ be the minimum hidden layer width, we find that

$$\epsilon_{\text{RF}} = \begin{cases} (1-\alpha)\left(1 + \sigma^2 \prod_{l=1}^{\ell} \frac{\gamma_l - \alpha}{\gamma_l} + \sum_{l=1}^{\ell} \frac{\alpha}{\gamma_l - \alpha}\right) & \\ \quad + \left(\frac{\alpha}{1-\alpha} + \sum_{l=1}^{\ell} \frac{\alpha}{\gamma_l - \alpha}\right) \eta^2, & \text{if } \alpha < \min\{1, \gamma_{\min}\} \\ \alpha \frac{1-\gamma_{\min}}{\alpha - \gamma_{\min}} + \frac{\gamma_{\min}}{\alpha - \gamma_{\min}} \eta^2, & \text{if } \alpha > \gamma_{\min} \text{ and } \gamma_{\min} < 1 \\ \frac{1}{\alpha - 1} \eta^2, & \text{if } \alpha > 1 \text{ and } \gamma_{\min} > 1. \end{cases} \tag{1}$$

The corresponding result ϵ_{LR} for simple linear regression can be obtained by setting $\ell = 0$, which recovers previous results [20, 47]. For a deep network, we find that

$$\epsilon_{\text{NN}} = \epsilon_{\text{LR}} + \begin{cases} z - \sigma^2(1-\alpha), & \text{if } \alpha < 1 \\ 0, & \text{if } \alpha > 1, \end{cases} \tag{2}$$

where $z = z(\alpha, \sigma^2, \eta^2, \gamma_1, \ldots, \gamma_\ell)$ is a non-negative real root of the polynomial

$$z^{\ell+1} = \sigma^2(1-\alpha) \prod_{l=1}^{\ell} \left[(\gamma_l - \alpha) z/\gamma_l + \alpha(1 - \alpha + \eta^2)/\gamma_l\right]. \tag{3}$$

The detailed conditions on which root should be selected are given in [43]; this result is consistent with that of a heuristic approximation by [35]. We note immediately that, in the absence of label noise ($\eta = 0$), simple linear regression and NNs do not display double-descent, while RF models can still display divergent generalization error if $\gamma_{\min} < 1$. This is illustrated in Figure 1.

At large widths $\gamma_l \to \infty$, both ϵ_{RF} and ϵ_{NN} tend to ϵ_{LR}, reflecting the fact that the deep network does not learn features in this limit [4, 13, 14, 16, 31, 33]. If $\sigma^2 < 1 + \eta^2/(1-\alpha)$, wider (and shallower) RF and NN models always generalize better. If $\sigma^2 > 1 + \eta^2/(1-\alpha)$, narrower (and deeper) NN models always generalize better, while RF models have an optimal width and depth. We illustrate

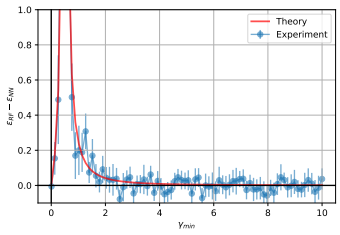

Figure 3: Generalization gap between two-layer RF and NN models as a function of hidden layer width γ_{min}. The details of the numerical experiments will be provided elsewhere.

optimal RF model architectures in Figure 2. Optimal network architecture therefore depends on the match between the prior and target scales.

Solving the polynomial equation for ϵ_{NN} in order-by-order in α/γ, we recover the dataset average of the leading-order perturbative correction to this limit we previously computed in [33]:

$$\epsilon_{\text{NN}} = \epsilon_{\text{LR}} + [(1-\alpha)(1-\sigma^2)+\eta^2] \sum_{l=1}^{\ell} \frac{\alpha}{\gamma_l} + \mathcal{O}\left(\frac{\alpha^2}{\gamma^2}\right). \tag{4}$$

However, upon expanding (1) one finds that the leading correction to ϵ_{RF} is identical, hence leading-order perturbation theory cannot distinguish the effect of representational flexibility from other finite-width effects. In particular, the gap in generalization for equal widths $\gamma_1 = \cdots = \gamma_\ell = \gamma$ is

$$\frac{\epsilon_{\text{RF}} - \epsilon_{\text{NN}}}{1-\alpha+\eta^2} = \frac{\ell(\ell+1)}{2\tilde{\sigma}^2}\frac{\alpha^2}{\gamma^2} + \mathcal{O}\left(\frac{\alpha^3}{\gamma^3}\right), \tag{5}$$

for $\tilde{\sigma}^2 = \sigma^2/[1+\eta^2/(1-\alpha)]$. The leading term in this expansion is positive, hence at very large widths training both layers should produce a small benefit relative to simply training the readout. For $\ell = 1$, one can show that $\epsilon_{\text{RF}} \geq \epsilon_{\text{NN}}$ at any width, with equality iff $\alpha = 0$ or $\gamma_1 \to \infty$. This behavior, and the excellent agreement of our theory with numerical experiment, is illustrated in Figure 3.

3 Conclusion

We have characterized how representational flexibility affects generalization performance in deep linear Bayesian regression models. We showed that mismatch in the prior and target scales determines when wider models generalize better, and that representational flexibility has a subleading effect on generalization at large widths.

We conclude by noting that our work has several important limitations beyond the non-rigorous nature of the replica method, which will be interesting to address in future work. First, our approach is highly specialized to deep linear networks, and would not extend easily to nonlinear models. Though the utility of linear networks as a model system for studying the effect of depth on inference has been clearly established [33, 35, 37, 39], rigorous characterization of the effect of nonlinearity on inference in deep Bayesian neural networks remains a largely open problem [4, 30, 31, 33, 35, 36, 48]. Second, we have assumed that the covariates are drawn from an isotropic Gaussian distribution. Though this is a standard generative model in theoretical studies of inference [21, 38, 39, 47], it is undoubtedly not reflective of real-world data. Extending results of this form to more realistic generative models will be an interesting objective for future work [20, 49]. Finally, while BNNs are finding practical applications in physics and elsewhere [50], another important direction for future work will be to develop a rigorous theoretical understanding of how results on the generalization performance of BNNs, like those obtained here, relate to the generalization performance of networks trained with stochastic gradient-based algorithms, a link that remains incompletely understood [40, 42, 51, 52].

4 Broader impacts

As our work is purely theoretical, we do not anticipate that it will have direct societal impacts.

Acknowledgments and Disclosure of Funding

This work was supported by a Google Faculty Research Award and NSF DMS-2134157.

References

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115, 2021. doi: 10.1145/3446776.

[2] 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. doi: https://doi.org/10.1073/pnas.1903070116.

[3] 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. URL https://proceedings.mlr.press/v139/yang21c.html.

[4] Laurence Aitchison. Why bigger is not always better: on finite and infinite 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 156–164. PMLR, July 2020. URL http://proceedings.mlr.press/v119/aitchison20a.html.

[5] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021. doi: https://doi.org/10.1088/1742-5468/ac3a74.

[6] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[7] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborova. Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed. 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 8936–8947. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/refinetti21b.html.

[8] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

[9] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, nov 2020. doi: 10.1088/1742-5468/abc4de. URL https://doi.org/10.1088/1742-5468/abc4de.

[10] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

[11] Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996. doi: https://doi.org/10.1007/978-1-4612-0745-0_2.

[12] Christopher KI Williams. Computing with infinite networks. *Advances in Neural Information Processing Systems*, pages 295–301, 1997. URL https://papers.nips.cc/paper/1996/hash/ae5e3ce40e0404a45ecacaaf05e5f735-Abstract.html.

[13] 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. URL https://openreview.net/forum?id=B1EA-M-0Z.

[14] 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. URL https://openreview.net/forum?id=H1-nGgWC-.

[15] Arthur Jacot, Franck Gabriel, and Clément 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, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

[16] Jiri Hron, Yasaman Bahri, Roman Novak, Jeffrey Pennington, and Jascha Sohl-Dickstein. Exact posterior distributions of wide Bayesian neural networks. *arXiv preprint arXiv:2006.10541*, 2020.

[17] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75, 2019. doi: 10.1002/cpa.22008.

[18] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

[19] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher-student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, dec 2020. doi: 10.1088/1742-5468/abc61d.

[20] 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. doi: https://doi.org/10.1038/s41467-021-23103-1.

[21] Jean Barbier, Wei-Kuo Chen, Dmitry Panchenko, and Manuel Sáenz. Performance of Bayesian linear regression in a model with mismatch. *arXiv preprint arXiv:2107.06936*, 2021.

[22] Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where and why do they appear? *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124002, 2021. doi: https://doi.org/10.1088/1742-5468/ac3909.

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

[24] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 11022–11032, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/7d420e2b2939762031eed0447a9be19f-Abstract.html.

[25] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020. URL https://proceedings.mlr.press/v119/d-ascoli20a.html.

[26] Hui Jin, Pradeep Kr. Banerjee, and Guido Montufar. Learning curves for Gaussian process regression with power-law priors and targets. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KeI9E-gsoB.

[27] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Learning curves of generic features maps for realistic datasets with a teacher-student model. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9704a4fc48ae88598dcbdcdf57f3fdef-Abstract.html.

[28] Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In *Advances in Neural Information Processing Systems*, volume 33, pages 15156–15172, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ad086f59924fffe0773f8d0ca22ea712-Abstract.html.

[29] Sho Yaida. Non-Gaussian processes and neural networks at finite widths. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 165–192, Princeton University, Princeton, NJ, USA, July 2020. PMLR. URL http://proceedings.mlr.press/v107/yaida20a.html.

[30] Jacob A Zavatone-Veth and Cengiz Pehlevan. Exact marginal prior distributions of finite Bayesian neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1baff70e2669e8376347efd3a874a341-Abstract.html.

[31] Daniel A Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, 2022.

[32] James Halverson, Anindita Maiti, and Keegan Stoner. Neural networks and quantum field theory. *Machine Learning: Science and Technology*, 2(3), 2021. doi: https://doi.org/10.1088/2632-2153/abeca3.

[33] Jacob A Zavatone-Veth, Abdulkadir Canatar, Benjamin S Ruben, and Cengiz Pehlevan. Asymptotics of representation learning in finite Bayesian neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cf9dc5e4e194fc21f397b4cac9cc3ae9-Abstract.html.

[34] Jacob A Zavatone-Veth and Cengiz Pehlevan. Depth induces scale-averaging in overparameterized linear Bayesian neural networks. In *Asilomar Conference on Signals, Systems, and Computers*, volume 55, 2021. doi: 10.1109/IEEECONF53345.2021.9723137.

[35] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11:031059, 09 2021. doi: 10.1103/PhysRevX.11.031059.

[36] Gadi Naveh and Zohar Ringel. A self consistent theory of Gaussian processes captures feature learning effects in finite CNNs. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b24d21019de5e59da180f1661904f49a-Abstract.html.

[37] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[38] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[39] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. doi: https://doi.org/10.1016/j.neunet.2020.08.022.

[40] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc., 2020.

[41] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? 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 10248–10259. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wenzel20a.html.

[42] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? 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 4629–4640. PMLR, 07 2021.

[43] Jacob A. Zavatone-Veth, William L. Tong, and Cengiz Pehlevan. Contrasting random and learned features in deep Bayesian linear regression. *Physical Review E*, 105:064118, Jun 2022. doi: 10.1103/PhysRevE.105.064118. URL https://link.aps.org/doi/10.1103/PhysRevE.105.064118.

[44] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. World Scientific Publishing Company, 1987. doi: https://doi.org/10.1142/0271.

[45] Andreas Engel and Christian van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. doi: https://doi.org/10.1017/CBO9781139164542.

[46] Jacob A. Zavatone-Veth and Cengiz Pehlevan. Replica method for eigenvalues of real Wishart product matrices. *arXiv*, 2022. doi: 10.48550/ARXIV.2209.10499. URL https://arxiv.org/abs/2209.10499.

[47] Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992. doi: https://doi.org/10.1088/0305-4470/25/5/020.

[48] Jacob A Zavatone-Veth and Cengiz Pehlevan. Activation function dependence of the storage capacity of treelike neural networks. *Physical Review E*, 103(2):L020301, 2021. doi: https://doi.org/10.1103/PhysRevE.103.L020301.

[49] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10:041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044.

[50] Miles Cranmer, Daniel Tamayo, Hanno Rein, Peter Battaglia, Samuel Hadden, Philip J Armitage, Shirley Ho, and David N Spergel. A Bayesian neural network predicts the dissolution of compact planetary systems. *Proceedings of the National Academy of Sciences*, 118(40), 2021.

[51] Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is SGD a Bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(79):1–64, 2021. URL http://jmlr.org/papers/v22/20-676.html.

[52] Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1010–1022. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0b1ec366924b26fc98fa7b71a9c249cf-Paper.pdf.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] As noted in Section 4, we do not anticipate that our work will have negative social impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are given in this abstract and in the long-form version of our work [43].
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs of the stated results are provided in the long-form version of our work [43].
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]