A Unified Approach to Feature Learning in Bayesian Neural Networks

Noa Rubin NOA.RUBIN@MAIL.HUJI.AC.IL
Racah institute of physics, Hebrew University, Jerusalem, Israel

Zohar Ringel ZOHAROZ@GMAIL.COM
Racah institute of physics, Hebrew University, Jerusalem, Israel

Inbar Seroussi INBARSER@GMAIL.COM
School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel

Moritz Helias M.HELIAS@FZ-JUELICH.DE
Institute for Advanced Simulation (IAS-6), Jülich Research Centre, Jülich, Germany & Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany

Abstract

The power of neuronal networks comes from their adaptation to training data, known as feature learning. We consider feature learning within Bayesian learning and derive the two prominent high dimensional theories, kernel scaling and kernel adaptation, respectively, from a unified large deviation approach. We then show when feature learning escapes the scaling approach, but is captured by kernel adaptation.

1. Introduction

A central quest of the theory of deep learning is to understand the inductive bias of network architectures, which underlays their ability to find solutions that generalize well despite networks being highly overparametrized. One process that enables generalization is feature learning, a process where the network learns useful representations of the data. The success of deep learning is often attributed to this process. This is reflected, in part, by the performance gap between actual deep neural networks and their infinite-width Gaussian Process (GP) counterparts, [5, 7, 12, 13, 15, 21] where only a minimal change to the weights is observed [9]. Moreover, feature learning is necessary to understand transfer learning, the central mechanism that enables modern foundation models [4].

Despite the importance of feature learning, there is no consensus on how to analytically describe this process. The two prominent Bayesian approaches derived using statistical physics type analysis— kernel scaling [10, 14] and kernel adaptation [18] — lead to seemingly contradictory descriptions, while both make successful predictions. Our main contributions are as follows: **I.** We introduce a formalism of feature learning that unifies the two approaches based on large deviation theory. **II.** We perform a numerical comparison of the two approaches in mean-field scaling (MFS) [11] and standard scaling (SS) i.e. weights $\propto 1/\sqrt{\text{width}}$. **III.** We identify a feature-learning scenario captured by kernel adaptation, which kernel scaling fails to describe.

2. Unified Approach to Feature Learning

We start by presenting the posterior output distribution, which we show to be the common origin of the two theories. Consider a network with a single hidden layer

$$h = V\,x, \quad f = w^{\mathrm{T}}\phi(h), \quad y = f + \xi\,, \tag{1}$$

where ϕ is a point-wise applied activation, $V \in \mathbb{R}^{M\times d}, w \in \mathbb{R}^M$ function and ξ is a Gaussian noise $\xi_\alpha \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \kappa M^{1-\gamma})$. Here $\boldsymbol{x} \in \mathbb{R}^d$ is the feature vector and $f \in \mathbb{R}$ is the scalar output and we consider n tuples of training data $\mathcal{D} = \{(x_\alpha, y_\alpha)\}_{1\le\alpha\le n}$. The data points $\{x_\alpha\}_{\alpha=1}^n$ are combined in the matrix $X \in \mathbb{R}^{n\times d}$. We assume Gaussian i.i.d. priors $V_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_V/d)$ and $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_w/M^\gamma)$ for all trainable weights and consider two cases, $\gamma = 1$, which we denote as standard scaling (SS) and $\gamma = 2$, which we denote as mean-field scaling (MFS). We will be interested in the posterior distribution of the outputs f after conditioning on the training data $\mathcal{D}$ in the proportional limit where $n \propto M \to \infty$ at fixed ratio n/M. We note that both theories can be extended to the deep case; here we focus on single hidden layer networks to clarify the comparison.

The joint prior distribution of the readouts $f = \{f_\alpha\}_{1\le\alpha\le n}$ and the labels $y = \{y_\alpha\}_{1\le\alpha\le n}$ follows from standard manipulations, as

$$p(y, f|C^{(xx)}) = \mathcal{N}(y|f, \kappa M^{1-\gamma}) \int \mathcal{D}\tilde{f}\, \exp\big(-f^{\mathrm{T}}\tilde{f} + W(\tilde{f}|C^{(xx)})\big)\,, \tag{2}$$

$$W(\tilde{f}|C^{(xx)}) = \ln\Big\langle \exp\Big(\sum_{\alpha=1}^n \tilde{f}_\alpha \sum_{i=1}^M w_i\,\phi(h_{\alpha i})\Big)\Big\rangle_{w_i, h_{\alpha i}}, \tag{3}$$

where the i.i.d. distribution of the input weights V_{ij} implies that $h_{\alpha i} \overset{\text{i.i.d. over } i}{\sim} \mathcal{N}(0, C^{(xx)})$, where $\mathbb{R}^{n\times n} \ni C^{(xx)} = g_V d^{-1}\, XX^{\mathrm{T}}$. We will show that both the kernel scaling approach and the kernel adaptation approach, follow from (3) noting that the posterior $p(f|y, C^{(xx)}) \propto p(y, f|C^{(xx)})$ and employing a large deviation principle. The difference between approaches is the order in which the expectation over $h_{\alpha i}$ and w_i is taken, resulting in two different order parameters, the kernel scale, or the network output, respectively. In this work, we are interested in comparing the predictions of each theory regarding the mean posterior discrepancy defined as

$$\langle\Delta_\alpha\rangle := y_\alpha - \langle f_\alpha\rangle_{p(f|y,C^{(xx)})} = -\kappa M^{1-\gamma}\,\frac{\partial}{\partial y_\alpha}\,\ln p(y|C^{(xx)}) = \kappa M^{1-\gamma}\langle\tilde{f}_\alpha\rangle, \tag{4}$$

where $p(y|C^{(xx)}) = \int df p(y, f|C^{(xx)})$ is the marginal over f.

2.1. Scaling Approach

For the special case of $\phi(h) = h$, by averaging first over the preactivations in eq. (3), W is given by the cumulant-generating function of a Gaussian- $W(\tilde{f}|w) = 1/2\,\tilde{f}^{\mathrm{T}}C^{(xx)}\tilde{f}\,\|w\|^2$. This shows that the readout weights only appear in the form of the squared norm – we hence define $Q := M^{\gamma-1}\|w\|^2$. By performing a change of variables, the posterior distribution can be rewritten as (see App. A.1)

$$p(y|C^{(xx)}) = \int dQ\,\mathcal{N}(y|0, M^{1-\gamma}QC^{(xx)} + \kappa M^{1-\gamma}\mathbb{I})\,p(Q)\,, \tag{5}$$

where $p(Q) = \left\langle \delta[-Q + \|w\|^2] \right\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{g_w}{M})}$. As detailed in App. (A.1), $p(Q)$ can be approximated using the large deviation approach with Q as an order parameter so that- $-\ln p(Q) \simeq \Gamma(Q)$, with $\Gamma(Q)$ being a rate function defined in eq. (13). Intuitively, the large deviation approach corresponds to a saddle point approximation, which is possible because Q concentrates. Finally, because the two terms in $S(Q) := \ln \mathcal{N}(y|0, M^{1-\gamma} Q C^{(xx)} + \kappa M^{1-\gamma} \mathbb{I}) - \Gamma(Q)$ are proportional to n and M, respectively, the integral over Q in the proportional limit $n, M \to \infty$ is dominated by the maximum Q^* of S, resulting in the average discrepancy

$$\langle \Delta \rangle = \kappa \left(Q^* \, C^{(xx)} + \kappa \mathbb{I} \right)^{-1} y \, . \tag{6}$$

As a consequence, this feature learning theory [10] recovers the same mean discrepancy as the GP limit [12], but with a kernel scaled by the factor Q^*/g_w. We comment that for a single hidden layer linear network in SS the expression for Q^* can be derived explicitly [17], however, the derivation brought here can be extended to deep networks as well where the explicit derivation does not hold.

For a non-linear activation function, the result by [1] can be recovered if one performs an additional cumulant expansion of eq. (3) up to Gaussian order $W(\tilde{f}|w) = \frac{1}{2} \sum_{\alpha\beta} \tilde{f}_\alpha \, C^{(\phi\phi)}_{\alpha\beta} \, \tilde{f}_\beta \sum_i w_i^2 + \mathcal{O}(\tilde{f}^4)$, where $C^{(\phi\phi)}_{\alpha\beta} := \langle \phi(h_\alpha)\phi(h_\beta) \rangle_{h_\alpha \sim \mathcal{N}(0, C^{(xx)})}$ and where we used the pairwise independence of the $\phi(h_{\alpha i})$ across i. This approximation corresponds to stating that the ϕ_α be jointly Gaussian distributed. In brief, we recover the action Eq. (33) in [1] from [10] by replacing $C^{(xx)}$ by $C^{(\phi\phi)}$ in the equations for the linear activations, showing the tight relation to the case of linear networks. However, from this presentation, it is unclear when this approximation is accurate, since the next to leading order terms can be $\propto O(1)$ for $n \propto M$ due to correlations in the data summation. In certain cases, a Gaussian equivalence property justifies this step [3] (as discussed in App. A.1.1).

2.2. Adaptive Kernel Approach

In the case of a linear network in mean-field scaling $\gamma = 2$, the cumulant-generating function $W(\tilde{f}|C^{(xx)})$ can be computed by taking the expectation with respect to w first, and then with h,

$$-2W(\tilde{f}|C^{(xx)})/M = \ln\left(1 - g_w/M^2 \, \tilde{f}^{\mathrm{T}} C^{(xx)} \tilde{f}\right) = \ln\det\left(\left[[C^{(xx)}]^{-1} - g_w/M^2 \, \tilde{f}\tilde{f}^{\mathrm{T}}\right]\right), \tag{7}$$

which shows that $W(M\circ)/M$ as a function of $\circ$ is independent of M, having scaling form, indicating that the network readout f concentrates; formally, a large deviation approach with respect to the parameter $\tilde{f}$ yields the prior distribution from (2)

$$-\ln \, p(y|C^{(xx)}) \simeq \sup_{\tilde{f}} \{ -\tilde{f}^{\mathrm{T}} y - \frac{\kappa}{2M} \tilde{f}^{\mathrm{T}} \tilde{f} - W(\tilde{f}|C^{(xx)}) \} \, , \tag{8}$$

where the quadratic term $\tilde{f}^{\mathrm{T}} \tilde{f}$ comes from the average over the noise. Now fixing y by the training labels, the $\tilde{f}^*$ that satisfies the supremum condition in eq. (8) yields the mean posterior discrepancy (4) given by $\langle \Delta \rangle = \kappa/M \langle \tilde{f} \rangle = \kappa/M \tilde{f}^*$

$$\langle \Delta \rangle = \kappa \left(Q C^{(xx)} + \kappa \, \mathbb{I} \right)^{-1} y = \kappa \left(\kappa \mathbb{I} + g_w \left([C^{(xx)}]^{-1} - g_w/M^2 \, \tilde{f}^* \, \tilde{f}^{*\mathrm{T}} \right)^{-1} \right)^{-1} y \, , \tag{9}$$

where the latter two expressions come from the two different forms (7) for W, and $Q^{-1} = g_w^{-1} - \tilde{f}^{*\mathrm{T}} C^{(xx)} \tilde{f}^*/M^2$. The result is similar to the NNGP prediction, but with a kernel that is rescaled by Q/g_w, or, correspondingly, changed into a rank-one direction $\tilde{f}^* \tilde{f}^{*\mathrm{T}}$.

In SS ($\gamma = 1$) the cumulant-generating function (3) does not possess a scaling form, implying that fluctuations of f and hence fluctuations of the discrepancies become important. To treat these systematically, we define the effective action $\gamma(\tilde{f}_*) := \sup_y\{-y^{\mathrm{T}}\tilde{f}_* - \ln p(y|C^{(xx)}\}$, such that $\partial\gamma(\tilde{f}_*)/\partial\tilde{f}_* = -y$. This action expanded to include the leading order fluctuation corrections [8] takes the form, with W'' being the hessian,

$$\gamma(\tilde{f}_*) = -\frac{\kappa}{2}\tilde{f}^{*\mathrm{T}}\tilde{f}^* - W(\tilde{f}_*) + \frac{1}{2}\ln\det\left(\kappa + W''(\tilde{f}_*)\right). \tag{10}$$

Explicit expressions for this implicit equation are given in App. (A.2.4).

For nonlinear networks, the same large deviations approach is applied both in standard and mean field scaling, see App. (A.2.3) and App. (A.2.1) for further details. Crucially, since the adaptive approach tracks many parameters, this allows it to capture richer feature learning effects than in the scaling picture, such as Grokking [16], and reduction in sample complexity described in section 3.

2.3. Experimental Results

For a linear network, the kernel scaling theory and the adaptive theory can be made to agree regarding the output statistics ($\langle\Delta\rangle$). This is true for mean-field scaling at leading order, and in standard scaling by including fluctuation corrections, as detailed in App. (A.2.4). We demonstrate this for a single hidden layer **linear** network trained on a **linear** target function. Fig. (1) shows the accuracy of the theories compared to the experiment for both SS and MFS, for different values of κ, M and $n = d = 300, g_v = g_w = 1$. In standard scaling, the fluctuation corrections to the adaptive approach improve its accuracy, closely matching it to the scaling approach, and also in MFS both theories are in close agreement.

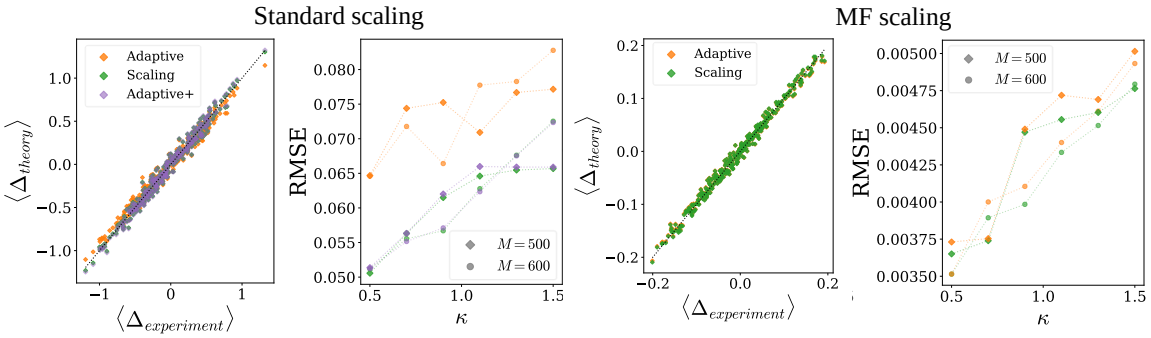

Figure 1: Comparison between numerics and theoretical predictions for $\langle\Delta\rangle$ ("adaptive" approach (9) / "adaptive+" with fluctuation corrections (46) / "scaling" approach (6)), each for standard scaling and for mean-field scaling (right) for different combinations of regularization noise κ and network width M. Scatter plots for $\kappa = 1$ and $M = 500$, RMSE $= \sqrt{\|(\langle\Delta^{\text{th.}}\rangle - \langle\Delta^{\text{exp.}}\rangle\|^2/n}$.

3. Sample Complexity Reduction in Nonlinear Networks

Even though the two theories predict the same result for linear networks, the fundamental difference in their description of feature learning has significant repercussions. In particular, we here show that

for a nonlinear network, the presence of strong feature learning results in a dimensional reduction which in turn leads to a reduction in sample complexity. Since the scaling approach considers only a scalar order parameter it cannot address such strong feature learning effects, and consequentially fails to capture the reduction in the complexity. This is in contrast to the adaptive approach, where the high dimensional order parameter allows for a richer picture of feature learning and the resulting sample complexity reduction.

3.1. Model

Here we consider a two-layer neural network as defined in eq. (2) with ReLU activation, trained on a target function $y(x) = v^{*\mathrm{T}}x + \epsilon\left(|v^{*\mathrm{T}}x| - \sqrt{2/\pi}\right)$, where v^* is normalized. We introduce the sample complexity measure, p-learnability ratio ($\mathcal{R}_p$) which is defined as the fraction of the p-th component of the target that was learned: $\mathcal{R}_p = \frac{\overline{f}(X)\cdot H_p(Xv^*)}{y(X)\cdot H_p(Xv^*)}$ where H_p is the p-th Hermite polynomial. Here $\mathcal{R}_p = 0$ indicates no learning of the p-th component and $\mathcal{R}_p = 1$ perfect learning. We note that this measure is different from the notion of information exponent [2], since it measures the learnability of higher order nonlinear components and it can also be defined per dataset. We use the kernel scaling theory [1] re-derived in Sec. 2.1. For the adaptive approach, we apply additional approximations as described in App. (B.1). This approach requires solving a single nonlinear equation, making it similar to the scaling approach in terms of computational complexity, yet because the kernel adapts, it nonetheless captures the qualitative behaviour of the system.

3.2. Experimental Results

Fig. (2) comparses the experimental learnability ratios to the predictions of the scaling approach, and to an approximate solution of the adaptive approach, for different values of β scaling n, M, d, where the value of the learnability ratio is averaged over datasets. The accuracy of the approximate adaptive approach improves with β, as expected from approching the proportional limit. While both, the scaling approach and the GP approximation, predict $\mathcal{R}_2 = 0$, this contradicts the experiment. The adaptive approach on the other hand correctly predicts the value of $\mathcal{R}_2$. For further details, see App. (B.2).

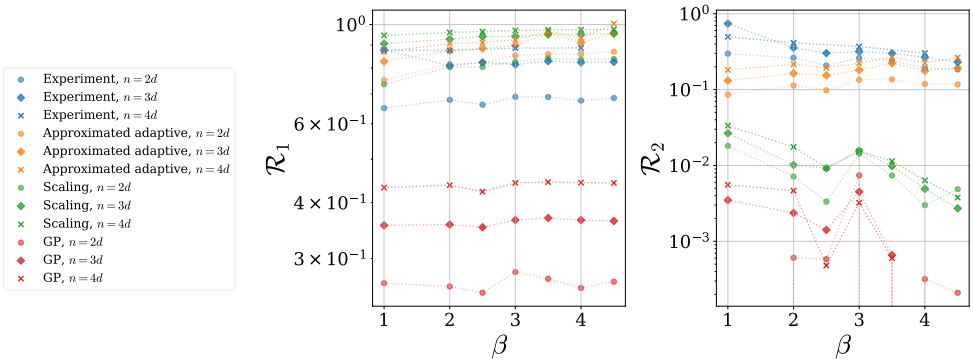

Figure 2: Learnability ratios $\mathcal{R}_{1,2}$ predicted by the approximate adaptive approach (orange (54)), scaling approach (green (6)), GP (red) compared to experimental values (blue), as a function of the scaling factor β, for different values of n/d.

4. Conclusion

We compare two statistical physics theories of feature learning, the kernel scaling approach, where the NNGP kernel is rescaled by a single self-consistently determined scalar, and the kernel adaptation approach, where the kernel changes more flexibly. We present a unified derivation of both theories from the same mechanistic starting point of the Bayesian posterior distribution, each following in the proportional limit from the application of ideas from large deviation theory to the kernel scaling variable or to the network readout as order parameters, respectively. We show that for a linear network in MFS, the two approaches yield consistent results. In SS, both theories agree if leading-order fluctuation corrections of the discrepancies are taken into account in the adaptive approach. For a nonlinear network in MFS, we find that feature learning may result in phenomena that escape the kernel scaling theory and require the richer, kernel adaptation approach. Our work thus provides a stepping stone in the development of a coherent view of contemporary feature learning theories.

References

[1] S Ariosto, R Pacelli, M Pastore, F Ginelli, M Gherardi, and P Rotondo. Statistical mechanics of deep learning beyond the infinite-width limit. *arXiv preprint arXiv:2209.04882*, 2022.

[2] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

[3] Hugo Cui, Florent Krzakala, and Lenka Zdeborova. Bayes-optimal learning of deep random networks of extensive-width. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6468–6521. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cui23b.html.

[4] Rishi Bommasani et. al. On the opportunities and risks of foundation models. (arXiv:2108.07258), July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

[5] 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, November 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc4de.

[6] 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. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

[7] Boris Hanin. Random neural networks in the infinite width limit as gaussian processes. *The Annals of Applied Probability*, 33(6A):4798–4819, 2023.

[8] Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-46444-8.

[9] 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. URL https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

[10] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, Sep 2021. doi: 10.1103/PhysRevX.11.031059. URL https://link.aps.org/doi/10.1103/PhysRevX.11.031059.

[11] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/content/115/33/E7665.

[12] Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

[13] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes. *arXiv e-prints*, art. arXiv:1810.05148, October 2018.

[14] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. A statistical mechanics framework for bayesian deep neural networks beyond the infinite-width limit. *Nature Machine Intelligence*, 5(12):1497–1507, December 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00767-6.

[15] Leonardo Petrini, Francesco Cagnetta, Eric Vanden-Eijnden, and Matthieu Wyart. Learning sparse features can lead to overfitting in neural networks. (arXiv:2206.12314), October 2022. URL http://arxiv.org/abs/2206.12314. arXiv:2206.12314 [cs, stat].

[16] Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3ROGsTX3IR.

[17] Inbar Seroussi and Ofer Zeitouni. Lower bounds on the generalization error of nonlinear learning models. *IEEE Transactions on Information Theory*, 2022.

[18] 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, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y.

[19] 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, 2023.

[20] Hugo Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478 (1-3):1–69, 2009.

[21] Christopher Williams. Computing with infinite networks. *Advances in neural information processing systems*, 9, 1996.

Appendix A. Derivation of Unified Formalism

A.1. Kernel Scaling Approach

For the special case of a linear activation function $\phi(h) = h$, (3) becomes the cumulant-generating function of a Gaussian $W(\tilde{f}|w) = 1/2\, \tilde{f}^{\mathrm{T}} C^{(xx)} \tilde{f}\, \|w\|^2$. This shows that the readout weights only appear in the form of the squared norm $Q := \|w\|^2$. The distribution of the output of the network (2) for standard scaling $\gamma = 1$ is hence

$$p(y, f|C^{(xx)}) = \mathcal{N}(y|f, \kappa) \int dQ\, \mathcal{N}(f|0, Q\, C^{(xx)})\, p(Q), \tag{11}$$
$$p(Q) = \left\langle \delta[-Q + \|w\|^2] \right\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{g_w}{M^\gamma})}.$$

Here $p(Q)$ is the distribution of the Euclidean length of the vector of $w \in \mathbb{R}^M$ with M i.i.d. Gaussian entries of variance g_w/M^γ each.

For large M and for standard scaling $\gamma = 1$ we may approximate its distribution by the Gärtner-Ellis theorem as

$$p(Q) \overset{\text{l.d.p.}}{\simeq} e^{-\Gamma(Q)}, \tag{12}$$
$$\Gamma(Q) = \sup_{\tilde{Q}} \{\tilde{Q}Q - W(\tilde{Q})\} = -\frac{M}{2} \left[\frac{Q}{g_w} - \ln \frac{Q}{g_w}\right] + \text{const.}, \tag{13}$$

because its cumulant-generating function has the scaling form $W(\tilde{Q}) = M \ln \left\langle \exp\left(\tilde{Q} w^2\right)\right\rangle_{w \sim \mathcal{N}(0, g_w/M)} = M\, \lambda_Q(\tilde{Q}/M)$ with $\lambda_Q(k) = -\frac{1}{2} \ln\left[1 - 2 g_w k\right]$ independent of M. Intuitively this implies that the mean $\langle Q \rangle \propto M^0$ and its variance is $\langle\langle Q^2 \rangle\rangle \propto M^{-1}$, so Q concentrates, where $\langle\langle ... \rangle\rangle$ refers to cumulant expectation. So the network prior for the labels $p(y|C^{(xx)}) \equiv \int df\, p(y, f|C^{(xx)})$ follows from (11) as

$$p(y|C^{(xx)}) \overset{\text{l.d.p}}{\simeq} \int dQ\, e^{S(Q|y)} \tag{14}$$
$$S(Q|y) = \ln \mathcal{N}(y|0, QC^{(xx)} + \kappa \mathbb{I}) - \Gamma(Q), \tag{15}$$

where one may call $S(Q|y)$ the "action" for Q. Its form agrees to Li & Sompolinsky [10], their Eq. A11 for a single layer. We obtain an extension for mean-field scaling $\gamma = 2$ by considering, instead of Q, the distribution of $\mathcal{Q} := MQ$. Analogous to (13) its distribution

$$p(\mathcal{Q}) = \left\langle \delta[-\mathcal{Q} + M\, \|w\|^2] \right\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{g_w}{M^2})} \overset{\text{l.d.p.}}{\simeq} \sup_{\tilde{Q}} \{\tilde{Q}\mathcal{Q} - W(\tilde{Q})\} \equiv \Gamma(\mathcal{Q})$$

so $\mathcal{Q}$ for mean-field scaling is distributed identical to Q in standard scaling. The network prior for mean-field scaling is hence written as

$$p(y|C^{(xx)}) \overset{\text{l.d.p}}{\simeq} \int d\mathcal{Q}\, e^{S(\mathcal{Q}|y)} \tag{16}$$
$$S(\mathcal{Q}|y) = \ln \mathcal{N}(y|0, \mathcal{Q}/MC^{(xx)} + \kappa \mathbb{I}) - \Gamma(\mathcal{Q}) \tag{17}$$

In the following, we will again rename $\mathcal{Q}$ by Q.

In both expressions, (17) and (15), the exponent S contains two terms, both of which are extensive in n or M, respectively. The integral over Q, in the proportional limit $n, M \to \infty$ while n/M fixed, will hence be dominated by the maximum with regard to Q. This maximum a posteriori estimate for Q, given by the conditional on the training data $p(Q|y) = p(y|Q)\, p(Q)/p(y)$ only depends on the numerator and is therefore given by the stationary point Q^* of (14) given by $\partial S/\partial Q \stackrel{!}{=} 0$. To obtain predictions beyond the length of the readout $Q = \|w\|^2$ one notes that the mean discrepancies between label and network output are $\langle \Delta_\alpha \rangle := y_\alpha - \langle f_\alpha \rangle_{p(f|y,C^{(xx)})}$ are given with $\ln\, p(y|C^{(xx)}) \overset{\text{MAP } Q}{\simeq} \sup_Q\, S(Q|y)$

$$\langle \Delta_\alpha \rangle = -\kappa \frac{\partial}{\partial y_\alpha} \ln p(y|C^{(xx)}) \overset{(11)}{=} \left\langle y_\alpha - f_\alpha \right\rangle \tag{18}$$
$$\overset{\text{MAP } Q}{\simeq} -\kappa \frac{\partial}{\partial y_\alpha} S(Q^*|y) = \kappa \left(Q^* M^{1-\gamma}\, C^{(xx)} + \kappa \mathbb{I} \right)^{-1} y,$$

where the inner derivative $\partial S/\partial Q\, \partial Q/\partial z_\alpha = 0$ due to stationarity of Q^*. A consequence of this feature learning theory is that the mean predictor is the same as the one of the NNGP with a different regularization noise $\kappa/(Q^* M^{1-\gamma})$

$$\begin{aligned} \langle f_{x_*} \rangle &= \left[\mathbb{I} - \kappa \left(Q^* M^{1-\gamma}\, C^{(xx)} + \kappa \mathbb{I} \right)^{-1} \right] y \\ &= \left[C^{(xx)}_{x_* \circ} \right] \left(C^{(xx)} + \kappa/(Q^* M^{1-\gamma}) \mathbb{I} \right)^{-1}_{\circ\circ} y_\circ, \end{aligned}$$

where $\circ$ refer to training indices and x_* to the test point. Along similar lines follows the variance as

$$\begin{aligned} \langle\!\langle \Delta_\alpha, \Delta_\beta \rangle\!\rangle =& \langle\!\langle f_\alpha, f_\beta \rangle\!\rangle \\ &= \kappa \delta_{\alpha\beta} - \kappa^2 \left(C + \kappa \mathbb{I} \right)^{-1}_{\alpha\beta} = C - C \left[C + \kappa \mathbb{I} \right]^{-1} C \Big|_{C = Q^* M^{1-\gamma} C^{(xx)}}, \end{aligned}$$

which is the same expression as in the NNGP, but with the NNGP kernel rescaled by $Q^* M^{1-\gamma}$.

This derivation only employed methods from large deviation theory and can be made rigorous. It enables by large M and approximated Q by its maximum a posteriori estimate.

A.1.1. CUMULANT EXPANSION FOR NON-LINEAR ACTIVATION

For a non-linear activation function we perform a cumulant expansion of eq. (3) up to Gaussian order, writing the weights in standard scaling as $w = \bar{w}/\sqrt{M}$ with $\bar{w} = \mathcal{O}(1)$

$$W(\tilde{f}|w) = \frac{1}{2} \sum_{\alpha\beta} \tilde{f}_\alpha\, C^{(\phi\phi)}_{\alpha\beta}\, \tilde{f}_\beta\, \frac{1}{M} \sum_i \bar{w}_i^2 + \mathcal{O}\big(1/M^2 \tilde{f}^4\big)\,, \tag{19}$$

where we assumed a point-symmetric activation function, so that only even cumulants appear, and $C^{(\phi\phi)}_{\alpha\beta} = \langle \phi_\alpha \phi_\beta \rangle_{h \sim \mathcal{N}(0, C^{(xx)})}$ is the second cumulant. Even though the suppressed terms are $\propto M^{-1}$, it is unclear when this approximation is accurate, because the next order suppressed term explicitly reads $\frac{1}{4! M^2} \sum_i \overline{w}_i^4 \sum_{\alpha,\beta,\gamma,\delta} \langle\!\langle \phi_\alpha \phi_\beta \phi_\delta \phi_\delta \rangle\!\rangle f_\alpha f_\beta f_\gamma f_\delta$ where $\langle\!\langle \phi_\alpha \phi_\beta \phi_\delta \phi_\delta \rangle\!\rangle$ is the fourth cumulant of the postactivations. It is a priori not clear that this term is subleading compared to the Gaussian term, because of the summation $\sum_{\gamma,\delta} \propto n^2 \propto M^2$. A justification must therefore come from additional assumptions. An example is a Bayes-optimal setting, where the students weights are Gaussian i.i.d. and the teacher uses the same distribution as a prior. Then a Gaussian equivalence principle holds [6] due to the Nishimori property which has been exploited in [3]

A.2. Kernel Adaptation Approach

For the kernel adaptation approach, we first take the marginalization over f in (2) to obtain

$$p(y|C^{xx}) = \int df \mathcal{N}(f|y, \frac{\kappa}{M^{\gamma-1}}\mathbb{I}) \int \mathcal{D}\tilde{f}\, \exp\big(-f^T\tilde{f} + W(\tilde{f}|C^{xx})\big) \tag{20}$$

$$= \int \mathcal{D}\tilde{f}\, \exp\big(-y^{\mathrm{T}}\tilde{f} + \frac{\kappa}{2M^{\gamma-1}}\tilde{f}^{\mathrm{T}}\tilde{f} + W(\tilde{f}|C^{xx})\big)\,, \tag{21}$$

where W follows from (3) by taking the expectation value over the w to obtain

$$W(\tilde{f}|C^{(xx)}) := \ln \Big\langle \exp\big(\sum_\alpha \tilde{f}_\alpha \sum_i w_i\, \phi(h_{\alpha i})\big)\Big\rangle_{w_i\sim\mathcal{N}(0,g_w/M^\gamma),h_{\alpha i}} \tag{22}$$

$$= M\, \ln \Big\langle \exp\big(\frac{1}{2}\sum_{\alpha\beta} \tilde{f}_\alpha\tilde{f}_\beta \frac{g_w}{M^\gamma}\, \phi(h_\alpha)\phi(h_\beta)\big)\Big\rangle_{h_\alpha}. \tag{23}$$

A.2.1. MEAN-FIELD SCALING

Now assume mean-field scaling $\gamma = 2$ and correspondingly scale the regularization noise κ/M so that it will not dominate the signal part. Then the cumulant-generating function in (23) has scaling form

$$\mathcal{W}(\tilde{f}) := \frac{\kappa}{2\,M}\tilde{f}^{\mathrm{T}}\tilde{f} + W(\tilde{f}) = M\,\lambda_f(\tilde{f}/M), \tag{24}$$

where $\lambda_f(k) = \kappa/2\, k^{\mathrm{T}}k + \ln \Big\langle \exp\big(g_w/2\, \sum_{\alpha\beta} k_\alpha k_\beta\, \phi(h_\alpha)\phi(h_\beta)\big)\Big\rangle_{h_\alpha\sim\mathcal{N}(0,C^{(xx)})}$ is independent of M with $k := \tilde{f}/M$ This implies that for large M, the probability distribution $p(y|C^{(xx)})$ can be approximated by the Gärtner-Ellis theorem [20] as

$$-\ln\, p(y|C^{(xx)})/M = \sup_k\{k^{\mathrm{T}}y - \lambda_f(k)\} =: \gamma_f(y|C^{(xx)}). \tag{25}$$

The supremum condition $0 = y - \nabla_k\lambda_f(k)$ yields with $k^* = \langle\Delta\rangle/\kappa$

$$k^* = \langle\Delta\rangle/\kappa = \big(\overline{C} + \kappa\mathbb{I}\big)^{-1}\, y, \tag{26}$$

where $\overline{C} := g_w\, \Big[\phi(h)\phi^{\mathrm{T}}(h)\Big]$ plays the role of the kernel and the expectation $\big[\,\ldots\,\big]$ is with regard to the measure

$$\big[\,\ldots\,\big] \propto \big\langle\, \ldots\, \exp\big(g_w/2 \sum_{\alpha\beta} k_\alpha k_\beta\, \phi(h_\alpha)\phi(h_\beta)\big)\big\rangle_{h\sim\mathcal{N}(0,C^{(xx)})}, \tag{27}$$

where the proportionality constant is given by the proper normalization. We note that when taking $k = 0$, then $\bar{C}$ is simply the GP kernel- $g_w C^{(\phi\phi)}$. The expression (26) has a similar form as the discrepancies in the case of kernel scaling (18), replacing $Q\,C^{(xx)} \to \overline{C}$. In particular, the limit $M \to \infty$ exists and k^* assumes a finite value, making the measure (27) non-Gaussian. Compared to kernel scaling, where one obtains a single-order parameter Q, one here has the vector of discrepancies $\langle\Delta\rangle = \kappa k^* \in \mathbb{R}^n$ as order parameters.

A.2.2. LINEAR ACTIVATION

For the linear case considered in the main text, the expectation over h in (23) is a Gaussian integral with the solution

$$W\big(\tilde{f}|C^{(xx)}\big) = -\frac{M}{2}\left[\ln\det\left(\left[[C^{(xx)}]^{-1} - \frac{g_w}{M^2}\tilde{f}\tilde{f}^{\mathrm{T}}\right]\right) + \ln\det(C^{(xx)})\right] \tag{28}$$

$$= -\frac{M}{2}\left[\ln\left[\left(1 - \frac{g_w}{M^2}\tilde{f}^{\mathrm{T}}C^{(xx)}\tilde{f}\right)/\det(C^{(xx)})\right] + \ln\det(C^{(xx)})\right] \tag{29}$$

$$= -\frac{M}{2}\ln\left(1 - \frac{g_w}{M^2}\tilde{f}^{\mathrm{T}}C^{(xx)}\tilde{f}\right), \tag{30}$$

where we used the matrix determinant lemma for the rank-one part $\tilde{f}\tilde{f}^{\mathrm{T}}$. Since each appearance of $\tilde{f}$ comes with one factor $1/M$, the cumulant-generating function has scaling form, so we define with $k := \tilde{f}/M$ the scaled cumulant-generating function $\lambda_f(k) = W(Mk)/M$

$$\lambda_f(k) = -\frac{1}{2}\left[\ln\left[\left(1 - g_w k^{\mathrm{T}}C^{(xx)}k\right)/\det(C^{(xx)})\right] + \ln\det(C^{(xx)})\right] \tag{31}$$

$$= -\frac{1}{2}\ln\left(1 - g_w k^{\mathrm{T}}C^{(xx)}k\right), \tag{32}$$

and the supremum condition corresponding to (26) for the first form of W in (31) takes the explicit form

$$0 = y_\alpha - \kappa k^*_\alpha - g_w \sum_\beta \left\{\left[[C^{(xx)}]^{-1} - g_w k^*(k^*)^{\mathrm{T}}\right]^{-1}\right\}_{\alpha\beta} k^*_\beta, \tag{33}$$

which is the second form of eq. (9) given in the main text. The second form of the cumulant-generating function in (31) yields the supremum condition

$$0 \stackrel{!}{=} y_\alpha - \kappa\, k^*_\alpha - \frac{g_w[C^{(xx)}k^*]_\alpha}{1 - g_w(k^*)^{\mathrm{T}}C^{(xx)}k^*}, \tag{34}$$

which yields the first form of eq. (9) in the main text.

A.2.3. STANDARD SCALING

We here investigate the kernel adaptation approach in standard scaling $\gamma = 1$. Then the cumulant-generating function (23) does not possess scaling form. This implies that fluctuations of f and hence fluctuations of the discrepancies become important. To treat these systematically, we use that by (4) $\ln\, p(y|C^{(xx)})$ acts as a cumulant-generating function for the discrepancies. So we define the cumulant-generating function from (4)

$$w(j) := \ln p(-j|C^{(xx)}) = \ln \int \mathcal{D}\tilde{f}\, \exp\left(j^{\mathrm{T}}\tilde{f} + \mathcal{W}(\tilde{f})\right), \tag{35}$$

fluctuation corrections by defining the effective action

$$\gamma(\tilde{f}_*) := \sup_j\{j^{\mathrm{T}}\tilde{f}_* - w(j)\}, \tag{36}$$

which obeys the equation of state

$$\frac{\partial \gamma(\tilde{f}_*)}{\partial \tilde{f}_*} = j + \underbrace{\tilde{f}_*^{\mathrm{T}} \frac{\partial j}{\partial \tilde{f}_*} - \frac{\partial w}{\partial j}^{\mathrm{T}} \frac{\partial j}{\partial \tilde{f}_*}}_{=0} = j \stackrel{!}{=} -y, \tag{37}$$

where we use that $\partial w / \partial j = \tilde{f}_*$ and that we need to insert for $-j$ the training labels y to obtain the posterior conditioned on the training data.

Now we would like to expand γ to include the leading order fluctuation corrections.

For a theory of the form $p \propto \exp(S(\phi))$ the fluctuation correction to γ would take the form $\frac{1}{2} \ln \det(-S'')$. (see, e.g., [8]), which comes from integrating over the Gaussian fluctuations around the self-consistently determined saddle point. This situation is different, because the integral $\int \mathcal{D}\tilde{f} = \prod_\alpha \int_{-i\infty}^{i\infty} \frac{d\tilde{f}}{2\pi i}$ proceeds along the imaginary axis for each sample coordinate α. Also, the cumulant-generating function $\mathcal{W}$ in (35) here plays the role of what normally is the action S. As a function of a real-valued $\tilde{f}$, the Hessian of $\mathcal{W}$ must be non-negative, because it is a covariance. Taking the integral along t_α where $\tilde{f}_\alpha = i t_\alpha$, the fluctuations are controlled by $-\mathcal{W}''$, which is non-positive, as it has to be. As result, the leading order fluctuation corretions to γ are

$$\gamma(\tilde{f}_*) = -\mathcal{W}(\tilde{f}_*) + \frac{1}{2} \ln \det \left(\mathcal{W}''(\tilde{f}_*) \right).$$

Where $\mathcal{W}''$ is the Hessian of $\mathcal{W}$ concerning $\tilde{f}_*$. If the $\ln \det$-term was absent, the equation of state $\partial \mathcal{W} / \partial \tilde{f}_* = y$ would be identical to the supremum condition in (25). Including the fluctuation determinant, the equation of state yields an implicit equation to determine $\tilde{f}_*$ as

$$-y = \frac{\partial \gamma(\tilde{f}_*)}{\partial \tilde{f}_*} = -\mathcal{W}'_\alpha(\tilde{f}_*) + \frac{1}{2} \sum_{\beta\gamma} \mathcal{W}'''_{\alpha\beta\gamma}(\tilde{f}_*) \left[\mathcal{W}''(\tilde{f}_*)\right]^{-1}_{\beta\gamma}. \tag{38}$$

Where here we used the notation $\mathcal{W}'_\alpha = \frac{\partial \mathcal{W}}{\partial \tilde{f}_\alpha}, \mathcal{W}''_{\alpha\beta} = \frac{\partial \mathcal{W}}{\partial \tilde{f}_\alpha \partial \tilde{f}_\beta}$ etc. So one needs the first to third derivatives of $\mathcal{W}$, which follow with the measure

$$\Big[\dots \Big]_h \propto \Big\langle \dots \exp\big(\frac{1}{2} \sum_{\alpha\beta} \tilde{f}_{*\alpha} \tilde{f}_{*\beta} \frac{g_w}{M} \phi(h_\alpha) \phi(h_\beta) \big) \Big\rangle_{h \sim \mathcal{N}(0, C^{(xx)})}, \tag{39}$$

where the proportionality is fixed by the proper normalization.

A.2.4. FLUCTUATION CORRECTIONS IN LINEAR NETWORK

For the linear network, in analogy to (29), the cumulant generating function W takes the form

$$W\big(\tilde{f} | C^{(xx)}\big) = -\frac{M}{2} \left[\ln \left[\left(1 - \frac{g_w}{M} \tilde{f}^{\mathrm{T}} C^{(xx)} \tilde{f} \right) / \det(C^{(xx)}) \right] + \ln \det(C^{(xx)}) \right] \tag{40}$$

$$= -\frac{M}{2} \ln \left(1 - \frac{g_w}{M} \tilde{f}^{\mathrm{T}} C^{(xx)} \tilde{f} \right). \tag{41}$$

The action $\mathcal{W}$ appearing in (35), correspondingly, takes the form

$$\mathcal{W}(\tilde{f}) = \kappa / 2 \tilde{f}^{\mathrm{T}} \tilde{f} - M/2 \ln \left(1 - g_w / M \tilde{f}^{\mathrm{T}} C^{(xx)} \tilde{f} \right). \tag{42}$$

Since, on the other hand, $\mathcal{W}$ evaluated at real-valued $\tilde{f}$, is a cumulant-generating function its Hessian is non-negative.

The explicit terms for the derivatives of $\mathcal{W}$ are given by:

$$\mathcal{W}\left(\tilde{f}\right) = \frac{\kappa}{2}\tilde{f}^T\tilde{f} - \frac{M}{2}\left[\ln\left(1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}\right)\right] \tag{43}$$

$$\mathcal{W}'_\alpha\left(\tilde{f}\right) = \left[\kappa\tilde{f} + \frac{g_w C^{(xx)}\tilde{f}}{1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}}\right]_\alpha$$

$$\mathcal{W}''_{\alpha\beta}\left(\tilde{f}\right) = \left[\kappa I + \frac{2g_w^2}{M}\frac{C^{(xx)}\tilde{f}\tilde{f}^T C^{(xx)}}{\left(1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}\right)^2} + g_w\frac{C^{(xx)}}{1 - \frac{g_w}{M}\tilde{f}^T C\tilde{f}}\right]_{\alpha\beta}$$

$$\mathcal{W}'''_{\alpha\beta\gamma}\left(\tilde{f}\right) = \frac{2g_w^2}{M\left(1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}\right)^2}\underbrace{\left[\frac{4g_w}{M\left(1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}\right)}C\tilde{f}\tilde{f}^T C^{(xx)} + C^{(xx)}\right]_{\alpha\beta}}_{T_{\alpha\beta}}\left[C^{(xx)}\tilde{f}\right]_\gamma$$

$$+ \frac{2g_w^2}{M}\frac{C^{(xx)}_{\alpha\gamma}\left[C^{(xx)}\tilde{f}\right]_\beta + C^{(xx)}_{\beta\gamma}\left[C^{(xx)}\tilde{f}\right]_\alpha}{\left(1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}\right)^2}$$

Using the Sherman-Morris formula, we obtain-

$$\left[\mathcal{W}''\left(\tilde{f}\right)\right]^{-1} = A - \frac{AC^{(xx)}\tilde{f}\tilde{f}^T C^{(xx)}A}{1 + \tilde{f}^T C^{(xx)}AC^{(xx)}\tilde{f}^T}; \quad A = \left[\kappa I + g_w\frac{C^{(xx)}}{1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}}\right]^{-1} \tag{44}$$

Note that-

Giving us the equation for y:

$$y = \left(\kappa + \frac{g_w C^{(xx)}}{1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}}\right)\tilde{f} \tag{45}$$

$$+ \frac{g_w^2 C^{(xx)}}{M\left(1 - \frac{g_w}{M}\tilde{f}^T C^{(xx)}\tilde{f}\right)}\left(tr\left(\left[\mathcal{W}''\right]^{-1}T\right)I + 2\left[I - \frac{AC^{(xx)}\tilde{f}\tilde{f}^T C^{(xx)}}{1 + \tilde{f}^T C^{(xx)}AC^{(xx)}\tilde{f}}\right]AC\right)\tilde{f}\,, \tag{46}$$

which needs to be solved for $\tilde{f}$ to obtain the discrepancies shown in Fig. **??**.

Appendix B. Sample Complexity Reduction

B.1. Theoretical Derivation

Assume a two-layer neural network with ReLU activation, trained on a target function-

$$y(x) = v^* \cdot x + \epsilon \left(|v^* \cdot x| - \sqrt{\frac{2}{\pi}} \right) \tag{47}$$

Where v^* is normalized for simplicity. As shown in eq. (26) The discrepancy obeys:

$$\langle \Delta \rangle = \kappa \left(\overline{C} + \kappa I \right)^{-1} y. \tag{48}$$

Where $\bar{C}$ is given by

$$\overline{C} = g_w \left\langle \phi(h) \phi(h)^{\mathrm{T}} \exp\left(\frac{1}{2} \frac{g_w}{M^2} \tilde{f}^{\mathrm{T}} \phi(h) \phi(h)^{\mathrm{T}} \tilde{f} \right) \right\rangle_{h \sim \mathcal{N}\left(0, C^{(xx)}\right)} \tag{49}$$

We note that h can be written in terms of the weights as $h = Xv$, so that in this space $\bar{C}$ is given by-

$$\begin{aligned} \overline{C}_{\alpha\beta} &= g_w \left\langle \phi(Xv) \phi(Xv)^{\mathrm{T}} \exp\left(\frac{1}{2} \frac{g_w}{M^2} \tilde{f}^{\mathrm{T}} \phi(Xv) \phi(Xv)^{\mathrm{T}} \tilde{f} \right) \right\rangle_{v \sim \mathcal{N}(0, g_v/d)} \\ &\propto g_w \int dv \phi(x_\alpha v) \phi(x_\beta v)^{\mathrm{T}} \exp \underbrace{\left(-\frac{d}{2g_v} v^{\mathrm{T}} v + \frac{1}{2} \frac{g_w}{M^2} \sum_{\alpha\beta} \tilde{f}_\alpha \tilde{f}_\beta \phi(x_\alpha v) \phi(x_\beta v) \right)}_{:= S_v} \end{aligned} \tag{50}$$

up to normalizing constants, where here x_α is the α-th row of the input data X. In general, it is not possible to obtain a closed form expression for S_v, and therefore we make the following simplifications:

- By (26), we replace $\tilde{f}/M \sim \langle \Delta \rangle / \kappa$, which is exact in mean-field scaling and consistent with Serrousi et. al. [19].
- We assume that for sufficiently small ϵ the discrepancy can be approximated as follows- $\langle \Delta \rangle \sim aXv^*$, where a is unknown.
- Taking the continuum limit/ equivalent kernel (EK) approximation, allowing the substitution of summations over the data points with integrals with respect to the data measure.

By comparison between numerics and theory in 3, we show that these approximations in fact capture the essential phenomena and lead to accurate predictions. Within these approximations, the distribution of the hidden layer weights is given by

$$\begin{aligned} S_v &= \frac{d}{2g_v} \left(v^{\mathrm{T}} v - \frac{g_v g_w n^2 a^2}{\kappa^2 d} \left[\int d\mu_{\boldsymbol{x}} \phi(vx) \left((v^*)^{\mathrm{T}} x \right) \right]^2 \right) \\ &= \frac{d}{2g_v} \left(v^{\mathrm{T}} v - \frac{g_v g_w n^2 a^2}{\kappa^2 d} \left(\frac{v^{\mathrm{T}} v^*}{2} \right)^2 \right) := v^{\mathrm{T}} \Sigma^{-1}(a) v \,. \end{aligned} \tag{51}$$

Similarly, note that in the continuum limit, the linear function is an eigenfunction of $\overline{C}$, as we obtain:

$$\begin{aligned}\overline{C}Xv^* &= \int dx\, \overline{C}\,(y,x)\, x^{\mathrm{T}}v = g_w \int dx \int dv \phi\,(vy)\,\phi\,(vx)\, x^{\mathrm{T}} v e^{-\frac{1}{2}v^{\mathrm{T}}\Sigma^{-1}(a)v} \\ &= \frac{g_w}{2}\int dv\; v^{\mathrm{T}}v^*\phi\,(vy)\, e^{-\frac{1}{2}v^{\mathrm{T}}\Sigma^{-1}(a)v} = \frac{g_w}{4}\underbrace{v^{*\mathrm{T}}\Sigma^{-1}\,(a)\,v^*}_{:=\Lambda_*(a)} y^{\mathrm{T}}v^*\,,\end{aligned} \tag{52}$$

where $\Lambda_* = \frac{g_v}{d}\left(1 - \frac{g_v g_w n^2 a^2}{4\kappa^2 d}\right)^{-1}$. Again assuming that ϵ is sufficiently small, the target is dominated by the linear term, resulting in the following equation for a:

$$a = \kappa\left[\frac{n g_w}{4}\Lambda_*\,(a) + \kappa\right]^{-1} \tag{53}$$

Which can be solved numerically for a. Finally, we obtain the expression for $\mathcal{R}_p$ for the test data

$$\mathcal{R}_p = \frac{\sum_{\alpha\beta}\left[\overline{C}^a_{\star\circ}\left(\overline{C}^a_{\circ\circ} + \kappa I\right)^{-1}\right]_{\alpha\beta} y\,(x_\beta)\, H_p\,(x_\alpha v^*)}{\sum_\alpha y\,(x_\alpha)\, H_p\,(x_\alpha v^*)}\,, \tag{54}$$

where in the weight space description, the kernel $\overline{C}$ is given by

$$\begin{aligned}\left[\overline{C}^a_{\star\circ}\right]_{\alpha\beta} &= \left\langle \phi\left(v^{\mathrm{T}}x^{\text{test}}_\alpha\right)\phi\left(v^{\mathrm{T}}x^{\text{train}}_\beta\right)\right\rangle_{v\sim\mathcal{N}(0,\Sigma(a))} \\ \left[\overline{C}^a_{\circ\circ}\right]_{\alpha\beta} &= \left\langle \phi\left(v^{\mathrm{T}}x^{\text{train}}_\alpha\right)\phi\left(v^{\mathrm{T}}x^{\text{train}}_\beta\right)\right\rangle_{v\sim\mathcal{N}(0,\Sigma(a))}\end{aligned} \tag{55}$$

and a is such that it solves the self-consistent equation (53) above.

B.2. Experimental Results

We choose a network configuration in which the theory predicts that the variance in the v^* direction will be $\sim \sqrt{d}$ times that of the variance in other directions. This choice implies that there will be significant feature learning and that the Gaussian equivalence principle, that the weights are identically distributed, will no longer hold. We would expect that in this scenario, the feature learning results in a significant effective dimensional reduction of the task, which in turn would result in a reduction of sample complexity. Thus, we would expect that the network begins to learn high-order polynomials at n points which is an order of magnitude lower than that of the GP limit and the scaling approach. In particular, we expect that the value of $\mathcal{R}_2$ will no longer be negligible.

B.2.1. Experimental Details

We trained five ensembles of 10 networks, each ensemble trained on a different dataset. We used the following network parameters:

$$M = \beta 750,\;\; d_0 = \beta 96,\;\; n_0 \in \{2\beta d_0, 3\beta d_0, 4\beta d_0\} \tag{56}$$

$$\kappa = g_w = 0.3,\;\; g_v = 0.5,\;\; \epsilon = 0.1 \tag{57}$$

Where we changed the values of β. With this parameter choice, we observe that there is in fact strong feature learning, as can be seen by the difference in the variance of the hidden layer weights aligned with v^* compared to the perpendicular one.

In fig. (3) the distribution of the weights in the hidden layer is shown, both in the same direction as the target (v^*) and in a direction perpendicular to it. In this figure, evidence of strong feature learning is observed, as the variance of the hidden layer weights in the v^* direction is $\sim 15 \sim \mathcal{O}(\sqrt{d})$ times that in the perpendicular dimension. The comparison between numerics and theory shows accurate agreement, a posteriori justifying the assumptions of the theory.

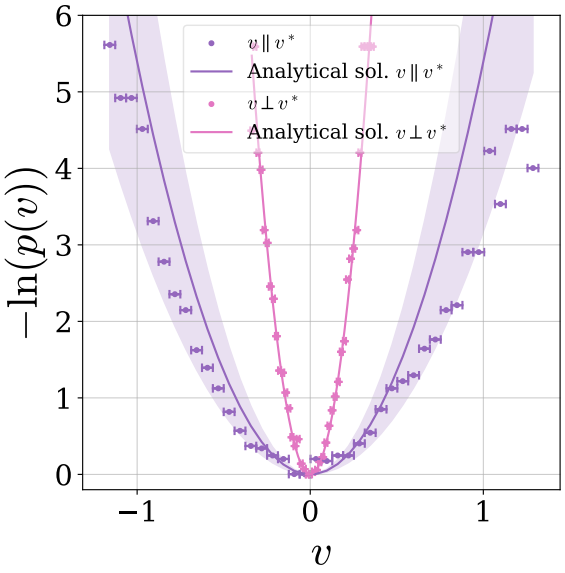

Figure 3: Negative log probability of hidden layer weight distribution. The distribution of the weights in the same direction of the target is significantly wider than the distribution in orthogonal directions.