From Kernels to Features: A Multi-Scale Adaptive Theory of Feature Learning

Noa Rubin [*1] **Kirsten Fischer** [*2 3] **Javed Lindner** [*2 4 5] **Inbar Seroussi** [6] **Zohar Ringel** [1] **Michael Krämer** [5]
Moritz Helias [2 4]

Abstract

Feature learning in neural networks is crucial for their expressive power and inductive biases, motivating various theoretical approaches. Some approaches describe network behavior after training through a change in kernel scale from initialization, resulting in a generalization power comparable to a Gaussian process. Conversely, in other approaches training results in the adaptation of the kernel to the data, involving directional changes to the kernel. The relationship and respective strengths of these two views have so far remained unresolved. This work presents a theoretical framework of multi-scale adaptive feature learning bridging these two views. Using methods from statistical mechanics, we derive analytical expressions for network output statistics which are valid across scaling regimes and in the continuum between them. A systematic expansion of the network's probability distribution reveals that mean-field scaling requires only a saddle-point approximation, while standard scaling necessitates additional correction terms. Remarkably, we find across regimes that kernel adaptation can be reduced to an effective kernel rescaling when predicting the mean network output in the special case of a linear network. However, for linear and non-linear networks, the multi-scale adaptive approach captures directional feature learning effects, providing richer insights than what could be recovered from a rescaling of the kernel alone.

[*]Equal contribution [1]The Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, Israel [2] Institute for Advanced Simulation (IAS-6), Computational and Systems Neuroscience, Jülich Research Centre, Jülich, Germany [3]RWTH Aachen University, Aachen, Germany [4]Department of Physics, RWTH Aachen University, Aachen, Germany [5]Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen University, Aachen, Germany [6]Department of Applied Mathematics, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel. Correspondence to: Noa Rubin <noa.rubin@mail.huji.ac.il>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

By mistake, this document lacks David Dahmen, Institute for Advanced Simulation (IAS-6), Computational and Systems Neuroscience, Jülich Research Centre, Jülich, Germany, as an additional author; he has been an author on this project from its beginning. To properly reference this work, please see our arXiv preprint (Rubin et al., 2025).

A central phenomenon that is essential for explaining the power of neural networks (NNs) is feature learning (FL), where networks learn meaningful high-dimensional representations of the data (Bengio et al., 2013). FL plays an increasingly important role in our ability to understand and rationalize the behavior of large language models (LLMs). Sparse autoencoders can extract so called monosemantic features from LLMs that are given by a superposition of layer activations (Bricken et al., 2023); these features allow interpreting and even altering model behavior (Templeton et al., 2024). Beyond interpretability, FL is essential for efficient generalization with finite data, as it enhances informative directions in the learned representations, reducing the sample complexity of learning functions of these directions (Abbe et al., 2021; Paccolat et al., 2021; Dandi et al., 2024). Despite its significance, many open questions remain regarding the mechanisms underlying the emergence of these feature directions.

A well-characterized case in NN theory is the limit of infinite-width and finite sample size, where networks behave as Gaussian processes (GPs) (MacKay, 2003), characterized by the neural network Gaussian process (NNGP) kernel (Neal, 1996; Williams, 1998; Matthews et al., 2018; Lee et al., 2018). However, the NNGP does not capture FL, which emerges at finite network width, in the proportional limit, where both network width and sample size tend to infinity proportionally (Li & Sompolinsky, 2021), or in certain scaling regimes (Yang et al., 2024). Hence the NNGP fails to capture the networks' nuanced internal representations that arise from feature learning (van Meegen & Sompolinsky, 2024). Multiple theoretical approaches have emerged to describe this phenomenon, yet there is no consensus on how to characterize FL. A common approach is to study the change of the network kernel, though the existing frameworks differ in their predictions for this change.

One prominent class of theories, which are commonly referred to as rescaling theories (Li & Sompolinsky, 2021; Pacelli et al., 2023; Baglioni et al., 2024), predicts that the average network output and variance can be described by a rescaled NNGP kernel. Initially developed for linear networks in the standard scaling regime[1], this framework surprisingly yields impressively accurate predictions even in mean-field scaling[2]. Despite the strong FL in this regime, the average network outputs can be obtained from an output kernel that is a rescaled NNGP kernel.

However, FL is often considered a structural phenomenon, as in the case of Gabor filters (Gabor, 1946; Rai & Rivas, 2020) that emerge in the latent layers of convolutional neural networks (Luan et al., 2018). Thus, the expectation is that the effect of FL on the output should be directional as well. The rescaling result raises fundamental questions about how learned features are represented in network outputs and can be captured theoretically.

In contrast, adaptive theories of FL (Roberts et al., 2022; Seroussi et al., 2023; Bordelon & Pehlevan, 2023; Fischer et al., 2024b; van Meegen & Sompolinsky, 2024) consider learned features, predicting that the kernel undergoes a structural change and incorporates features explicitly. Consequentially, these theories are able to predict phenomena in networks that stem from FL such as a reduction in sample complexity – the required amount of samples to learn a given task – relative to that of a GP (Naveh & Ringel, 2021b) as well as grokking (Rubin et al., 2024). However, adaptive theories are significantly more complex computationally than rescaling theories, while yielding comparable predictions for quantities such as the network loss. A fundamental open question remains: How can two such different descriptions of FL be valid at the same time?

In this work, we address this pivotal question by systematically connecting different FL theories as well as discussing their differences. Using methods from statistical physics, we recast the theoretical description of the posterior distribution of network outputs into a minimization problem with respect to a quantity which we call the “order parameter”. We find that different theories result from different choices of order parameters, in particular with regard to their dimensionality (see Fig. 1a). Our main contributions are:

- We derive a multi-scale adaptive theory that is valid across the full range of scaling regimes, from mean-field to standard scaling, which allows us to systematically include finite-width corrections (see Fig. 1b). This generalizes previous adaptive approaches, which were restricted to specific scaling regimes, and holds for arbitrary tasks as well as non-linear networks.

[1] A scaling where readout weight variance scales as $1/\text{width}$.
[2] A scaling where readout weight variance scales as $1/\text{width}^2$.

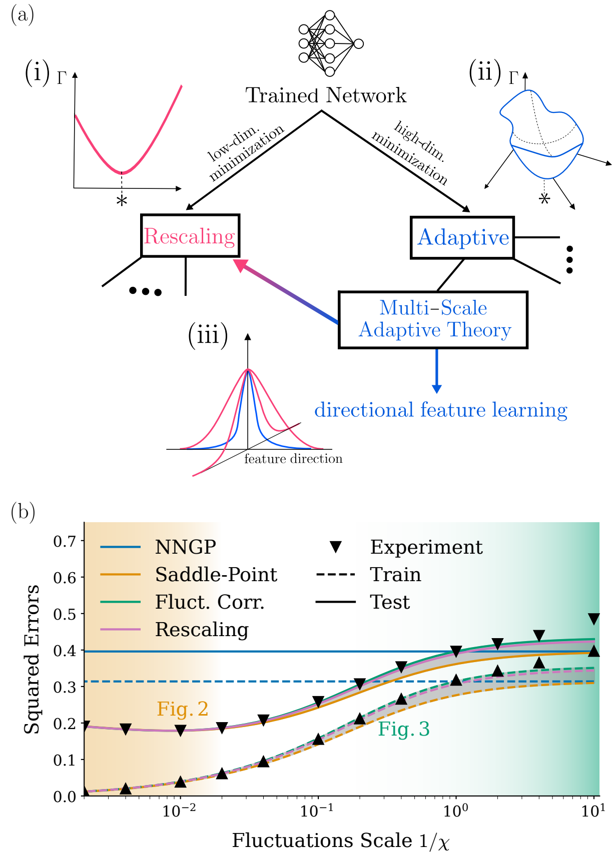

Figure 1. (a) The multi-scale adaptive theory bridges between rescaling and adaptive theories of feature learning. Starting from the distribution of network outputs for trained networks, the choice of order parameter decides whether a rescaling (red) or adaptive (blue) theory is obtained. The choice of order parameter recasts feature learning into either a (i) low-dimensional minimization or (ii) high-dimensional minimization problem. An approximation of the multi-scale adaptive theory in certain limits yields the result of the rescaling approach, but in addition describes (iii) directional aspects of feature learning. (b) Training (solid line) and test errors (dashed line) across scaling regimes for different approaches. While standard scaling (green shaded area) requires a one-loop approximation with fluctuation corrections (Fluct. Corr.), a saddle-point or tree-level approximation (Saddle-Point) is sufficient in mean-field scaling (orange shaded area). We show results for the kernel rescaling theory by (Li & Sompolinsky, 2021) as reference (Rescaling). We here show results for a linearly separable task; for results on MNIST see Fig. 6 in App. D. Parameters: $\gamma = 1$, $P_{\text{train}} = 80$, $N = 100$, $D = 200$, $\kappa_0 = 1$, $P_{\text{test}} = 10^3$, $g_v = g_w = 0.5$, $\Delta p = 0.1$.

- By analyzing the simplest non-trivial model, we reconcile adaptive and rescaling theories: we show that for the mean network output the multi-scale adaptive theory can be approximated in certain limits to yield an effective rescaling of the kernel. This explains why certain FL phenomena do not appear in rescaling theories.
- Our theory reveals how the two approaches differ. In linear networks, rescaling and adaptive theories yield equivalent predictions for the mean network output, but not for the output covariance. In mean-field scaling, the covariance exhibits clear adaptation to task-relevant directions, accurately captured by our adaptive theory and not by a rescaling theory. In non-linear networks, the disparity between the approaches emerges already at the level of the mean predictions: our adaptive theory correctly predicts a change in sample complexity class relative to the NNGP, a phenomenon that is not captured by rescaling theories.

Overall, our findings suggest that a comprehensive understanding of FL requires moving beyond kernel rescaling towards high-dimensional kernel adaptation.

2. Related works

The limit of infinite network width and finite amount of training data has been studied extensively, yielding among others the NNGP kernel (Neal, 1995; Williams, 1998; Lee et al., 2018; Matthews et al., 2018; Avidan et al., 2024). This theory relates network behavior at initialization to training dynamics (Poole et al., 2016; Pennington et al., 2017; Schoenholz et al., 2017; Xiao et al., 2018). However, the NNGP cannot explain the often superior performance of finite-width networks (Li et al., 2015; Chizat et al., 2019; Lee et al., 2020; Aitchison, 2020; Refinetti et al., 2021), requiring either the inclusion of finite-width effects or different infinite-width limits such as μP scaling (Yang et al., 2022; Vyas et al., 2023).

Describing FL in neural networks in a Bayesian framework has led to concurrent views: kernel rescaling (Li & Sompolinsky, 2021; 2022; Pacelli et al., 2023; Bassetti et al., 2024; Baglioni et al., 2024) and kernel adaptation (Aitchison, 2020; Naveh & Ringel, 2021a; Seroussi et al., 2023; Fischer et al., 2024b; Rubin et al., 2024; van Meegen & Sompolinsky, 2024). These differ in the choice of order parameters considered and in consequence also in the explained phenomena. A complementary perspective is provided by Yang et al. (2023), who propose a unified theoretical framework showing that FL extends classical kernel methods by enabling the learning of data-dependent feature maps in the infinite-width regime.

Various works study other aspects of networks in the Bayesian framework: Zavatone-Veth & Pehlevan (2021b) study properties of the network prior, whereas we focus on the network posterior. Hanin & Zlokapa (2023) obtain a rigorous non-asymptotic description of deep linear networks in terms of Meijer-G functions. Zavatone-Veth et al. (2022) study the same setting but consider explicit models on the input data in the limit of infinite pattern dimension. Zavatone-Veth & Pehlevan (2021a) investigate deep linear networks in different proportional limits, recovering the results from Li & Sompolinsky in an adaptive approach. Cui et al. (2023) study non-linear networks, exploiting the Nishimori conditions that hold for Bayes-optimal inference, where student and teacher have the same architecture and the student uses the teacher's weight distribution as a prior; the latter is assumed Gaussian i.i.d., which allows them to use the Gaussian equivalence principle (Goldt et al., 2020) to obtain closed-form solutions.

Our work is distinct from perturbative approaches such as (Antognini, 2019; Naveh et al., 2021; Cohen et al., 2021; Halverson et al., 2021; Roberts et al., 2022; Hanin & Zlokapa, 2023; Hanin, 2024) for the Bayesian setting or (Dyer & Gur-Ari, 2020; Huang & Yau, 2020; Aitken & Gur-Ari, 2020; Roberts et al., 2022; Bordelon & Pehlevan, 2023; Buzaglo et al., 2024) for gradient-based training that use the strength of non-Gaussian cumulants of the outputs as an expansion parameter. In contrast, we perform an expansion in terms of fluctuations around the mean outputs, which is able to capture phenomena that escape perturbative treatments such as phase transitions; this technique corresponds to an infinite resummation of perturbative terms. Our approach is similar to van Meegen & Sompolinsky (2024) with the difference that they scale weight variances as $1/N^3$, so that readout weights concentrate (see App. B for a comparison of the approaches).

Another line of work focuses on the dynamics of FL: Saxe et al. (2014) derive exact learning dynamics for deep linear networks, while Bordelon & Pehlevan (2023) use dynamical mean-field theory to describe network behavior in the early stages of training of gradient descent training in different scaling regimes; in contrast we consider networks at equilibrium. Yang & Hu (2020) consider the effect of network training dynamics and learning rate scales in networks. An experimental investigation of kernels in feature learning in gradient descent settings was performed by Canatar & Pehlevan (2022). Day et al. (2024) study the effect of weight initialization on generalization and training speed. A different viewpoint considers spectral properties of FL (Simon et al., 2023; Yang et al., 2024) as well as investigating the effects of learned representations directly (Petrini et al., 2023). Maillard et al. (2024) derive polynomial scaling limits of the required amount of training data.

3. Single hidden-layer network

We consider the following network architecture

$$h_\alpha = V x_\alpha,\ f_\alpha = w^\mathsf{T}\phi(h_\alpha),\ y_\alpha = f_\alpha + \xi_\alpha, \tag{1}$$

where ξ is Gaussian regularization noise $\xi \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\kappa)$. We consider P tuples of training data $\mathcal{D} = \{(x_\alpha, y_\alpha)\}_{1\le\alpha\le P}$ with $x_\alpha \in \mathbb{R}^D$ and $y_\alpha \in \mathbb{R}$ as well as an unseen test point (x_*, y_*) denoted by $*$. Here, ϕ denotes a non-linear activation function and $f_\alpha \in \mathbb{R}$ is the scalar network output. We study the Bayesian setting with Gaussian priors on the readin weights $V \in \mathbb{R}^{N\times D}$ as $V_{ij} \sim \mathcal{N}(0, g_v/D)$ and the readout weights $w \in \mathbb{R}^N$ as $w_i \sim \mathcal{N}(0, g_w/N^\gamma)$.

We differentiate between two cases: (a) standard scaling for $\gamma = 1$ and (b) mean-field scaling for $\gamma = 2$. Accordingly, we scale the regularization noise as $\kappa = \kappa_0 N^{1-\gamma}$ so that it does not dominate the network output in mean-field scaling. For concise notation, we use the shorthands $f_\mathcal{D} = (f_\alpha)_{1\le\alpha\le P}$, $X = (x_\alpha)_{1\le\alpha\le P}$ and $y = (y_\alpha)_{1\le\alpha\le P}$ in the following. Further, summations over repeated indices are implied $V_{kl}x_l \equiv \sum_{l=1}^N V_{kl}x_l$. The code for theory and experiments can be found in 10.5281/zenodo.15480898.

4. Multi-scale adaptive feature learning theory

We compute the network posterior on the test point (x_*, y_*) by conditioning on the training data $\mathcal{D}$ and derive a set of self-consistency equations for the average discrepancies $\langle\Delta\rangle$ between labels y and mean posterior network outputs $\langle f_\mathcal{D}\rangle$ on the training data. This description on the level of the discrepancies yields a high-resolution picture of the network behavior: it allows us to explain kernel rescaling results in the proportional limit as well as predict directional aspects of FL.

4.1. Predictor statistics of the neural network

We are interested in the Bayesian network posterior for the network output on training points $f_\mathcal{D}$ and a test point f_*, which corresponds to training the network with Langevin stochastic gradient descent (LSGD) (Welling & Teh, 2011; M et al., 2017; Naveh et al., 2021) until convergence (see App. C.3 for details). We denote the joint vector of outputs as $f \coloneqq (f_\mathcal{D}, f_*) \in \mathbb{R}^{P+1}$. Following along the lines of Segadlo et al. (2022a), we write the joint distribution as

$$\begin{aligned} p(f_\mathcal{D}, f_*, y) &= \mathcal{N}(y|f_\mathcal{D}, \kappa_0 N^{1-\gamma}) \\ &\times \int \mathrm{d}\tilde{f}_\mathcal{D} \int \mathrm{d}\tilde{f}_*\, \exp\left[-i\tilde{f}^\mathsf{T} f + W(i\tilde{f}_\mathcal{D}, i\tilde{f}_*)\right], \end{aligned} \tag{2}$$

with $\tilde{f} \coloneqq (\tilde{f}_\mathcal{D}, \tilde{f}_*)$ the conjugate fields to $f = (f_\mathcal{D}, f_*)$. The cumulant-generating function $W(i\tilde{f}_\mathcal{D}, i\tilde{f}_*)$ of the network prior is given by

$$W(i\tilde{f}_\mathcal{D}, i\tilde{f}_*) = \ln\left\langle \exp\left(\sum_{a=1}^{P+1} i\tilde{f}_a w_j \phi(h_{aj})\right)\right\rangle_{w_j, h_{aj}}, \tag{3}$$

where the average $\langle\dots\rangle_{w_j,h_j}$ is over the prior distribution on the network parameters and the hidden pre-activations $h_{aj} \overset{\text{i.i.d. over j}}{\sim} \mathcal{N}(0, C^{(xx)})$ with $C^{(xx)} = g_v/D\, XX^\mathsf{T} \in \mathbb{R}^{(P+1)\times(P+1)}$. The detailed derivation can be found in App. A. The statistics of the conjugate fields $\tilde{f}_\mathcal{D}$ are directly linked to the statistics of the network predictors $f_\mathcal{D}$ via the output discrepancies $\Delta = y - f_\mathcal{D}$ on the training data as

$$\langle\Delta\rangle = -i\kappa_0 N^{1-\gamma}\langle\tilde{f}_\mathcal{D}\rangle. \tag{4}$$

To obtain the statistics of the conjugate variables $(\tilde{f}_\mathcal{D}, \tilde{f}_*)$ and thus also of the network outputs $(f_\mathcal{D}, f_*)$, we define a conditional cumulant-generating function $\mathcal{W}(k, j_*|y) \coloneqq \ln\langle\exp(j_* f_* + ik^\mathsf{T}\tilde{f}_\mathcal{D})\rangle_{f_*, \tilde{f}_\mathcal{D}}$ which yields

$$\begin{aligned} \mathcal{W}(k, j_*|y) &= \ln\int \mathrm{d}\tilde{f}\, \exp\left[ik^\mathsf{T}\tilde{f}_\mathcal{D} + \mathcal{S}(\tilde{f}_\mathcal{D}, j_*|y)\right], \\ \mathcal{S}(\tilde{f}_\mathcal{D}, j_*|y) &= -iy^\mathsf{T}\tilde{f}_\mathcal{D} - \frac{\kappa_0}{2}N^{1-\gamma}\tilde{f}_\mathcal{D}^\mathsf{T}\tilde{f}_\mathcal{D} \\ &\quad + W(i\tilde{f}_\mathcal{D}, j_*). \end{aligned} \tag{5}$$

Here, we introduced source terms (k, j_*) with $k \in \mathbb{R}^P$, $j_* \in \mathbb{R}$ to obtain the statistics of $(\tilde{f}_\mathcal{D}, f_*)$ as derivatives. On the training points, we have

$$\langle\tilde{f}_\mathcal{D}\rangle = -i\nabla_k\mathcal{W}|_{k,j_*=0},\ \langle\langle\tilde{f}_\mathcal{D}\tilde{f}_\mathcal{D}^\mathsf{T}\rangle\rangle = -\nabla_k^2\mathcal{W}|_{k,j_*=0}, \tag{6}$$

with $\langle\langle\tilde{f}_\mathcal{D}\tilde{f}_\mathcal{D}^\mathsf{T}\rangle\rangle$ the covariance of $\tilde{f}_\mathcal{D}$. On the test point, we get

$$\langle f_*\rangle = \partial_{j_*}\mathcal{W}|_{k,j_*=0},\ \langle\langle f_*^2\rangle\rangle = \partial_{j_*}^2\mathcal{W}|_{k,j_*=0}. \tag{7}$$

However, the cumulant-generating function $\mathcal{W}(k, j_*|y)$ in (5) in general does not have an analytical solution. Instead, we perform a systematic expansion in terms of fluctuations of the network output using the Legendre transform of the cumulant-generating function $\mathcal{W}$ of the network posterior

$$\Gamma(\bar{\tilde{f}}, j_*|y) = \mathrm{extr}_k\, ik^\mathsf{T}\bar{\tilde{f}} - \mathcal{W}(k, j_*|y), \tag{8}$$

where we take the extremum with respect to k. This transform is a function of the mean conjugate field $\bar{\tilde{f}} = \langle\tilde{f}_\mathcal{D}\rangle$ (in the following, we drop the index $\mathcal{D}$ for readability), defined self-consistently by the stationary condition given by

$$\partial_{\bar{\tilde{f}}}\Gamma(\bar{\tilde{f}}, j_*|y) \overset{!}{=} 0. \tag{9}$$

The Legendre transform $\Gamma(\bar{\tilde{f}}, j_*|y)$ is thus a natural way of constructing a minimization problem that yields the quantity

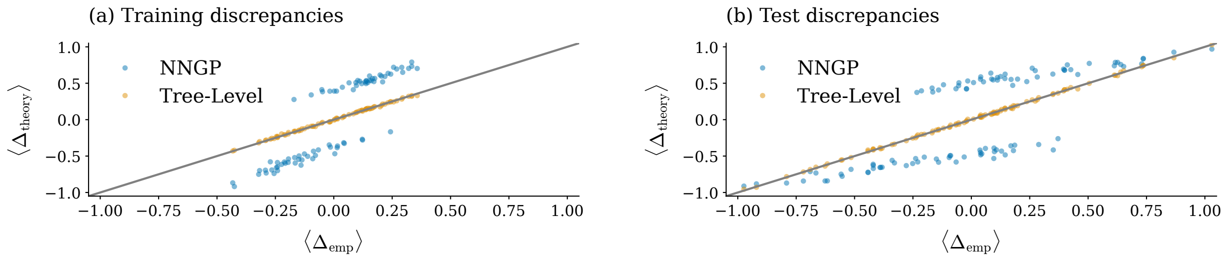

Figure 2. (a) Training discrepancies $\langle\Delta\rangle = y - \langle f_{\mathcal{D}}\rangle$ and (b) test discrepancies $\langle\Delta_*\rangle = y_* - \langle f_*\rangle$ on an Ising task in mean-field scaling. We show theoretical values for both NNGP and tree-level against empirical results, where the gray line marks the identity. In contrast to the NNGP, the tree-level approximation accurately matches the empirical values. While we use $\phi = \text{id}$ here, the non-linear case $\phi = \text{erf}$ yields similar results (see Fig. 9 in App. D). Parameters: $\gamma = 2$, $P_{\text{train}} = 80$, $N = 100$, $D = 200$, $\kappa_0 = 1$, $P_{\text{test}} = 10^3$, $g_v = g_w = 0.5$, $\Delta p = 0.1$.

we are interested in. It recasts the problem of computing the statistics of the posterior, which is the stationary solution of the stochastic minimization problem described by the LSGD training, into an effective deterministic optimization problem of Γ with regard to the mean discrepancies $\bar{\tilde{f}}$; intuitively, we may therefore think of Γ as an effective loss function that explicitly depends only on the mean discrepancies $\langle\Delta\rangle \propto \bar{\tilde{f}}$, but implicitly takes fluctuations of Δ into account. Moreover, it allows computing corrections to the mean network outputs in a systematic manner, building on a broad foundation of methods from statistical physics (Zinn-Justin, 1996; Helias & Dahmen, 2020).

Using the relationship between first-order parametric derivatives of the Legendre transform Γ and the cumulant-generating function $\mathcal{W}(k, j_*|y)$, we obtain

$$\langle f_*\rangle = -\partial_{j_*}\Gamma|_{k,j^*=0}. \tag{10}$$

Next, we consider systematic approximations of Γ for different scaling regimes, from which we determine the network output statistics.

4.2. Exact network prior for linear networks

The above expressions are exact but require knowledge of the cumulant-generating function W of the network prior in (3). For general non-linear activation functions ϕ, the cumulant-generating function W can only be approximated, e.g. using a cumulant expansion (see App. A.1.2 for details). For linear activations $\phi(h) = h$, however, we derive an exact expression

$$W(i\tilde{f}_{\mathcal{D}}, i\tilde{f}_*) = -\frac{N}{2}\ln\det\left[\mathbb{I} + \frac{g_w}{N^\gamma}\hat{C}^{(xx)}\tilde{f}\,\tilde{f}^{\mathsf{T}}\right]. \tag{11}$$

Since the goal of this work is to connect existing theories while also studying their differences, we consider the linear setting in the following as the simplest setting possible. Despite this choice, the theoretical framework presented here also applies to the non-linear case as discussed in App. A.

4.3. Saddle-point approximation in mean-field scaling

In mean-field scaling, the exponent $\mathcal{S}$ of the cumulant-generating function in (5) scales linearly with the network width N, while the fluctuations of the network output scale as $\langle\langle f f^{\mathsf{T}}\rangle\rangle \sim 1/N$ and become negligible. Thus, we can perform a saddle-point approximation for the integral in (5) and obtain the tree-level approximation of the Legendre transform (Helias & Dahmen, 2020) by replacing $\tilde{f}$ by its mean value $\tilde{f} \mapsto \bar{\tilde{f}}$ yielding

$$\Gamma(\bar{\tilde{f}}, j_*|y) \approx \Gamma_{\text{TL}}(\bar{\tilde{f}}, j_*|y) = -\mathcal{S}(\bar{\tilde{f}}, j_*|y). \tag{12}$$

We derive this result more rigorously in App. A.2 using a large deviation principle (Touchette, 2009). From the stationary condition in (9), we obtain a self-consistency equation for $\bar{\tilde{f}}$ given by

$$\bar{\tilde{f}} = -i\left(\kappa_0 N^{-1}\mathbb{I} + C_{\text{TL}}(\bar{\tilde{f}})\,C^{(xx)}_{\mathcal{D}\mathcal{D}}\right)^{-1} y, \tag{13}$$

$$C_{\text{TL}}(\bar{\tilde{f}}) = g_w N^{-1}\left[\mathbb{I} + \frac{g_w}{N^2}C^{(xx)}_{\mathcal{D}\mathcal{D}}\bar{\tilde{f}}\bar{\tilde{f}}^{\mathsf{T}}\right]^{-1}, \tag{14}$$

where $C^{(xx)}_{\mathcal{D}\mathcal{D}} \in \mathbb{R}^{P\times P}$ refers to the training data submatrix of $C^{(xx)}$. In the remainder of this section, $\bar{\tilde{f}}$ refers to the solution of (13). We obtain the discrepancies on the training points as

$$\langle\Delta\rangle_{\text{TL}} = \kappa_0\left(\kappa_0\mathbb{I} + C_{\text{TL}}(\bar{\tilde{f}})\,C^{(xx)}_{\mathcal{D}\mathcal{D}}\right)^{-1} y. \tag{15}$$

For the test point, we get

$$\begin{aligned}&\langle f_*\rangle_{\text{TL}} \qquad (16)\\ &= \left[C_{\text{TL}}(\bar{\tilde{f}})\,C^{(xx)}_{\mathcal{D}*}\right]^{\mathsf{T}}\left(\kappa_0 N^{1-\gamma}\mathbb{I} + C_{\text{TL}}(\bar{\tilde{f}})\,C^{(xx)}_{\mathcal{D}\mathcal{D}}\right)^{-1} y\end{aligned}$$

where $C^{(xx)}_{\mathcal{D}*} := \{g_v/D\, x_\alpha\cdot x_*\}_{1\le\alpha\le P} \in \mathbb{R}^{P\times 1}$, recovering results by Seroussi et al. (2023). In Fig. 2, we compare theoretical values for training and test discrepancies against empirical measurements for linear networks trained on a

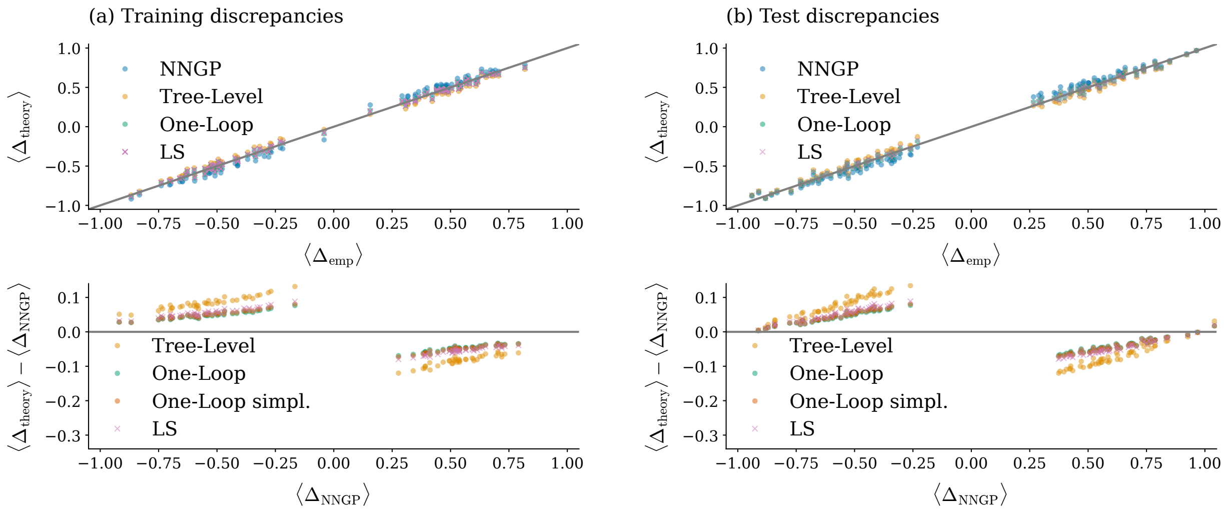

Figure 3. (a) Training discrepancies $\langle \Delta \rangle = y - \langle f_{\mathcal{D}} \rangle$ and (b) test discrepancies $\langle \Delta_* \rangle = y_* - \langle f_* \rangle$ on an Ising task in standard scaling. Upper row: theoretical values for different theories against empirical results; gray line marks the identity. Lower row: difference of theoretical values to the NNGP as a baseline against NNGP predictions, indicating small differences between the different approaches. Results of the kernel rescaling approach by Li & Sompolinsky (2021) are shown as reference (LS). Parameters: $\gamma = 1$, $P_{\text{train}} = 80$, $N = 100$, $D = 200$, $\kappa_0 = 0.4$, $P_{\text{test}} = 10^3$, $g_v = 0.5$, $g_w = 0.2$, $\Delta p = 0.1$.

linearly separable Ising task (see App. C.3 for details). Comparing to the NNGP as a baseline, we find that, while the NNGP fails to match network outputs, the multi-scale adaptive theory accurately predicts the values observed in trained networks. For similar results on non-linear networks, see Fig. 9 in App. D.

4.4. Fluctuation corrections in standard scaling

In standard scaling, output fluctuations are not scaled down by the network width N and instead become non-negligible. To obtain the leading-order fluctuation corrections, we expand the exponent $\mathcal{S}$ of the cumulant-generating function in (5) around its saddle-point $\bar{\tilde{f}}$ to second order as

$$\mathcal{S}(\tilde{f}_{\mathcal{D}}, j_*|y) \approx \mathcal{S}(\bar{\tilde{f}}, j_*) + \frac{1}{2}\big(\tilde{f}_{\mathcal{D}} - \bar{\tilde{f}}\big)^{\mathsf{T}} \mathcal{S}^{(2)} \big(\tilde{f}_{\mathcal{D}} - \bar{\tilde{f}}\big)\,, \quad (17)$$

where $\mathcal{S}^{(2)}$ denotes the Hessian of $\mathcal{S}(\tilde{f}, j_*|y)$ with respect to $\tilde{f}$ at the saddle-point $\bar{\tilde{f}}$. Calculating the Gaussian integral in (5), we obtain the one-loop approximation of the Legendre transform (Helias & Dahmen, 2020) as

$$\Gamma_{\text{1-Loop}}(\bar{\tilde{f}}, j_*|y) = -\mathcal{S}(\bar{\tilde{f}}, j_*) - \frac{1}{2}\log\det(-\mathcal{S}^{(2)})\,. \quad (18)$$

The self-consistency equation for $\bar{\tilde{f}}$ from the stationary condition in (9) is then given by

$$\bar{\tilde{f}}_\alpha = \Big[A\big(\bar{\tilde{f}}\big)\Big]^{-1}_{\alpha\beta} \left[-iy_\beta - \frac{1}{2}\Big[\mathcal{S}^{(2)}\Big]^{-1}_{\delta\epsilon} \mathcal{S}^{(3)}_{\epsilon\delta\beta}\Big|_{j^*=0}\right]\,, \quad (19)$$

where $A\big(\bar{\tilde{f}}\big) = \kappa_0 \mathbb{I} + C_{\text{TL}}\big(\bar{\tilde{f}}\big)\, C^{(xx)}_{\mathcal{D}\mathcal{D}}$ and $\mathcal{S}^{(n)}$ refers to the n-th derivative of the exponent $\mathcal{S}$ with respect to $\tilde{f}$ evaluated at $\bar{\tilde{f}}$ (see App. A.2 for details). In the remainder of this section, $\bar{\tilde{f}}$ refers to the self-consistent solution of (19), which is not necessarily the same as the one of (13) in the previous section. This yields for the training discrepancies $\langle \Delta \rangle_{\text{1-Loop}} = i\kappa_0 \bar{\tilde{f}}$ as in (4) and for the test point from (10)

$$\begin{aligned} &\langle f_* \rangle_{\text{1-Loop}} \quad (20)\\ &= \kappa_0^{-1} C^{(xx)}_{*\mathcal{D}} C_{\text{TL}}\big(\bar{\tilde{f}}\big)^{\mathsf{T}} \langle \Delta \rangle + \frac{1}{2} (\mathcal{S}^{(2)})^{-1}_{\beta\alpha} \mathcal{S}^{(3)}_{\alpha\beta*}\Big|_{j^*=0}\,. \end{aligned}$$

In the next section, we will see how these expressions reduce to a kernel rescaling theory in the proportional limit $N \propto P \to \infty$, which in linear networks we refer to as one-loop simplified in Fig. 3, where we compare theoretical predictions to empirical measurements on the Ising task. We show results for the multi-scale adaptive theory presented here as well as the rescaling theory by Li & Sompolinsky (2021), which was derived for the standard scaling regime. Due to the weak FL in standard scaling, all theories match the network behavior relatively well. However, by taking the NNGP as a reference, the differences between the theories become discernable: The tree-level solution shows deviations from the other solutions, predicting overly small test errors compared to the one-loop solution and compared to empirics. Furthermore, predictions of the one-loop solution agree to those of the rescaling theory by Li & Sompolinsky (2021).

The one-loop solution takes into account leading-order fluctuation corrections. The latter vanish in mean-field scaling, so one expects the one-loop approximation to converge to the tree-level result in this scaling regime. We show this explicitly in Fig. 1b, where we demonstrate how the different theories transition between the two scaling regimes by scaling $\kappa_0 \mapsto \kappa_0/\chi$ and $g_w \mapsto g_w/\chi$ with $0.1/N < 1/\chi < 10$ determining the scale of fluctuations. As expected, train and test errors decrease for increasing FL in the mean-field regime. Due to non-negligible fluctuations, the tree-level and one-loop solutions differ in standard scaling. When further increasing the fluctuations scale, even the one-loop solution does not accurately predict empirical measurements anymore since this regime requires fluctuation corrections beyond first order. In principle, the multi-scale adaptive approach allows computing these higher-order correction terms (Helias & Dahmen, 2020). When decreasing the fluctuations towards the mean-field scaling regime, the one-loop solution converges to the tree-level solution. Notably, the here presented multi-scale adaptive approach accurately predicts train and test errors across both scaling regimes, including the intermediate regime.

5. Kernel rescaling theory as an approximation of the multi-scale adaptive theory

Existing rescaling theories (Li & Sompolinsky, 2021; 2022; Pacelli et al., 2023; Bassetti et al., 2024; Baglioni et al., 2024) and adaptive theories (Naveh & Ringel, 2021a; Seroussi et al., 2023; Fischer et al., 2024b; Rubin et al., 2024; van Meegen & Sompolinsky, 2024) make both qualitatively and quantitatively different predictions regarding network behavior. On the one hand, rescaling approaches predict that the mean network output is equivalent to that obtained by a rescaled NNGP kernel. On the other hand, adaptive approaches such as the multi-scale adaptive theory presented here, as well as other existing approaches, predict that the kernel adapts to the data in a richer manner, showing changes in specific directions that are determined by the training data's statistics. While these approaches are quite different, we here expose the close relation between them in two respects: (i) We show that the adaptive and the rescaling approach can both be derived from the same starting point; the expression for the joint distribution of the network outputs (2). (ii) We show that for linear networks the adaptive approach in the proportional limit $N \propto P \to \infty$ can be approximated by a kernel rescaling for the mean outputs. For output fluctuations and for certain non-linear networks, however, such a reduction does not hold.

Technically, the differences between the two viewpoints stem from different choices of the order parameter used in the approximation of the posterior, utilizing either a saddle-point approximation or including fluctuation corrections. Specifically, with point (i), we show in App. A.4 that the equations obtained by Li & Sompolinsky (2021) can be obtained from (2) by marginalizing over the hidden preactivations h in (3) and performing a change of variables so that the posterior is a function of a single scalar order parameter $Q := \|w\|^2$. A saddle-point approximation with respect to this variable yields a self-consistency equation for Q and consequently expressions for the predictor statistics on test and training points, such as the mean and fluctuations. As the order parameter is scalar here, it is limited to describing scalar changes to the kernel.

Conversely, the choice of the high-dimensional order parameter in the multi-scale adaptive approach, which in mean-field scaling reproduces equations from the approach in (Seroussi et al., 2023), results in structural changes to the kernel. Notably, the choice of a high-dimensional order parameter results in the need to correct for fluctuations that arise in standard scaling, requiring us to go beyond the saddle-point approximation by using fluctuation corrections.

Surprisingly, as we have shown in the previous section, for a linear network on the level of the mean predictors, the multi-scale adaptive approach converges to that of the rescaling one, even though they have qualitatively different kernels. This observation motivates (ii), showing that for a linear network in the proportional limit $N \propto P \to \infty$, regardless of the initial choice of order parameter, the mean network output can be obtained from kernel regression (Rasmussen & Williams, 2006) with a rescaled NNGP kernel. For non-linear networks, however, differences already arise on the level of the mean predictors, as shown in the next section.

In the kernel rescaling case, the predictor for the mean output is obtained by replacing the NNGP kernel $K_{\text{NNGP}} = g_w N^{1-\gamma} C^{(xx)}_{\mathcal{D}\mathcal{D}}$ with a rescaled kernel

$$K_{\text{rescaling}} = Q/(g_w N^{1-\gamma})\, K_{\text{NNGP}} \,. \tag{21}$$

For the multi-scale adaptive approach presented here, the output statistics in mean-field scaling are obtained from

$$K_{\text{adaptive, TL}} = \left[\mathbb{I} + \frac{g_w}{N^\gamma} C^{(xx)}_{\mathcal{D}\mathcal{D}}\, \bar{\bar{f}}_{\text{TL}} \bar{\bar{f}}_{\text{TL}}^{\mathsf{T}}\right]^{-1} K_{\text{NNGP}} \,. \tag{22}$$

The appearing matrix product allows a non-trivial change of the NNGP kernel in certain meaningful directions, e.g. a teacher direction or dominant eigendirections of the input kernel that align with the target, thereby yielding additional insights. However, we derive an equivalent equation for the mean predictor by simplifying (11) using the matrix-determinant-lemma, which yields the mean output from a rescaled NNGP kernel given by

$$K_{\text{rescaling, TL}} = Q_{\text{TL}}\big(\bar{\bar{f}}\big)/(g_w N^{1-\gamma})\, K_{\text{NNGP}} \,, \tag{23}$$

where $Q_{\text{TL}}\big(\bar{\bar{f}}\big) = g_w N^{1-\gamma}/\big(1 + \frac{g_w}{N^\gamma} \bar{\bar{f}}^{\mathsf{T}} C^{(xx)} \bar{\bar{f}}\big)$ and $\bar{\bar{f}}$ satisfies (13). So even though the adaptive approach in mean-field scaling considers a directional change to the kernel, in terms of the mean output this is equivalent to a rescaled kernel.

In standard scaling, one cannot immediately express the mean output in terms of a rescaled kernel. However, in the proportional limit $N \propto P \to \infty$, certain fluctuation correction terms become negligible, reducing the expressions to a rescaling form again (see App. A.5). The rescaling factor is given by

$$\begin{aligned} &Q_{\text{1-loop}}(\bar{\bar{f}}) \\ &= Q_{\text{TL}}(\bar{\bar{f}}) - \frac{Q^2_{\text{TL}}(\bar{\bar{f}})}{N}\text{Tr}\left[A^{-1}(\bar{\bar{f}})\, C^{(xx)}_{\mathcal{DD}}\right] , \end{aligned} \tag{24}$$

where $A(\bar{\bar{f}}) := \kappa_0\mathbb{I} + Q_{\text{TL}}(\bar{\bar{f}})C^{(xx)}_{\mathcal{DD}}$, and $\bar{\bar{f}}$ satisfies (141).

We thus find that known theoretical approaches are all derived from the same original posterior distribution by considering different order parameters, while their resulting predictions for the mean network output behave like a rescaled NNGP. However the rescaling equivalence of mean predictors holds only for linear networks, as well as non-linear networks approximated using a cumulant expansion of the non-linearity as described in App. A.1, since one only needs to substitute $C^{(xx)}_{\mathcal{DD}} \mapsto C^{(\phi\phi)}_{\mathcal{DD}}$. However this equivalence does not necessarily hold for more fine-grained methods of approximating the non-linearity, such as the approach described in App. A.2.3. Applying such approximations within the adaptive approach allows predicting various phenomena that emerge in non-linear networks such as phase transitions (Rubin et al., 2024) and changes to sample complexity App. A.2.3, that escape a description by a rescaled kernel.

6. Directional feature learning emerges in adaptive description

The power of NNs stems from their ability to detect high-dimensional features in the data, implying that in the transition from the lazy to the rich regime this would be reflected in the network output statistics in a non-trivial manner. It is well established that the network weights adapt during training in an anisotropic manner, detecting relevant directions present in the training data (Seroussi et al., 2023; Fischer et al., 2024b); yet surprisingly, for the mean output of a linear network this adaptation seems to be equivalent to an isotropic rescaling of the NNGP kernel.

Teacher-student setting In this section, we demonstrate that the directional aspect of FL is nonetheless present in output fluctuations, which is only captured by the adaptive approach. Given a normalized feature direction $\hat{\phi}$, we define a directional FL measure $\Phi(\hat{\phi})$ that indicates to which degree this feature is represented by the network as

$$\Phi(\hat{\phi}) := \hat{\phi}^\mathsf{T}\langle\langle ff^\mathsf{T}\rangle\rangle\hat{\phi}. \tag{25}$$

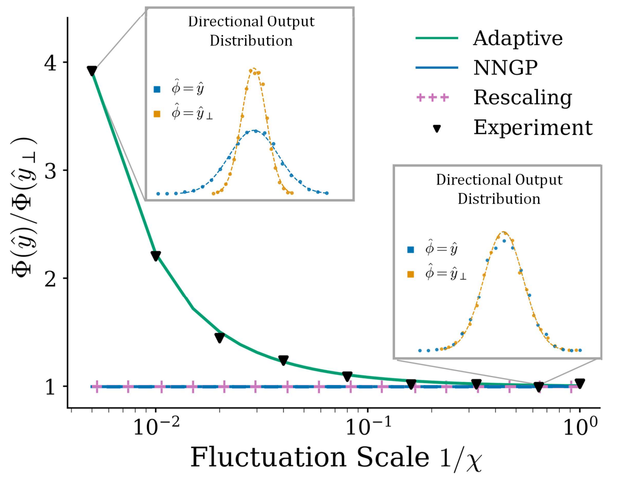

Figure 4. Relative directional feature learning in a teacher student setting as a function of the fluctuation scale $1/\chi$. Both NNGP and rescaling theory fail to capture directional feature learning, while the multi-scale adaptive theory accurately predicts network behavior. Insets show the output distribution in different directions; a detailed version can be found in Fig. 12 in the Appendix. Parameters: $P_{\text{train}} = 80$, $N = 200$, $D = 50$, $\kappa_0 = 2$, $g_v = 0.01$, $g_w = 2$.

Then, $\Phi(\hat{\phi}) \approx \text{Tr}\left(\langle\langle ff^\mathsf{T}\rangle\rangle\right)$ indicates that the feature direction $\hat{\phi}$ dominates the covariance, implying that this feature has been perfectly learned, whereas $\Phi(\hat{\phi}) \ll \text{Tr}\left(\langle\langle ff^\mathsf{T}\rangle\rangle\right)$ is an indication of weak directional FL. As derived in App. A.2, we obtain for the covariance of the network outputs on the training data

$$\langle\langle ff^\mathsf{T}\rangle\rangle_{\text{adaptive}} = \kappa\mathbb{I} - \kappa^2\left(A + \frac{2Q^2_{\text{TL}}}{N\kappa_0^2}F\right)^{-1}, \tag{26}$$

where we observe a structural change in the covariance matrix in form of the term $F := C^{(xx)}_{\mathcal{DD}}\langle\Delta\rangle\langle\Delta\rangle^\mathsf{T}C^{(xx)}_{\mathcal{DD}}$, which is not present in a rescaling of the NNGP, whose covariance is (see (129) in App. A.4 for details)

$$\langle\langle ff^\mathsf{T}\rangle\rangle_{\text{rescaling}} = \kappa\mathbb{I} - \kappa^2\left(\kappa\mathbb{I} + Q\, C^{(xx)}_{\mathcal{DD}}\right)^{-1}. \tag{27}$$

As evident from the expressions for the covariance, the directional FL measure Φ differs significantly between the two approaches, which is illustrated most easily for a kernel $C^{(xx)} \propto \mathbb{I}$: the isotropy of the rescaling theory then results in the same value of $\Phi(\hat{\phi})$ independent of the direction of $\hat{\phi}$, whereas the structural change of the covariance in the adaptive theory by the rank-one term $\langle\Delta\rangle\langle\Delta\rangle^\mathsf{T}$ in (26) may yield larger values of $\Phi(\hat{\phi})$ for features $\hat{\phi} \parallel \langle\Delta\rangle$.

We show the directional aspect of FL in a teacher-student setting, where the teacher is given by $y = Xw_*$ with $X \sim \mathcal{N}(0, \mathbb{I})$ and the student is a linear network as in Section 3 with $\phi(h) = h$. In this setting, the teacher defines a feature direction $\hat{y}_* = Xw_*/|Xw_*|$, and for comparison, we consider another possible feature direction $\hat{y}_\perp = Xw_\perp/|Xw_\perp|$, orthogonal to the former in the sense that $w_\perp \perp w_*$. The latter can be thought of as the direction of a randomly selected teacher that differs in the weights of the hidden layer from that of the actual target teacher. In Fig. 4, we show the relative directional FL measure $\Phi(\hat{y}_*)/\Phi(\hat{y}_\perp)$ between the target teacher and a random, orthogonal teacher direction. While the rescaling theory does not differentiate between these directions, the adaptive theory accurately predicts amplification of the teacher direction when entering the mean-field regime.

Amplification of kernel structure The expression (26) likewise exposes a cooperative effect caused by feature learning as a result of shaping the kernel jointly by the input and the output statistics. To illustrate this effect, we consider the Ising task (see (C.3)), where the input kernel $C^{(xx)}_{\mathcal{DD}} = g_v\,(\mathbb{I} + \epsilon\, yy^\mathsf{T})$ contains a rank-one term composed of the binary classification labels, unlike the teacher-student input kernel where $C^{(xx)}_{\mathcal{DD}} \propto \mathbb{I}$. Although the factor $\epsilon = 1 + 4p(1-p)$ is potentially small, as is the case for weakly distinguishable patterns ($0 \leq p - 1/2 \ll 1$), the matrix factor F appearing in (26) can significantly amplify this signal. Consider the case that the target is not learned well, for example at large ridge κ, so one has $\langle\Delta\rangle \simeq y$. The matrix factor

$$F \simeq g_V^2\; yy^\mathsf{T}\,(1 + \epsilon\, P)^2 \tag{28}$$

then shows that the weak structure $\propto \epsilon$ present in the input kernel has been amplified by a factor P (see Fig. 11 in App. D). For an unstructured kernel $C^{(xx)} \propto \mathbb{I}$ ($\epsilon = 0$) the rank-one component still persists $F = g_V^2\, yy^\mathsf{T}$, but it is not amplified by P here. The cooperative effect, in contrast, shows that feature learning may utilize coherent statistics between input and output.

Non-linear networks Directional effects in linear networks emerge only in the covariance of the outputs. For non-linear networks, however, these effects appear in the mean outputs, and have qualitative effects on network performance. The adaptive approach described in App. A.2.3 enables structural changes in the kernel which enhance specific features of the data. This enhancement, in turn, allows the network to learn non-linear functions of those features with significantly fewer data points than would typically be expected. In particular, introducing even a small non-linearity can lead to a fundamental difference in the behavior of approaches that rely on a rescaled or adapted kernel. InApp. A.2.3 we demonstrate this for a teacher-student setting similar to the one above but where we take both teacher and student to be non-linear. We find that the adaptive approach accurately predicts that the student will learn the non-linear component of the teacher, in contrast to the NNGP and kernel rescaling approaches which predict the non-linearity would not be learned (see Fig. 5 in App. A.2.3). These results reflect a change to the sample complexity class, and establish a principled connection between feature learning and improved network performance.

7. Discussion

In this work we present a unified theoretical framework to understand feature learning (FL) in the Bayesian setting across scaling regimes, from lazy to rich learning. This framework describes both effects of data adaptation in trained networks, i.e. directional changes of the network's output statistics in response to statistical dependencies present in the training data, as well as output rescaling phenomena that were described in previous works (Li & Sompolinsky, 2021; Pacelli et al., 2023). Our theory thus creates links between existing and so far unconnected previous theories. In the rich regime, the presented multi-scale adaptive theory clearly exposes directional aspects of FL, thus going beyond rescaling theories. By considering the simplest non-trivial case of a linear network, we finally reconcile the apparent contradiction between directional adaptation and rescaling by recovering the latter as an approximation of the former on the level of the mean output.

Furthermore, the multi-scale adaptive theory presented here applies to both standard and mean-field scaling and the entirety of the scaling spectrum. The latter is possible since the presented theoretical framework allows systematically computing fluctuation corrections depending on the scaling regime. While the tree-level solution and its equivalence to rescaling applies only to $\gamma = 2$, the one-loop theory applies to the full regime $\gamma \in [1, 2]$ and thus also the equivalence to rescaling (see Section 5). Note that this equivalence is restricted to the mean predictor, while the adaptive theory captures additional aspects, like directional FL in the variance terms (see Section 6). Moreover, this equivalence breaks down in the case of non-linear networks, where the presence of FL results in changes to the sample complexity class. In addition, our theory does not make any assumptions on the data set; we show results for an Ising task, a teacher student task and MNIST.

Outlook In this work, we study shallow networks, but we expect directional FL to be crucial for network performance in deeper networks as well. Beyond this, it will be valuable to extend the theoretical framework to other network architectures such as convolutional networks, residual networks, and transformers, using the respective network priors (Garriga-Alonso et al., 2019; Hron et al., 2020; Fischer et al., 2024a). To study the effect of noise in input data on FL (Lindner et al., 2023), we would like to include fluctuations of the input kernel in the theoretical framework.

Acknowledgements

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 368482240/GRK2416, and the Israeli Science Foundation - 374/23. The authors gratefully acknowledge the computing time granted by the JARA Vergabegremium and provided on the JARA Partition part of the supercomputer JURECA at Forschungszentrum Jülich (computation grant JINB33). MK would like to thank the Institute for Advanced Simulation (IAS-6) at Juelich Research Center and its directors Markus Diesmann and Sonja Grün for their hospitality during regular visits. We thank the reviewers for providing valuable feedback.

Impact Statement

This paper works towards understanding feature learning, thus aiming to advance explainability of networks. While the latter surely has societal impacts, these will be much further down the line.

References

10.5281/zenodo.15480898. URL https://doi.org/10.5281/zenodo.15480898. 3

Abbe, E., Boix-Adsera, E., Brennan, M. S., Bresler, G., and Nagaraj, D. The staircase property: How hierarchical structure can guide deep learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26989–27002. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e2db7186375992e729165726762cb4c1-Paper.pdf. 1

Aitchison, L. Why bigger is not always better: on finite and infinite neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 156–164. PMLR, 2020. URL https://proceedings.mlr.press/v119/aitchison20a.html. 2

Aitken, K. and Gur-Ari, G. On the asymptotics of wide networks with polynomial activations, 2020. URL https://arxiv.org/abs/2006.06687. 2

Antognini, J. M. Finite size corrections for neural network gaussian processes. pp. 1908.10030 [cs.LG], 2019. URL https://arxiv.org/abs/1908.10030. 2

Ariosto, S., Pacelli, R., Pastore, M., Ginelli, F., Gherardi, M., and Rotondo, P. Statistical mechanics of deep learning beyond the infinite-width limit. 2023. A.4, A.4.2

Avidan, Y., Li, Q., and Sompolinsky, H. Connecting ntk and nngp: A unified theoretical framework for wide neural network learning dynamics, 2024. URL https://arxiv.org/abs/2309.04522. 2

Baglioni, P., Pacelli, R., Aiudi, R., Di Renzo, F., Vezzani, A., Burioni, R., and Rotondo, P. Predictive power of a bayesian effective action for fully connected one hidden layer neural networks in the proportional limit. 133:027301, 2024. doi: 10.1103/PhysRevLett.133.027301. URL https://link.aps.org/doi/10.1103/PhysRevLett.133.027301. 1, 2, 5

Bassetti, F., Gherardi, M., Ingrosso, A., Pastore, M., and Rotondo, P. Feature learning in finite-width bayesian deep linear networks with multiple outputs and convolutional layers, 2024. URL https://arxiv.org/abs/2406.03260. 2, 5

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. 35 (8):1798–1828, aug 2013. doi: 10.1109/tpami.2013.50. URL https://doi.org/10.1109/tpami.2013.50. 1

Bordelon, B. and Pehlevan, C. Self-consistent dynamical field theory of kernel evolution in wide neural networks*. 2023(11):114009, nov 2023. doi: 10.1088/1742-5468/ad01b0. URL https://dx.doi.org/10.1088/1742-5468/ad01b0. 1, 2

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., and team, A. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html. 1

Buzaglo, G., Harel, I., Nacson, M. S., Brutzkus, A., Srebro, N., and Soudry, D. How uniform random weights induce non-uniform bias: Typical interpolating neural networks generalize with narrow teachers. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5035–5081. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/buzaglo24a.html. 2

Canatar, A. and Pehlevan, C. A kernel analysis of feature learning in deep neural networks. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8, 2022. doi: 10.1109/Allerton49937.2022.9929375. URL https://doi.org/10.1109/Allerton49937.2022.9929375. 2

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. volume 32, 2019. URL https://openreview.net/pdf?id=rkgxDVSlLB. 2

Cohen, O., Malka, O., and Ringel, Z. Learning curves for overparametrized deep neural networks: A field theory perspective. 3:023034, 2021. doi: 10.1103/PhysRevResearch.3.023034. URL https://link.aps.org/doi/10.1103/PhysRevResearch.3.023034. 2

Cui, H., Krzakala, F., and Zdeborova, L. Bayes-optimal learning of deep random networks of extensive-width. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 6468–6521. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cui23b.html. 2

Dandi, Y., Krzakala, F., Loureiro, B., Pesce, L., and Stephan, L. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024. URL http://jmlr.org/papers/v25/23-1543.html. 1

Day, H., Kahn, Y., and Roberts, D. A. Feature learning and generalization in deep networks with orthogonal weights, 2024. URL https://arxiv.org/abs/2310.07765. 2

Dyer, E. and Gur-Ari, G. Asymptotics of wide networks from feynman diagrams. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gFvANKDS. 2

Fischer, K., Dahmen, D., and Helias, M. Field theory for optimal signal propagation in resnets, 2024a. URL https://arxiv.org/abs/2305.07715. 7

Fischer, K., Lindner, J., Dahmen, D., Ringel, Z., Krämer, M., and Helias, M. Critical feature learning in deep neural networks, 2024b. 1, 2, 5, 6, A

Gabor, D. Theory of communication. *Journal of the Institution of Electrical Engineers - Part I: General*, 94:58–58, 1946. URL https://api.semanticscholar.org/CorpusID:61327032. 1

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklfsi0cKm. 7

Goldt, S., Reeves, G., Mézard, M., Krzakala, F., and Zdeborová, L. The Gaussian equivalence of generative models for learning with two-layer neural networks. 2020. URL https://arxiv.org/abs/2006.14709v1. 2

Halverson, J., Maiti, A., and Stoner, K. Neural networks and quantum field theory. *Machine Learning: Science and Technology*, 2(3):035002, apr 2021. doi: 10.1088/2632-2153/abeca3. URL https://doi.org/10.1088/2632-2153/abeca3. 2

Hanin, B. Random fully connected neural networks as perturbatively solvable hierarchies. *Journal of Machine Learning Research*, 25(267):1–58, 2024. URL http://jmlr.org/papers/v25/23-0643.html. 2

Hanin, B. and Zlokapa, A. Bayesian interpolation with deep linear networks. 120(23):e2301345120, 2023. doi: 10.1073/pnas.2301345120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2301345120. 2

Helias, M. and Dahmen, D. *Statistical Field Theory for Neural Networks*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-46444-8. 4.1, 4.3, 4.4, 4.4, A, A.2

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. Infinite attention: NNGP and NTK for deep attention networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4376–4386. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hron20a.html. 7

Huang, J. and Yau, H.-T. Dynamics of deep neural networks and neural tangent hierarchy. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4542–4551. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/huang20l.html. 2

Krogh, A. and Hertz, J. A simple weight decay can improve generalization. volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf. C.3

Lee, J., Sohl-Dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, Vancouver, 2018. OpenReview.net. URL https://openreview.net/forum?id=B1EA-M-0Z. 1, 2

Lee, J., Schoenholz, S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study. volume 33, pp. 15156–15172. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/

ad086f59924fffe0773f8d0ca22ea712-Paper.pdf. 2

Li, Q. and Sompolinsky, H. Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization. 11(3): 031059, 2021. doi: 10.1103/PhysRevX.11.031059. URL https://journals.aps.org/prx/abstract/10.1103/PhysRevX.11.031059. 1, 1, 2, 3, 4.4, 5, 7, A.4, A.4, A.4.2, A.5, 8

Li, Q. and Sompolinsky, H. Globally gated deep linear networks. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34789–34801. Curran Associates, Inc., 2022. 2, 5

Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. Convergent learning: Do different neural networks learn the same representations? In Storcheus, D., Rostamizadeh, A., and Kumar, S. (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL https://proceedings.mlr.press/v44/li15convergent.html. 2

Lindner, J., Dahmen, D., Krämer, M., and Helias, M. A theory of data variability in neural network bayesian inference. *arXiv preprint arXiv:2307.16695*, 2023. URL https://arxiv.org/abs/2307.16695. 7

Luan, S., Zhang, B., Zhou, S., Chen, C., Han, J., Yang, W., and Liu, J. Gabor convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1254–1262, 2018. doi: 10.1109/WACV.2018.00142. 1

M, S., t, Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. URL http://jmlr.org/papers/v18/17-214.html. 4.1

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 1

Maillard, A., Troiani, E., Martin, S., Krzakala, F., and Zdeborova, L. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=R8znYRjxj3. 2

Matthews, A. G. d. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-. 1, 2

Naveh, G. and Ringel, Z. A self consistent theory of gaussian processes captures feature learning effects in finite CNNs. virtual, 2021a. NeurIPS 2021. URL https://openreview.net/forum?id=vBYwwBxVcsE. 2, 5

Naveh, G. and Ringel, Z. A self consistent theory of gaussian processes captures feature learning effects in finite cnns. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=vBYwwBxVcsE. 1

Naveh, G., Ben David, O., Sompolinsky, H., and Ringel, Z. Predicting the outputs of finite deep neural networks trained with noisy gradients. 104: 064301, Dec 2021. doi: 10.1103/PhysRevE.104.064301. URL https://link.aps.org/doi/10.1103/PhysRevE.104.064301. 2, 4.1, C.3

Neal, R. M. *Bayesian learning for neural networks*. Springer, 1995. 2

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer New York, 1996. doi: 10.1007/978-1-4612-0745-0. URL https://doi.org/10.1007/978-1-4612-0745-0. 1

Paccolat, J., Petrini, L., Geiger, M., Tyloo, K., and Wyart, M. Geometric compression of invariant manifolds in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):044001, apr 2021. doi: 10.1088/1742-5468/abf1f3. URL https://dx.doi.org/10.1088/1742-5468/abf1f3. 1

Pacelli, R., Ariosto, S., Pastore, M., Ginelli, F., Gherardi, M., and Rotondo, P. A statistical mechanics framework for bayesian deep neural networks beyond the infinite-width limit. *Nat. Mach. Intell.*, 5(12):1497–1507, December 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00767-6. URL https://doi.org/10.1038/s42256-023-00767-6. 1, 2, 5, 7, A.4

Pennington, J., Schoenholz, S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d9fc0cdb67638d50f411432d0d41d0ba-Paper.pdf. 2

Petrini, L., Cagnetta, F., Vanden-Eijnden, E., and Wyart, M. Learning sparse features can lead to overfitting in neural networks*. 2023(11):114003, nov 2023. doi: 10.1088/1742-5468/ad01b9. URL https://dx.doi.org/10.1088/1742-5468/ad01b9. 2

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*. 2016. URL https://proceedings.neurips.cc/paper/2016/file/148510031349642de5ca0c544f31b2ef-Paper.pdf. 2

Rai, M. and Rivas, P. A review of convolutional neural networks and gabor filters in object recognition. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1560–1567, 2020. doi: 10.1109/CSCI51800.2020.00289. 1

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006. 5

Refinetti, M., Goldt, S., Krzakala, F., and Zdeborova, L. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8936–8947. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/refinetti21b.html. 2

Risken, H. *The Fokker-Planck Equation*. Springer Verlag Berlin Heidelberg, 1996. doi: 10.1007/978-3-642-61544-3_4. URL https://doi.org/10.1007/978-3-642-61544-3_4. C.3

Roberts, D. A., Yaida, S., and Hanin, B. *The Principles of Deep Learning Theory*. Cambridge University Press, May 2022. doi: 10.1017/9781009023405. URL https://doi.org/10.1017/9781009023405. 1, 2

Rubin, N., Seroussi, I., and Ringel, Z. 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. 1, 2, 5, 5, A.2.3

Rubin, N., Fischer, K., Lindner, J., Dahmen, D., Seroussi, I., Ringel, Z., KrÃ€mer, M., and Helias, M. From kernels to features: A multi-scale adaptive theory of feature learning. *arXiv preprint arXiv:2502.03210*, 2025. 1

Saxe, A., Mcclelland, J., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Represenatations*, 2014. 2

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=H1W1UN9gg. 2

Segadlo, K., Epping, B., van Meegen, A., Dahmen, D., Krämer, M., and Helias, M. Unified field theoretical approach to deep and recurrent neuronal networks. 2022 (10):103401, 2022a. URL https://dx.doi.org/10.1088/1742-5468/ac8e57. 4.1

Segadlo, K., Epping, B., van Meegen, A., Dahmen, D., Krämer, M., and Helias, M. Unified field theoretical approach to deep and recurrent neuronal networks. 2022b. accepted. A

Seroussi, I., Naveh, G., and Ringel, Z. Separation of scales and a thermodynamic description of feature learning in some cnns. 14(1):908, 2023. URL https://doi.org/10.1038/s41467-023-36361-y. 1, 2, 4.3, 5, 6, A.2.3

Simon, J. B., Dickens, M., Karkada, D., and Deweese, M. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=FDbQGCAViI. 2

Templeton, A., Conerly, T., and team, A. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity. 1

Touchette, H. The large deviation approach to statistical mechanics. 478(1):1–69, 2009. ISSN 0370-1573. URL https://www.sciencedirect.com/science/article/pii/S0370157309001410. 4.3, A, A.2, A.4.1

van Meegen, A. and Sompolinsky, H. Coding schemes in neural networks learning classification tasks. pp. 2406.16689, 2024. 1, 1, 2, 5, B, B, B, B, B

Vyas, N., Atanasov, A., Bordelon, B., Morwani, D., Sainathan, S., and Pehlevan, C. Feature-learning networks are consistent across widths at realistic scales. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=LTdfYIvbHc. 2

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 681–688, 2011. 4.1

Williams, C. K. Computation with infinite neural networks. 10(5):1203–1216, 1998. doi: 10.1162/089976698300017412. URL https://doi.org/10.1162/089976698300017412. 1, 2

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xiao18a.html. 2

Yang, A. X., Robeyns, M., Milsom, E., Anson, B., Schoots, N., and Aitchison, L. A theory of representation learning gives a deep generalisation of kernel methods. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39380–39415. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/yang23k.html. 2

Yang, G. and Hu, E. J. Feature Learning in Infinite-Width Neural Networks. 2020. URL https://arxiv.org/abs/2011.14522. 2

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. 2022. doi: 10.48550/arXiv.2203.03466. 2

Yang, G., Simon, J. B., and Bernstein, J. A spectral condition for feature learning, 2024. URL https://arxiv.org/abs/2310.17813. 1, 2

Zavatone-Veth, J. A. and Pehlevan, C. Depth induces scale-averaging in overparameterized linear bayesian neural networks. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 600–607. IEEE, 2021a. doi: 10.1109/ieeeconf53345.2021.9723137. URL http://dx.doi.org/10.1109/IEEECONF53345.2021.9723137. 2

Zavatone-Veth, J. A. and Pehlevan, C. Exact marginal prior distributions of finite bayesian neural networks. virtual, 2021b. NeurIPS 2021. URL https://openreview.net/forum?id=MxE7xFzv0N8. 2

Zavatone-Veth, J. A., Tong, W. L., and Pehlevan, C. Contrasting random and learned features in deep bayesian linear regression. 105:064118, Jun 2022. doi: 10.1103/PhysRevE.105.064118. URL https://journals.aps.org/pre/abstract/10.1103/PhysRevE.105.064118. 2

Zinn-Justin, J. *Quantum field theory and critical phenomena*. Clarendon Press, Oxford, 1996. 4.1

A. General approach to train and test statistics

We are interested in the training discrepancies $\langle\Delta_\alpha\rangle = y_\alpha - \langle f_\alpha\rangle$ with $\langle f_\alpha\rangle$ denoting the mean network output, and in the mean network output $\langle f_*\rangle$ for a test point x_* after conditioning on the training data $\mathcal{D} = \{(x_\alpha, y_\alpha)\}_{1\le\alpha\le P}$. For clarity, in the appendix we make all index notations explicit instead of using $\mathcal{D}$ as in the main text, and denote summations over training data points with Greek letters. We refer to the mean network outputs as predictors. The joint prior distribution for (f, f_*, y) can be computed as in (Segadlo et al., 2022b; Fischer et al., 2024b) and is given by

$$p(f, f_*, y) = \mathcal{N}(y|f, \kappa_0 N^{1-\gamma}) \int d^{(P)}\tilde{f} \int \mathrm{d}\tilde{f}_* \exp\Big(- \sum_{a=1}^{P+1} i\tilde{f}_a f_a + W(i\tilde{f}, i\tilde{f}_*)\Big), \tag{29}$$

$$W(\tilde{f}_{\mathcal{D}}, \tilde{f}_*) = \ln \Big\langle \exp\Big(\sum_{a=1}^{P+1} \tilde{f}_a \sum_{j=1}^{N} w_j\, \phi(h_{aj}) \Big) \Big\rangle_{w_i, h_{ai}}, \tag{30}$$

where we use the shorthands $\int \mathrm{d}^{(P)}\tilde{f} = \prod_{\alpha=1}^{P}\int_{-\infty}^{\infty} d\tilde{f}_\alpha/(2\pi)$, the $P+1$ index corresponds to the test point, and i is the imaginary unit. The i.i.d. distribution of the readin weights V_{kl} implies that $h_{\alpha j} \overset{\text{i.i.d. over j}}{\sim} \mathcal{N}(0, \hat{C}^{(xx)})$ with the covariance matrix of the hidden-layer representation given by

$$\hat{C}^{(xx)} = \begin{bmatrix} C^{(xx)} & \Big\{C^{(xx)}_{\alpha *}\Big\}_{\alpha=1}^{P} \\ \Big\{C^{(xx)}_{*\alpha}\Big\}_{\alpha=1}^{P} & C^{(xx)}_{**} \end{bmatrix}, \tag{31}$$

where $C^{(xx)} = g_v/D\, XX^\mathsf{T}$,and $C^{(xx)}_{*\alpha} = g_v/D\, x_\alpha \cdot x_*$. To keep notation concise, summations over repeated indices on the right are implied in the following.

We may obtain training discrepancies $\langle\Delta_\alpha\rangle$ and the test predictor $\langle f_*\rangle$ from the joint cumulant-generating function $\mathcal{W}$ for the test point defined as

$$\mathcal{W}(j_*|y) = \ln \int df_* \int df \exp(j_* f_*)\, p(f, f_*, y). \tag{32}$$

Taking its derivatives w.r.t. to either training labels y_α or the source term j_* yields the posterior of the desired quantities

$$\langle\Delta_\alpha\rangle = -\kappa_0 N^{1-\gamma} \frac{\partial \mathcal{W}(j_*|y)}{\partial y_\alpha}\Big|_{j_*}, \tag{33}$$

$$\langle f_*\rangle = \frac{\partial \mathcal{W}(j_*|y)}{\partial j_*}\Big|_{j_*=0}, \tag{34}$$

because the outer derivative of the logarithm produces the normalization by the model evidence (marginal likelihood) $1/p(y) = 1/\int df\ \int df_*\, p(f, f^*, y)$.

Likewise, the variances follow as

$$\langle\!\langle \Delta_\alpha \Delta_\beta \rangle\!\rangle = \kappa_0 N^{1-\gamma} - \kappa_0^2 N^{2-2\gamma} \frac{\partial^2 \mathcal{W}(j_*|y)}{\partial y_\alpha \partial y_\beta}\Big|_{j_*=0}, \tag{35}$$

$$\langle\!\langle f_*^2 \rangle\!\rangle = \frac{\partial^2 \mathcal{W}(j_*|y)}{\partial (j_*)^2}\Big|_{j_*=0}, \tag{36}$$

By inserting (29) into (32) and performing the integration over f, we can rewrite $\mathcal{W}$ as

$$\mathcal{W}(j_*|y) = \ln \int d^{(P)}\tilde{f}\, \exp\big(- i y_\alpha \tilde{f}_\alpha - \frac{\kappa_0}{2} N^{1-\gamma} \tilde{f}_\alpha \tilde{f}_\alpha + W(i\tilde{f}_{\mathcal{D}}, j_*)\big). \tag{37}$$

Comparing (33), (35), and (37), we note that y acts as a linear source term for $\tilde{f}_{\mathcal{D}}$, from which we see that the physical meaning of the field $\tilde{f}_{\mathcal{D}}$ is related to the discrepancy between target and network output

$$\langle\Delta_\alpha\rangle = \kappa_0 N^{1-\gamma} \langle i\tilde{f}_\alpha\rangle, \tag{38}$$

$$\langle\!\langle \Delta_\alpha \Delta_\beta \rangle\!\rangle = \kappa_0 N^{1-\gamma} \delta_{\alpha\beta} + \kappa_0^2 N^{2-2\gamma}\, \langle\!\langle \tilde{f}_\alpha \tilde{f}_\beta \rangle\!\rangle. \tag{39}$$

For computational convenience, we now introduce a source term k

$$\mathcal{W}(k,j_*|y)=\ln\int\mathrm{d}\tilde{f}\,\exp\Big(ik_\alpha\tilde{f}_\alpha\underbrace{-iy_\alpha\tilde{f}_\alpha-\frac{\kappa_0}{2}N^{1-\gamma}\tilde{f}_\alpha\tilde{f}_\alpha+W(i\tilde{f},j_*)}_{\mathcal{S}}\Big), \tag{40}$$

allowing us to compute moments of $\tilde{f}$ by differentiating by k instead of y and subsequently setting $k=0$. We define the latter part of the exponent of $\mathcal{W}$ as the action

$$\mathcal{S}(\tilde{f},j_*|y):=-iy_\alpha\tilde{f}_\alpha-\frac{\kappa_0}{2}N^{1-\gamma}\tilde{f}_\alpha\tilde{f}_\alpha+W(i\tilde{f},j_*|y). \tag{41}$$

Depending on the scaling in γ, the network outputs f fully concentrate on their mean values or require corrections due to non-negligible fluctuations. To treat both cases jointly and systematically, we introduce the so-called effective action (Helias & Dahmen, 2020) as

$$\Gamma(\bar{\bar{f}},j_*|y)=\mathrm{extr}_k\,ik^\mathsf{T}\bar{\bar{f}}-\mathcal{W}(k,j_*|y), \tag{42}$$

where we explicitly keep the dependence on the source term j_* for the test point in order to compute parametric derivatives to obtain test point statistics. This corresponds to the Legendre transform of the cumulant-generating function $\mathcal{W}$; in the case that $\mathcal{W}(k,j_*|y)$ has a scaling form, a large deviation principle can be applied and the effective action corresponds to the rate function (Touchette, 2009).

The argument $\bar{\bar{f}}$ is implicitly defined by the stationary point (sometimes referred to as the equation of state)

$$\frac{\partial\Gamma(\bar{\bar{f}},j_*|y)}{\partial\bar{\bar{f}}_\alpha}=ik_\alpha=0, \tag{43}$$

as we set the source term k to 0 by definition. Using the definition of Γ in (42), the extremum condition yields a self-consistency equation for $\bar{\bar{f}}$

$$i\bar{\bar{f}}(j_*)=\frac{\partial\mathcal{W}(k,j_*|y)}{\partial k}. \tag{44}$$

In the following we determine approximations of the Legendre transform $\Gamma(\bar{\bar{f}},j_*|y)$ to different orders of statistical fluctuations, corresponding to different scaling regimes. From the definition of the effective action Γ follows as well that we obtain the mean output on that test point from

$$\langle f_*\rangle=\frac{\partial\mathcal{W}(k,j_*|y)}{\partial j_*}\Big|_{k,j_*=0}=-\frac{\partial\Gamma(\bar{\bar{f}},j_*|y)}{\partial j_*}\Big|_{j^*=0}. \tag{45}$$

A.1. Cumulant-generating function W of the network prior

We recall that $h_{\alpha j}$ is i.i.d Gaussian in j with covariance given by (31). Likewise, the weights w_i are i.i.d., so that we may factorize the expectation to get an overall factor of N in (30)

$$W(\tilde{f}_\mathcal{D},\tilde{f}_*)=N\,\ln\Big\langle\exp\Big(\sum_{a=1}^{P+1}\tilde{f}_a w\,\phi(h_a)\Big)\Big\rangle_{w,h_a}, \tag{46}$$

reducing the expectation over w and h to scalars with regard to the former neuron index j.

A.1.1. Linear activation function

For linear activation function $\phi(h)=h$, we compute first the Gaussian integral in (46) over w and then over the hidden-layer representations h, yielding

$$\begin{aligned}W(\tilde{f}_\mathcal{D},\tilde{f}_*)&=N\,\ln\Big\langle\exp\Big(\tilde{f}^\mathsf{T}h\,w\Big)\Big\rangle_{w,h_a}\\&=-\frac{N}{2}\,\ln\det\Big[\mathbb{I}-\frac{g_w}{N^\gamma}\hat{C}^{(xx)}\tilde{f}\tilde{f}^\mathsf{T}\Big].\end{aligned} \tag{47}$$

Taking the integrals in the opposite order yields

$$
\begin{aligned}
W(\tilde{f}_{\mathcal{D}},\tilde{f}_*) &= N\,\ln\,\Big\langle\,\exp\Big(\frac{1}{2}\tilde{f}^{\mathsf{T}}\,\hat{C}^{(xx)}\,\tilde{f}\,w^2\Big)\Big\rangle_w \\
&= -\frac{N}{2}\,\ln\Big[1-\frac{g_w}{N^\gamma}\tilde{f}^{\mathsf{T}}\hat{C}^{(xx)}\tilde{f}\Big]\,. \qquad (48)
\end{aligned}
$$

The two expressions are identical also by the matrix-determinant lemma.

A.1.2. Non-linear activation function

For a non-linear activation function $\phi(h)$, the integral in (46) in general cannot be solved in a closed form. However, one may perform a cumulant-expansion of the cumulant-generating function (46) in terms of the first two cumulants of ϕ as

$$
\begin{aligned}
W(\tilde{f}_{\mathcal{D}},\tilde{f}_*) &= N\,\ln\,\Big\langle\,\exp\Big(\tilde{f}^{\mathsf{T}}\phi(h)w\Big)\Big\rangle_{w,h_a}\,, \qquad (49)\\
&= N\,\ln\,\Big\langle\,\exp\Big(w\,\tilde{f}^{\mathsf{T}}m+\frac{1}{2}w^2\,\tilde{f}^{\mathsf{T}}\hat{C}^{(\phi\phi)}\,\tilde{f}\,\Big)\Big\rangle_w\,,
\end{aligned}
$$

where we introduced the short hands for the cumulants of ϕ as

$$
\begin{aligned}
m_a &:= \langle\phi(h_a)\rangle_{h_a\sim\mathcal{N}(0,\hat{C}^{(xx)})}\,, \qquad (50)\\
\hat{C}^{(\phi\phi)}_{ab} &:= \big\langle\phi(h_a)\phi(h_b)\big\rangle_{(h_a,h_b)\sim\mathcal{N}(0,\hat{C}^{(xx)})}\,.
\end{aligned}
$$

Taking the expectation over w of (49) yields

$$
\begin{aligned}
W(\tilde{f}_{\mathcal{D}},\tilde{f}_*) &= -\frac{N}{2}\,\ln\Big[1-\frac{g_w}{N^\gamma}\tilde{f}^{\mathsf{T}}\hat{C}^{(\phi\phi)}\tilde{f}\Big] \qquad (51)\\
&\quad+\frac{N}{2}\,\big[\tilde{f}^{\mathsf{T}}m\big]^2\,\Big[\frac{N^\gamma}{g_w}-\tilde{f}^{\mathsf{T}}\hat{C}^{(\phi\phi)}\tilde{f}\Big]^{-1}\,.
\end{aligned}
$$

For point-symmetric activation functions $\phi(-h)=-\phi(h)$, such as erf or tanh activation, the mean $m\equiv 0$ vanishes and comparing to (48) the only replacement that appears is $\hat{C}^{(xx)}\to\hat{C}^{(\phi\phi)}$.

A.2. Tree-level approximation

To compute the output statistics, one technically requires the exact effective action Γ in (42). However, in general it does not have an analytical solution and we instead determine a systematic expansion. A well-established method from both statistical physics and quantum field theory is the loopwise expansion (Helias & Dahmen, 2020), expands the effective action $\Gamma(\bar{\tilde{f}})$ in terms of fluctuations of $\tilde{f}$ around its mean value $\bar{\tilde{f}}$. The lowest-order term of the loopwise expansion is called the tree-level approximation, which hence corresponds to a standard mean-field approximation: one replaces $\tilde{f}_{\mathcal{D}}$ by its mean $\bar{\tilde{f}}$in the action itself

$$
\begin{aligned}
\Gamma_{\mathrm{TL}}(\bar{\tilde{f}},j_*|y) &= -S(\bar{\tilde{f}},j_*|y) \qquad (52)\\
&= iy_\alpha\bar{\tilde{f}}_\alpha+\frac{\kappa_0}{2}N^{1-\gamma}\bar{\tilde{f}}_\alpha\bar{\tilde{f}}_\alpha-W(i\bar{\tilde{f}},j_*). \qquad (53)
\end{aligned}
$$

The average value of $\bar{\tilde{f}}_\alpha$ is given by the equation of state (43) of the effective action

$$
\frac{\partial\Gamma_{\mathrm{TL}}(\bar{\tilde{f}},j_*|y)}{\partial\bar{\tilde{f}}_\alpha}\bigg|_{j_*=0}=0. \qquad (54)
$$

In mean-field scaling ($\gamma=2$) and for $N\to\infty$ this result becomes exact using the Gärtner-Ellis theoreom: the output cumulant-generating function $\mathcal{W}$ in (5) has a scaling form as

$$
iy_\alpha\tilde{f}_\alpha+\frac{\kappa_0}{2N}\tilde{f}_\alpha\tilde{f}_\alpha-W(i\tilde{f}_{\mathcal{D}},j_*)=N\lambda_f(\tilde{f}_{\mathcal{D}}/N) \qquad (55)
$$

with $\lambda_f(k) = iy_\alpha k_\alpha + \frac{\kappa_0}{2} k_\alpha k_\alpha - W(ik, j_*)$. Thus, we can approximate the probability distribution of network outputs as (Touchette, 2009)

$$-p(y|C^{(xx)})/N \approx \Gamma_{\mathrm{TL}}(\bar{\bar{f}}, j_*|y). \tag{56}$$

Due to the strong suppression of fluctuations in mean-field scaling with $N \to \infty$, the tree-level approximation is sufficient to describe the network behavior and in particular

$$\lim_{N\to\infty} -p(y|C^{(xx)})/N = \Gamma_{\mathrm{TL}}(\bar{\bar{f}}, j_*|y). \tag{57}$$

However, in the case of larger output fluctuations as in standard scaling ($\gamma = 1$), we need to take into account the output fluctuations systematically by including higher-order corrections to the tree-level result. We derive the leading-order correction in the following section A.3.

A.2.1. Linear activation function

From the equation of state (54) we obtain a self-consistency equation for $\bar{\bar{f}}_\alpha$ as

$$\bar{\bar{f}} = -i\left(\kappa_0 N^{1-\gamma}\mathbb{I} + C_{\mathrm{TL}}(\bar{\bar{f}})C^{(xx)}\right)^{-1} y, \tag{58}$$

$$C_{\mathrm{TL}}(\bar{\bar{f}}) = g_w N^{1-\gamma}\left[\mathbb{I} + \frac{g_w}{N^\gamma}C^{(xx)}\bar{\bar{f}}\bar{\bar{f}}^\mathsf{T}\right]^{-1}. \tag{59}$$

Using the relation between the statistics of the discrepancies Δ and $\tilde{f}_\mathcal{D}$ (38), we obtain for the training discrepancies

$$\begin{aligned}\langle\Delta_\alpha\rangle &= -\kappa_0 N^{1-\gamma}\frac{\partial\mathcal{W}(k, j_*|y)}{\partial k_\alpha}|_{j_*=0} &(60)\\ &= \kappa_0\left(\kappa_0\mathbb{I} + C_{\mathrm{TL}}(\bar{\bar{f}})C^{(xx)}\right)^{-1}_{\alpha\beta} y_\beta. &(61)\end{aligned}$$

For the test point, we get

$$\begin{aligned}\langle f_*\rangle_{\mathrm{TL}} &= -\frac{\partial\Gamma_{\mathrm{TL}}(\bar{\bar{f}}, j_*|y)}{\partial j_*}|_{j_*=0} = \frac{\partial W(i\bar{\bar{f}}, j_*)}{\partial j_*}|_{j_*=0} &(62)\\ &= \frac{g_w}{N^{1-\gamma}}C^{(xx)}_{*\alpha}\left(\mathbb{I} + \frac{g_w}{N^\gamma}C^{(xx)}\bar{\bar{f}}\bar{\bar{f}}^\mathsf{T}\right)^{-1}_{\alpha\beta} i\bar{\bar{f}}_\beta. &(63)\end{aligned}$$

By substituting the self-consistency equation for $\bar{\bar{f}}$, we obtain

$$\langle f_*\rangle_{\mathrm{TL}} = C_{\mathrm{TL}}(\bar{\bar{f}})_{\delta\alpha}C^{(xx)}_{*\delta}\left[\left(\kappa_0 N^{1-\gamma}\mathbb{I} + C_{\mathrm{TL}}(\bar{\bar{f}})\right)^{-1}\right]_{\alpha\beta} y_\beta\,. \tag{64}$$

For the covariance of the network predictors, we use that $\Delta_\alpha = y_\alpha - f_\alpha$ implies

$$\begin{aligned}\langle\langle\Delta_\alpha\Delta_\beta\rangle\rangle &= \langle(\Delta_\alpha - \langle\Delta_\alpha\rangle)(\Delta_\beta - \langle\Delta_\beta\rangle)\rangle\\ &= \langle(f_\alpha - \langle f_\alpha\rangle)(f_\beta - \langle f_\beta\rangle)\rangle = \langle\langle f_\alpha f_\beta\rangle\rangle, &(65)\end{aligned}$$

so that we may obtain the covariance from (39) as

$$\langle\langle f_\alpha f_\beta\rangle\rangle = \kappa_0 N^{1-\gamma}\delta_{\alpha\beta} - \kappa_0^2 N^{2-2\gamma}\frac{\partial^2\mathcal{W}(j_*|y)}{\partial y_\alpha\partial y_\beta}|_{j_*=0}. \tag{66}$$

Using the involutive property of the Legendre transform, we can express the Hessian of $\mathcal{W}$ by the inverse of the Hessian of Γ using $\mathcal{W}^{(2)} = -(\Gamma^{(2)})^{-1}$, yielding

$$\langle\langle f_\alpha f_\beta\rangle\rangle = \kappa_0 N^{1-\gamma}\delta_{\alpha\beta} + \kappa_0^2 N^{2-2\gamma}\left(\frac{\partial^2\Gamma(\bar{\bar{f}}|y)}{\partial\bar{\bar{f}}\partial\bar{\bar{f}}}|_{j_*=0}\right)^{-1}. \tag{67}$$

In tree level approximation with $\Gamma_{\mathrm{TL}} = -\mathcal{S}$, we then have

$$\langle\langle f_\alpha f_\beta \rangle\rangle = \kappa_0 N^{1-\gamma}\delta_{\alpha\beta} + \kappa_0^2 N^{2-2\gamma} \left(\frac{\partial^2(-\mathcal{S}(\bar{\tilde{f}}|y))}{\partial\bar{\tilde{f}}\partial\bar{\tilde{f}}}|_{j_*=0} \right)^{-1}. \tag{68}$$

Computing the Hessian of the action $\mathcal{S}$, we get

$$\begin{aligned} \frac{\partial^2(-\mathcal{S}(\bar{\tilde{f}}, j_*))}{\partial\bar{\tilde{f}}_\alpha\partial\bar{\tilde{f}}_\beta}\Big|_{j^*=0} &= \kappa_0\delta_{\alpha\beta} + Q_{\mathrm{TL}}\big(\bar{\tilde{f}}\big)\, C_{\alpha\beta} - \frac{2}{N} Q_{\mathrm{TL}}\big(\bar{\tilde{f}}\big)\, C_{\alpha\delta}\bar{\tilde{f}}_\delta \, Q_{\mathrm{TL}}\big(\bar{\tilde{f}}\big)\, C_{\beta\epsilon}\bar{\tilde{f}}_\epsilon & (69)\\ &= A\big(\bar{\tilde{f}}\big)_{\alpha\beta} - \frac{2}{N} Q_{\mathrm{TL}}^2\big(\bar{\tilde{f}}\big)\, \Big[C\bar{\tilde{f}}\bar{\tilde{f}}^{\mathsf{T}} C\Big]_{\alpha\beta}, & (70)\end{aligned}$$

where we use the shorthand $A\big(\bar{\tilde{f}}\big) = \kappa_0\mathbb{I} + Q_{\mathrm{TL}}\big(\bar{\tilde{f}}\big)\, C$. Using the relation between $\tilde{f}$ and the mean deviations from (38) $\langle\Delta_\alpha\rangle = i\kappa_0\bar{\tilde{f}}_\alpha$, we may rewrite this expression as

$$\frac{\partial^2(-\mathcal{S}(\bar{\tilde{f}}, j_*))}{\partial\bar{\tilde{f}}_\alpha\partial\bar{\tilde{f}}_\beta}\Big|_{j^*=0} = A\big(\bar{\tilde{f}}\big)_{\alpha\beta} + \frac{2}{N} Q_{\mathrm{TL}}^2\big(\bar{\tilde{f}}\big)\,\kappa_0^{-2} \Big[C_{\mathcal{D}\mathcal{D}}^{(xx)}\langle\Delta\rangle\langle\Delta\rangle^\top C_{\mathcal{D}\mathcal{D}}^{(xx)}\Big]_{\alpha\beta}, \tag{71}$$

which yields the expression from the main text (26) for the covariance of the network output on training data with $\kappa = \kappa_0 N^{1-\gamma}$ and hence

$$\langle\langle f f^\top \rangle\rangle = \kappa\mathbb{I} - \kappa^2 \left(A\big(\bar{\tilde{f}}\big) + \frac{2Q_{\mathrm{TL}}^2}{N\kappa_0^2} F \right) \tag{72}$$

with $F = C_{\mathcal{D}\mathcal{D}}^{(xx)}\langle\Delta\rangle\langle\Delta\rangle^{\mathsf{T}} C_{\mathcal{D}\mathcal{D}}^{(xx)}$.

A.2.2. Non-linear activation function

Here we consider a point-symmetric non-linear activation function, so that the mean $m = 0$; the extension to the case with $m \neq 0$ is straightforward. Due to the similarity of the expressions (48) and (51), the final expressions here have the same structure as in the case of the linear activation function, Eqs. (60) - (62), but with the replacement $C^{(xx)} \to C^{(\phi\phi)}$ throughout. Because these two cases throughout lead to identical expressions except for this replacement, in the following we denote $C^{(xx)}$ or $C^{(\phi\phi)}$ simply by C.

A.2.3. Non-linear activation function - beyond cumulant expansion

In this section we consider a more fine-grained approach compared to the cumulant expansion in the previous section. Following results from previous sections, the self-consistency equation for $\bar{\tilde{f}}$ in a non-linear network is given by

$$\bar{\tilde{f}} = -i\,(K + \kappa_0/N\mathbb{I})^{-1}\, y \tag{73}$$

with

$$K = g_w/N \frac{\int d^P h\, \phi(h)\, \phi(h)^{\mathsf{T}} \exp\left(-\frac{1}{2}h^T C^{(xx)} h - \frac{g_w}{2N^2}\bar{\tilde{f}}^{\mathsf{T}}\phi(h)\,\phi(h)^T\,\bar{\tilde{f}}\right)}{\int d^P h \exp\left(-\frac{1}{2}h^{\mathsf{T}} C^{(xx)} h - \frac{g_w}{2N^2}\bar{\tilde{f}}^{\mathsf{T}}\phi(h)\,\phi(h)^T\,\bar{\tilde{f}}\right)}. \tag{74}$$

In the case of linear networks, the integrals in (74) are tractable, and are replaced in previous sections by their explicit solution, as can be seen in (47). In the case of non-linear networks, the value of K must be approximated, either via a cumulant-expansion as in the previous section, via a variational Gaussian approximation (VGA) as in (Seroussi et al., 2023), or via higher-order distributional approximations such as the variational Gaussian mixture approximation as in (Rubin et al., 2024), as well as potentially richer approximations. Here we consider the VGA, but we emphasize that our framework is in no way limited to a Gaussian assumption.

The VGA defines a matrix Σ which is chosen such that it minimizes the KL divergence between the distribution of h as it appears in (74) and a Gaussian distribution with covariance Σ. The matrix Σ is then determined by the following equation

$$\left[\Sigma^{-1}\right]_{ij} = \left[C^{(xx)}\right]_{ij} - y^{\mathsf{T}}\,(K_\Sigma + \kappa_0/N\mathbb{I})^{-1}\,\frac{\partial K_\Sigma}{\partial\Sigma_{ij}}\,(K_\Sigma + \kappa_0/N\mathbb{I})^{-1}\, y \tag{75}$$

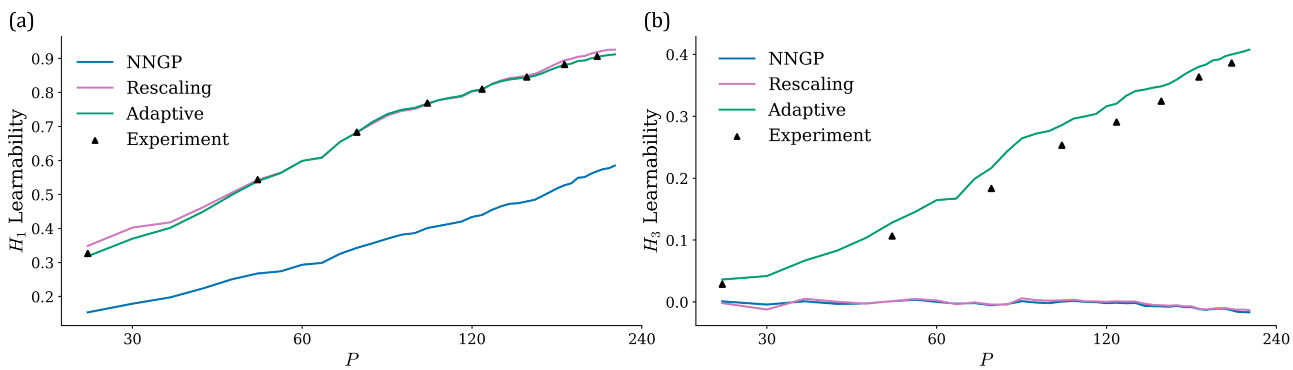

Figure 5. g-learnability of target components in two-layer erf network, trained on the target given in (77), where H_i learnability is defined according to (78), with $g(X) = H_i(Xw_*)$. As can be seen in panel (b), the adaptive approach as derived using VGA predicts that the network will begin to learn higher-order components of y at $P \sim \mathcal{O}(D)$, due to the manifestation of directionally dependent feature learning. On the other hand, kernel methods such as the NNGP and the rescaling approach predict that the network output will be linear, and the cubic component of the target would require $P \sim (D^3)$ training samples. Parameters: $P_{\text{test}} = 4.000$, $N = 1.000$, $D = 30$, $\kappa_0 = 2$, $g_v = g_w = 1$, $\gamma = 1$, $\epsilon = -0.1$.

where $K_\Sigma := g_w/N \left\langle \phi\left(h\right)\phi\left(h\right)^\mathsf{T}\right\rangle_{h\sim\mathcal{N}(0,\Sigma)}$ (which is tractable for activations such as Erf, ReLU). Thus the equation for $\bar{\bar{f}}$ can be written as

$$\bar{\bar{f}}_{\text{VGA}} = -i\left(K_\Sigma + \kappa_0/N\mathbb{I}\right)^{-1} y. \tag{76}$$

We note that for a linear network, the VGA approximation is exact, as the distribution of h is indeed Gaussian. The application of the VGA approximation allows for structural changes to the kernel which do not emerge from a simple cumulant expansion. The structural encoding of certain directions in the kernel could allow the network to learn complex functions of these directions with significantly less data points than would be naively expected. Thus, even a small non-linearity could lead to fundamentally different predictions by the adaptive approach compared to rescaling approaches. To demonstrate this, we consider a teacher-student setting, where the teacher is given by

$$y(x) = H_1(w_*^\mathsf{T} x) + \epsilon H_3(w_*^\mathsf{T} x), \tag{77}$$

with $H_{1,3}$ being the first- and third-order Hermite polynomials, and $x \sim \mathcal{N}(0,\mathbb{I})$. We quantify the ability of a network to learn the different target components (linear or cubic) with a parameter we call *learnability*, defined so that g-learnability is given by

$$g\text{-learnability} := \frac{f\left(X_{\text{test}}\right)^\mathsf{T} g\left(X_{\text{test}}\right)}{y\left(X_{\text{test}}\right)^\mathsf{T} g\left(X_{\text{test}}\right)}, \tag{78}$$

where f, y, g are applied row-wise, and g is any function. Having a learnability = 1 for a given component implies that the network has successfully learned this component. In Fig. 5 we show the theoretical predictions as well as experimental observations for the learnability of both target components ($H_{1,3}(w_*^\mathsf{T} x)$). The experimentally measured learnability of the non-linear component is in good agreement with the adaptive predictions (obtained by VGA), which significantly outperforms the predictions of the kernel approaches (both rescaling and NNGP).

A.3. One-Loop corrections in standard scaling

While in mean-field scaling ($\gamma = 2$) the cumulant-generating function has a scaling form and the network outputs f concentrate, we need to account for their fluctuations in standard scaling ($\gamma = 1$). In the following, we thus set $\gamma = 1$. To leading order, also called one-loop approximation, we have

$$\Gamma_{\text{1-Loop}}(\bar{\bar{f}}, j_*|y) = -\mathcal{S}(\bar{\bar{f}}, j_*) - \frac{1}{2}\log\det(-\mathcal{S}^{(2)}). \tag{79}$$

The self-consistency equation for $\bar{\bar{f}}$ then becomes

$$\frac{\partial\Gamma_{\text{1-Loop}}(\bar{\bar{f}},j_*|y)}{\partial\bar{\bar{f}}_\delta}\bigg|_{j^*=0}=-\frac{\partial\mathcal{S}(\bar{\bar{f}},j_*)}{\partial\bar{\bar{f}}_\delta}\bigg|_{j^*=0}-\frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial\bar{\bar{f}}_\delta}\bigg|_{j^*=0}\overset{!}{=}0. \tag{80}$$

Given the form (48) of the cumulant-generating function W, taking the derivative of $\mathcal{S}$ with respect to $\bar{\bar{f}}$ in (54) yields a different expression than in the previous section

$$-\frac{\partial\mathcal{S}(\bar{\bar{f}},j_*)}{\partial\bar{\bar{f}}_\rho}\bigg|_{j^*=0}=iy_\rho+\kappa_0\bar{\bar{f}}_\rho+Q_{\text{TL}}(\bar{\bar{f}})\,C_{\rho\delta}\bar{\bar{f}}_\delta, \tag{81}$$

$$Q_{\text{TL}}(\bar{\bar{f}})=\frac{g_w}{1+\frac{g_w}{N^\gamma}\bar{\bar{f}}^{\mathsf{T}}C\bar{\bar{f}}}, \tag{82}$$

where now Q_{TL} is a scalar. Note that in this form the tree-level equation for $\bar{\bar{f}}$ (62) can be written as

$$\bar{\bar{f}}=-i\left(\kappa_0N^{1-\gamma}\mathbb{I}+Q_{\text{TL}}(\bar{\bar{f}})\,C\right)^{-1}y, \tag{83}$$

thereby obtaining an expression in which the input kernel C is only rescaled by a scalar, which we call a kernel rescaling expression. For the second and third derivatives, we obtain

$$\begin{aligned}\frac{\partial^2(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta}\bigg|_{j^*=0}&=\kappa_0\delta_{\alpha\beta}+Q_{\text{TL}}(\bar{\bar{f}})\,C_{\alpha\beta}-\frac{2}{N}Q_{\text{TL}}(\bar{\bar{f}})\,C_{\alpha\delta}\bar{\bar{f}}_\delta\,Q_{\text{TL}}(\bar{\bar{f}})\,C_{\beta\epsilon}\bar{\bar{f}}_\epsilon &(84)\\&=A(\bar{\bar{f}})_{\alpha\beta}-\frac{2}{N}Q^2_{\text{TL}}(\bar{\bar{f}})\,\left[C\bar{\bar{f}}\bar{\bar{f}}^{\mathsf{T}}C\right]_{\alpha\beta}, &(85)\\\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial\bar{\bar{f}}_\delta}\bigg|_{j^*=0}&=-\frac{2}{N}Q^2_{\text{TL}}(\bar{\bar{f}})\,\left[C_{\alpha\beta}C_{\delta\epsilon}\bar{\bar{f}}_\epsilon+C_{\alpha\delta}C_{\beta\epsilon}\bar{\bar{f}}_\epsilon+C_{\beta\delta}C_{\alpha\epsilon}\bar{\bar{f}}_\epsilon\right] &(86)\\&\quad+\frac{8}{N^2}Q^3_{\text{TL}}(\bar{\bar{f}})\,C_{\alpha\alpha'}\bar{\bar{f}}_{\alpha'}\,C_{\beta\beta'}\bar{\bar{f}}_{\beta'}\,C_{\delta\delta'}\bar{\bar{f}}_{\delta'} &(87)\\&=-\frac{2}{N}Q^2_{\text{TL}}(\bar{\bar{f}})\,\left(C_{\alpha\beta}\left[C\bar{\bar{f}}\right]_\delta+C_{\alpha\delta}\left[C\bar{\bar{f}}\right]_\beta+C_{\beta\delta}\left[C\bar{\bar{f}}\right]_\alpha\right) &(88)\\&\quad+\frac{8}{N^2}Q^3_{\text{TL}}(\bar{\bar{f}})\,\left[C\bar{\bar{f}}\right]_\alpha\left[C\bar{\bar{f}}\right]_\beta\left[C\bar{\bar{f}}\right]_\delta. &(89)\end{aligned}$$

Here, we use the shorthand $A(\bar{\bar{f}})=\kappa_0\mathbb{I}+Q_{\text{TL}}(\bar{\bar{f}})\,C$. Overall, we obtain

$$\bar{\bar{f}}_\delta=\left[A(\bar{\bar{f}})^{-1}\right]_{\delta\epsilon}\left[-iy_\epsilon+\frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial\bar{\bar{f}}_\epsilon}\bigg|_{j^*=0}\right]. \tag{90}$$

Similarly to the previous section, the training discrepancies are given by

$$\langle\Delta_\alpha\rangle=i\kappa_0\bar{\bar{f}}_\alpha. \tag{91}$$

For the test predictor, we have

$$\begin{aligned}\langle f^*\rangle_{\text{1-Loop}}&=-\frac{\partial\Gamma_{\text{1-Loop}}(\bar{\bar{f}},j_*|y)}{\partial j^*}\bigg|_{j^*=0} &(92)\\&=-\frac{\partial W(\bar{\bar{f}},j_*)}{\partial j^*}\bigg|_{j^*=0}-\frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial j_*}\bigg|_{j^*=0} &(93)\\&=Q_{\text{TL}}(\bar{\bar{f}})\,C_{*\alpha}\bar{\bar{f}}_\alpha-\frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial j_*}\bigg|_{j^*=0}. &(94)\end{aligned}$$

The appearing derivatives of the action are structurally similar but we replace the training point x_γ by the test point x_*, yielding

$$\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial j_*}\Big|_{j^*=0} = -\frac{2}{N}Q_{\mathrm{TL}}^2(\bar{\bar{f}})\left(C_{\alpha\beta}\left[C\bar{\bar{f}}\right]_* + C_{\alpha*}\left[C\bar{\bar{f}}\right]_\beta + C_{\beta*}\left[C\bar{\bar{f}}\right]_\alpha\right) \tag{95}$$

$$+\frac{8}{N^2}Q_{\mathrm{TL}}^3(\bar{\bar{f}})\,\left[C\bar{\bar{f}}\right]_\alpha\left[C\bar{\bar{f}}\right]_\beta\left[C\bar{\bar{f}}\right]_*. \tag{96}$$

When solving these equations, we backtransform to the imaginary variables $\bar{\bar{f}} \mapsto i\bar{\bar{f}}$, which changes multiple signs and absorbs the appearing imaginary units.

A.4. Kernel rescaling approach

We here derive the results by Li & Sompolinsky (2021) and Ariosto et al. (2023) in our multi-scale adaptive theory: Using that $h_{\alpha j} \sim \mathcal{N}(0, C^{(xx)})$ i.i.d. over the neuron index j, we can rewrite the cumulant-generating function W (30) for the case of a linear activation function $\phi(h) = h$ conditioned on readout weights w as

$$W(\tilde{f}_{\mathcal{D}}|w) = \ln\left\langle \exp(-\tilde{f}_\alpha w_j h_{\alpha j})\right\rangle_{h_{\alpha j}} = \frac{1}{2}\tilde{f}_\alpha C^{(xx)}_{\alpha\beta}\tilde{f}_\beta\,\|w\|^2, \tag{97}$$

where we drop the test point here to keep notation concise. The result for the test point will follow naturally later. Likewise, performing a cumulant expansion up to second order in ϕ for a point-symmetric activation function as in (49), we obtain

$$W(\tilde{f}_{\mathcal{D}}|w) = \ln\left\langle \exp(\tilde{f}_\alpha w_j \phi(h_{\alpha j})\right\rangle_{h_{\alpha j}} = \frac{1}{2}\tilde{f}_\alpha C^{(\phi\phi)}_{\alpha\beta}\tilde{f}_\beta\,\|w\|^2, \tag{98}$$

where $C^{(\phi\phi)}$ is defined as in (50). As in the adaptive approach, we here again write C for short to refer to $C^{(xx)}$ in the case of linear activation function and to $C^{(\phi\phi)}$ in the case of the non-linear point symmetric activation.

We observe that the readout weights only appear in the form of the squared norm $\|w\|^2$. The distribution of the network output is hence

$$p(y,f|C) = \mathcal{N}(y|f,\kappa_0)\int \mathrm{d}\tilde{f}_{\mathcal{D}}\,\big\langle \exp\big(-i\tilde{f}_\alpha f_\alpha - \frac{1}{2}\tilde{f}_\alpha C_{\alpha\beta}\tilde{f}_\beta\,\|w\|^2\big)\big\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{g_w}{N})}. \tag{99}$$

Since both, the prior measure of the weights $w \sim \mathcal{N}(w|0, g_w N^{-1}) \propto \exp(N\,\|w\|^2/2g_w)$ and the explicit appearance of w, is only in the form of $\|w\|^2$, we may introduce this quantity as an auxiliary variable, which we name $Q := \|w\|^2 = \sum_{i=1}^N w_i^2$ and which corresponds to the Euclidean norm of the readout weight vector w. Note that, given $\|w\|^2$, the integral over $\tilde{f}_{\mathcal{D}}$ simply yields $f|_{\|w\|^2} \sim \mathcal{N}(0, \|w\|^2 C)$, so

$$p(y,f|C) = \mathcal{N}(y|f,\kappa_0)\int dQ\,\mathcal{N}(f|0,Q\,C)\,p(Q). \tag{100}$$

Here the distribution of the squared norm is

$$p(Q) = \big\langle\delta[-Q+\|w\|^2]\big\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{g_w}{N})} \tag{101}$$

$$= \int_{-i\infty}^{i\infty}\frac{d\tilde{Q}}{2\pi i}\,\big\langle \exp\big(\tilde{Q}\big[-Q+\|w\|^2\big]\big)\big\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{g_w}{N})} \tag{102}$$

$$= \int_{-i\infty}^{i\infty}\frac{d\tilde{Q}}{2\pi i}\,\exp\big(-\tilde{Q}\,Q + W(\tilde{Q})\big), \tag{103}$$

where $W(\tilde{Q}) = \ln\langle\exp\big(\tilde{Q}\,\|w\|^2\big)\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{g_w}{N})}$ is the cumulant-generating function of Q . Using that the w_j are i.i.d, we get

$$W(\tilde{Q}) = N\,\ln\big\langle\exp\big(\tilde{Q}\|w\|^2\big)\big\rangle_{w\sim\mathcal{N}(0,g_w/N)} \tag{104}$$

$$= -\frac{N}{2}\,\ln\big[1-\frac{2g_w}{N}\tilde{Q}\big], \tag{105}$$

where we performed the one-dimensional Gaussian integral over w. Up to here, all steps are exact.

(100) shows that the auxiliary variable Q being a scalar may only carry fluctuations of the overall scaling of the kernel and hence all descriptions and approximations in terms of Q can only change the scale of the kernel, which is consistent with the results in (Li & Sompolinsky, 2021; Pacelli et al., 2023).

A.4.1. Approximation of network prior for wide networks

One expects that Q concentrates for large N according to the central limit theorem since $Q = \|w\|^2 = \sum_{i=1}^N w_i^2$ with i.i.d. $w_i \sim \mathcal{N}(0, g_w/N)$. The cumulant-generating function W can be written as a scaling form $\lambda_N(k) := N^{-1}\, W(N\, k) = -\frac{1}{2}\, \ln\big[1 - 2g_w k\big]$ and its limit $N \to \infty$ then exists trivially, so that we may approximate $p(Q)$ with the Gärtner-Ellis theorem (Touchette, 2009) as

$$\ln\, p(Q) \simeq \sup_{\tilde{Q}}\, -Q\tilde{Q} + W(\tilde{Q}) \tag{106}$$

$$= -\frac{N}{2g_w}\,\big[1 - \frac{g_w}{Q}\big]\, Q - \frac{N}{2}\, \ln\big[\frac{g_w}{Q}\big] \tag{107}$$

$$= -\frac{N}{2}\,\big[\frac{Q}{g_w} - 1 - \ln\, \frac{Q}{g_w}\big] =: -\Gamma(Q). \tag{108}$$

Intuitively, by the scaled cumulant-generating function of the form $N\, W(\tilde{Q}/N) = -\frac{N}{2}\, \ln\big[1 - 2g_w\, \frac{\tilde{Q}}{N}\big]$ the mean of is of order $\langle Q\rangle = \mathcal{O}(1)$ and all higher-order cumulants of Q are being suppressed by at least $\mathcal{O}(N^{-1})$. So on exponential scales, one may parametrize the probability by the mean, namely one obtains the distribution of Q from the rate function as $e^{-\Gamma(Q)}$. To obtain (106), the supremum condition has been used $0 \overset{!}{=} -Q + g_w\,\big[1 - 2g_w\, \frac{\tilde{Q}}{N}\big]^{-1}$, solved for $1 - \frac{2g_w}{N}\,\tilde{Q} = \frac{g_w}{Q}$ and $\tilde{Q} = \frac{N}{2g_w}\,\big[1 - \frac{g_w}{Q}\big]$ and inserted into the first line of (106) to obtain the second line. The rate function, being the Legendre transform of W, obeys the equation of state

$$\frac{d}{dQ}\Gamma(Q) = \tilde{Q} = \frac{N}{2g_w}\,\big[1 - \frac{g_w}{Q}\big]. \tag{109}$$

So the final expression for the joint probability of y and f, the network prior, is

$$\begin{aligned} p(y, f|C) &\simeq \mathcal{N}(y|f, \kappa_0) \int dQ\, \mathcal{N}(f|0, Q\, C)\, e^{-\Gamma(Q)}, \\ &= \int dQ\, e^{S(Q|f,y)}, \end{aligned} \tag{110}$$

where the action $S(Q|f, y)$ is

$$\begin{aligned} S(Q|f, y) = &-\frac{\|y - f\|^2}{2\kappa_0} - \frac{P}{2}\ln\kappa_0 \\ &- \frac{1}{2} f^{\mathsf{T}}\,\big(QC\big)^{-1} f - \frac{1}{2}\ln\det\big(QC\big) - \Gamma(Q) + \text{const.} \end{aligned} \tag{111}$$

A.4.2. Maximum a posteriori estimate for Q

To obtain the posterior distribution for Q we marginalize (110) over the network outputs f, which yields

$$\begin{aligned} p(y|C) &\equiv \int df\, p(y, f|C) \\ &= \int dQ\, \exp\big(S(Q|y)\big), \end{aligned} \tag{112}$$

which yields the action

$$S(Q|y) = -\frac{1}{2} y^{\mathsf{T}}\,\big(QC + \kappa_0\mathbb{I}\big)^{-1} y - \frac{1}{2}\ln\det\big(QC + \kappa_0\mathbb{I}\big) - \Gamma(Q), \tag{113}$$

and which reproduces Eq. A11 in Li & Sompolinsky (2021) after inserting the rate function (106) and using $C = C^{(xx)}$ for the linear network. It likewise reproduces Eq. (33) in (Ariosto et al., 2023) when inserting $C = C^{(\phi\phi)}$ for the non-linear activation function.

When computing the maximum a posteriori value Q_{LS}, it only depends on the numerator of

$$p(Q|y) = \frac{p(y|Q)\,p(Q)}{p(y)}, \tag{114}$$

since the form of (112) is $p(y) = \int dQ\, p(y|Q)\,p(Q)$. Thus, computing the Q-integral in saddle point approximation comprises to the maximum a posteriori (MAP) Q_{LS} as $\ln\, p(y|Q)\,p(Q) = S(Q|y) + \mathrm{const}$ has the same stationary point as $p(y|Q)\,p(Q)$.

The length $Q = \|w\|^2$ in their theory is obtained by the maximum of (113), which is given by

$$\begin{aligned} 0 \overset{!}{=} \frac{\partial S}{\partial Q} = \frac{1}{2} y^{\mathsf{T}} \big(QC + \kappa_0 \mathbb{I}\big)^{-1} C \,\big(QC + \kappa_0 \mathbb{I}\big)^{-1} y \\ - \operatorname{tr} C\big(QC + \kappa_0 \mathbb{I}\big)^{-1} - \frac{N}{2}\big(\frac{1}{g_w} - \frac{1}{Q}\big). \end{aligned} \tag{115}$$

This self-consistency equation yields the tree-level approximation for $\|w\|^2 = Q_{\mathrm{LS}}$.

A.4.3. Predictor statistics

To obtain predictions beyond the length of the readout $\|w\|$, we start from 112. We obtain statistics of the training discrepancies $\Delta = y_\alpha - f_\alpha$ from

$$\begin{aligned} \frac{\partial}{\partial y_\alpha} \ln p(y|C) &\overset{\text{MAP}\,Q}{\simeq} \frac{d}{dy_\alpha} \sup_Q S(Q|y) \qquad (116) \\ &= \frac{\partial}{\partial y_\alpha} S(Q_{\mathrm{LS}}|y) + \underbrace{\frac{\partial S}{\partial Q}}_{=0} \frac{\partial Q}{\partial y_\alpha}\Big|_{Q=Q^*}, \qquad (117) \end{aligned}$$

where the derivative by Q vanishes because Q_{LS} has been determined by the supremum condition as the stationary point of the action. The partial derivative by y_α only acts on $-y^{\mathsf{T}}\left(QC + \kappa_0 \mathbb{I}\right)^{-1} y/2$ in the expression for (113)

$$\langle \Delta_\alpha \rangle = \kappa_0 \left(Q_{\mathrm{LS}} C + \kappa_0 \mathbb{I}\right)^{-1} y. \tag{118}$$

In consequence, the test predictor is identical to the NNGP predictor with a different regularizer κ_0/Q_{LS}

$$\langle f_* \rangle_{\mathrm{LS}} = \left[C_{*\alpha}\right] \left(C + \kappa_0/Q_{\mathrm{LS}} \mathbb{I}\right)^{-1}_{\alpha\beta} y_\beta. \tag{119}$$

To compute the variance of the predictor, we generalize(99) such that instead of the variance $\kappa_0 \mathbb{I}$ in $\mathcal{N}(y|f, \kappa_0 \mathbb{I})$, we insert a general covariance matrix K into the Gaussian measure $\mathcal{N}(y|f, K)$ and perform an integration over f

$$p(y|K, C^{(xx)}) = \int df\, \mathcal{N}(y|f, K) \Big\langle \prod_{\alpha=1}^{P} \delta\left[f_\alpha - \sum_{i=1}^{N} w_i\, \phi(h_{\alpha i})\right] \Big\rangle_{w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{g_w}{N}),\quad h_{\alpha i} \overset{\text{i.i.d. over } i}{\sim} \mathcal{N}(0, C^{(xx)})}. \tag{120}$$

The presence of the general matrix K allows us to measure the statistics of the discrepancies $\Delta_\alpha = y_\alpha - z_\alpha$, because writing the Gaussian $\mathcal{N}(y|f, K) \propto \exp\big(-\frac{1}{2}(y-f)^{\mathsf{T}} K^{-1}(y-f) + \frac{1}{2}\, \ln\, \det\left(K^{-1}\right)\big)$ explicitly we observe that derivatives by $\left[K^{-1}\right]_{\alpha\beta}$ yield

$$\frac{\partial}{\partial [K]^{-1}_{\alpha\beta}} \ln p(y|K, C)\Big|_{K=\kappa_0 \mathbb{I}} = -\frac{1}{2} \langle (y-f)_\alpha (y-f)_\beta \rangle + \frac{1}{2}\, \kappa_0\, \delta_{\alpha\beta}. \tag{121}$$

With the same manipulations that led to (112) one then has

$$p(y|K,C) \equiv \int dQ\, \exp\big(S(Q|K,z)\big), \tag{122}$$

where the action, corresponding to (113), is

$$S(Q|y,K) = -\frac{1}{2}y^\mathsf{T}\left(C+K\right)^{-1}y - \frac{1}{2}\ln\det\left(C+K\right) - \frac{N}{2}\left(\frac{Q}{g_w} - \ln Q\right)\Big|_{C=QC}. \tag{123}$$

So in the approximation replacing Q by its MAP Q_{LS} we get

$$\frac{\partial}{\partial[K]^{-1}_{\alpha\beta}}\ln p(y|K,C)\Big|_{K=\kappa_0\mathbb{I}} = \frac{d}{d[K]^{-1}_{\alpha\beta}}\sup_Q S(Q|y,K)\Big|_{K=\kappa_0\mathbb{I}} \tag{124}$$

$$= \frac{\partial}{\partial[K]^{-1}_{\alpha\beta}} S(Q_{\text{LS}}|y,K)\Big|_{K=\kappa_0\mathbb{I}}, \tag{125}$$

where the inner derivative by $\partial S/\partial Q$ drops out due to stationarity at Q_{LS}, which is given by the solution of (115). The latter partial derivative evaluates to

$$\frac{\partial}{\partial[K]^{-1}_{\alpha\beta}} S(Q|y,K)\Big|_{K=\kappa_0\mathbb{I}} = \Big[-\frac{1}{2}K\left[c+K\right]^{-1}yy^\mathsf{T}\left[c+K\right]^{-1}K + \frac{1}{2}K\left(c+K\right)^{-1}K\Big]_{\alpha\beta}\Big|_{K=\kappa_0\mathbb{I},\,c=Q_{\text{LS}}C} \tag{126}$$

$$= \kappa_0^2\Big[-\frac{1}{2}\left[c+\kappa_0\mathbb{I}\right]^{-1}yy^\mathsf{T}\left[c+\kappa_0\mathbb{I}\right]^{-1} + \frac{1}{2}\left(c+\kappa_0\mathbb{I}\right)^{-1}\Big]_{\alpha\beta}\Big|_{c=Q_{\text{LS}}C}, \tag{127}$$

where we used that $\partial K_{\gamma\delta}/\partial[K]^{-1}_{\alpha\beta} = -K_{\gamma\alpha}\,K_{\beta\delta}$, which follows by symmetry from $\partial K^{-1}_{\gamma\delta}/\partial K_{\alpha\beta} = -K^{-1}_{\gamma\alpha}\,K^{-1}_{\beta\delta}$.

So the second moment of the discrepancies with (121) is

$$\begin{aligned}\langle\Delta_\alpha\Delta_\beta\rangle &= \kappa_0\delta_{\alpha\beta} + \kappa_0^2\Big[\left[c+\kappa_0\mathbb{I}\right]^{-1}yy^\mathsf{T}\left[c+\kappa_0\mathbb{I}\right]^{-1} - \left(c+\kappa_0\mathbb{I}\right)^{-1}\Big]_{\alpha\beta}\Big|_{c=Q_{\text{LS}}C}\\ &= \langle\Delta_\alpha\rangle\langle\Delta_\beta\rangle + \kappa_0\delta_{\alpha\beta} - \kappa_0^2\left(c+\kappa_0\mathbb{I}\right)^{-1}_{\alpha\beta}\Big|_{c=Q_{\text{LS}}C},\end{aligned} \tag{128}$$

where we used (118) in the last step. Because $\Delta = y - f$ and the target label y do not fluctuate, the latter two terms in (128) are the variance

$$\langle\langle\Delta_\alpha,\Delta_\beta\rangle\rangle = \langle\langle f_\alpha, f_\beta\rangle\rangle \tag{129}$$

$$= \kappa_0\delta_{\alpha\beta} - \kappa_0^2\left(c+\kappa_0\mathbb{I}\right)^{-1}_{\alpha\beta}\Big|_{c=Q_{\text{LS}}C} \tag{130}$$

$$= c - c\left[c+\kappa_0\mathbb{I}\right]^{-1}c\Big|_{c=Q_{\text{LS}}C}, \tag{131}$$

which is the usual expression for the variance of the NNGP predictor of a Gaussian process with the kernel $c = Q_{\text{LS}}\,C$.

A.5. Connecting kernel rescaling and adaptive approach

While the kernel rescaling approach holds in the proportional limit $N \propto P \to \infty$, the one-loop approximation holds also for large but finite $P, N \gg 1$. As we have seen in Fig. 3 in the main text, they yield almost identical results in certain settings. By considering the proportional limit, we may connect these two approaches: some correction terms vanish in this limit, leaving only a scalar term.

To this end, we look at the scaling of each correction term with both P and N. We have

$$-S^{(2)} = \kappa_0\mathbb{I} + Q_{\text{TL}}\,C^{(xx)} - \frac{2}{N}Q_{\text{TL}}^2\,C^{(xx)}\,\bar{\bar{f}}\bar{\bar{f}}^\mathsf{T}C^{(xx)} \tag{132}$$

$$= \kappa_0\mathbb{I} + Q_{\text{TL}}\,C^{(xx)} + \mathcal{O}(1/N), \tag{133}$$

since $(Q\,C^{(xx)}+\kappa)\,\bar{\bar{f}}\propto y=\mathcal{O}(1)$ and thus also $C^{(xx)}\bar{\bar{f}}=\mathcal{O}(1)$. Here, we drop the dependence of Q_{TL} on $\bar{\bar{f}}$ for brevity. The fluctuation correction is given by

$$\frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S}(\bar{\bar{f}},j_*))}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial\bar{\bar{f}}_\delta}\bigg|_{j^*=0} \tag{134}$$

$$=-\frac{1}{N}Q_{\mathrm{TL}}^2\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\left(C^{(xx)}_{\alpha\beta}\left[C^{(xx)}\bar{\bar{f}}\right]_\gamma+C^{(xx)}_{\alpha\gamma}\left[C^{(xx)}\bar{\bar{f}}\right]_\beta+C^{(xx)}_{\beta\gamma}\left[C^{(xx)}\bar{\bar{f}}\right]_\alpha\right) \tag{135}$$

$$+\frac{4}{N^2}Q_{\mathrm{TL}}^3\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\left[C^{(xx)}\bar{\bar{f}}\right]_\alpha\left[C^{(xx)}\bar{\bar{f}}\right]_\beta\left[C^{(xx)}\bar{\bar{f}}\right]_\gamma \tag{136}$$

Looking at the individual terms, we have for the first term in the second line

$$\frac{1}{N}Q_{\mathrm{TL}}^2\,\mathrm{Tr}\left[(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}C^{(xx)}+\mathcal{O}(1/N))^{-1}\,C^{(xx)}\right]C^{(xx)}_{\gamma\delta}\bar{\bar{f}}_\delta=\mathcal{O}(P/N), \tag{137}$$

where the factor P results from the appearing trace. Assuming the regularization noise κ_0 to be small compared to the kernel $C^{(xx)}$, we see for the other terms that they scale as

$$\frac{1}{N}Q_{\mathrm{TL}}^2\sum_{\alpha\beta}(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}C^{(xx)}+\mathcal{O}(1/N))^{-1}_{\beta\alpha}\left(C^{(xx)}_{\alpha\gamma}C^{(xx)}_{\beta\delta}\bar{\bar{f}}_\delta+C^{(xx)}_{\beta\gamma}C^{(xx)}_{\alpha\delta}\bar{\bar{f}}_\delta\right)\approx\frac{4}{N}Q_{\mathrm{TL}}C^{(xx)}_{\gamma\delta}\bar{\bar{f}}_\delta=\mathcal{O}(1/N), \tag{138}$$

$$\frac{4}{N^2}Q_{\mathrm{TL}}^3\sum_{\alpha\beta}(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}C^{(xx)}+\mathcal{O}(1/N))^{-1}_{\beta\alpha}\left[C^{(xx)}\bar{\bar{f}}\right]_\alpha\left[C^{(xx)}\bar{\bar{f}}\right]_\beta\left[C^{(xx)}\bar{\bar{f}}\right]_\gamma=\mathcal{O}(P/N^2). \tag{139}$$

In the proportional limit $P\propto N\to\infty$, only the first term does not vanish and the self-consistency equation for $\bar{\bar{f}}$ becomes

$$iy+\kappa_0\bar{\bar{f}}+Q_{\mathrm{TL}}(\bar{\bar{f}})\,C^{(xx)}\bar{\bar{f}}-\frac{1}{N}Q_{\mathrm{TL}}^2(\bar{\bar{f}})\,\mathrm{Tr}\left[(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}(\bar{\bar{f}})\,C^{(xx)})^{-1}C^{(xx)}\right]C^{(xx)}\,\bar{\bar{f}}\stackrel{!}{=}0, \tag{140}$$

yielding

$$i\bar{\bar{f}}=\left(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}(\bar{\bar{f}})\left(1-\frac{1}{N}Q_{\mathrm{TL}}(\bar{\bar{f}})\,\mathrm{Tr}\big[(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}(\bar{\bar{f}})\,C^{(xx)})^{-1}C^{(xx)}\big]\right)C^{(xx)}\right)^{-1}y. \tag{141}$$

The rescaling factor is thus given by

$$Q_{\text{1-Loop}}\coloneqq Q_{\mathrm{TL}}-\frac{1}{N}Q_{\mathrm{TL}}^2\mathrm{Tr}\big[(\kappa_0\mathbb{I}+Q_{\mathrm{TL}}C^{(xx)})^{-1}C^{(xx)}\big], \tag{142}$$

where $Q_{\mathrm{TL}}=Q_{\mathrm{TL}}(\bar{\bar{f}})$ depends on the self-consistent solution in (141). The tree-level solution is the leading term here and receives a correction due to the output fluctuations. We cannot directly compare the expression for this rescaling factor to the one in (Li & Sompolinsky, 2021), since the latter is given by the self-consistency equation (113) and the former by the self-consistency equation (141) for the training discrepancies inserted into (142). Nevertheless, Fig. 3 in the main text shows that these two expressions yield the same value and thus the same predictions for the mean discrepancies numerically.

B. Other scaling regimes

We here expose the relation of our theory to the work by van Meegen & Sompolinsky (2024) where yet another scaling regime is considered. We start from the effective action in tree-level approximation (53) and the form for W given by (46)

$$\Gamma_{\mathrm{TL}}(\bar{\bar{f}},j_*|y)=iy_\alpha\bar{\bar{f}}_\alpha+\frac{\kappa_0}{2}N^{-1}\bar{\bar{f}}_\alpha\bar{\bar{f}}_\alpha-W(i\bar{\bar{f}},j_*) \tag{143}$$

$$W(\tilde{f}_{\mathcal{D}},\tilde{f}_*)=N\,\ln\left\langle\int dw\,\exp\left(\sum_{a=1}^{P+1}\tilde{f}_a\frac{w}{N}\,\phi(h_a)-P\,\frac{w^2}{2g_w}\right)\right\rangle_{h_a\sim\mathcal{N}(0,C^{(xx)})}+\text{const.},$$

where, as in the original work, we scale the variance of the readout weights by an additional factor P (see ref. (van Meegen & Sompolinsky, 2024), their Section "1. Weight Posterior" second last paragraph). In the proportional limit $P\propto N$ it entails $w \propto N^{-\frac{3}{2}}$, which scales down the readout weights even more strongly than mean-field scaling. This scaling allows taking the integral over w in saddle point approximation because the exponent $\sum_{a=1}^{P+1} \tilde{f}_a w\, \phi(h_a) - P\, w^2/2g_w$ scales with P. With (57) one therefore has with $l = (\bar{\bar{f}}, j_*)/N$

$$\begin{aligned} -\ln\, p(y|C^{(xx)})/N &= \Gamma_{\mathrm{TL}}(N\, l\,|y)/N \\ &= iy^{\mathsf{T}} l + \frac{\kappa_0}{2}\, l^{\mathsf{T}} l \\ &- \ln \int dw \left\langle \exp\Big(- P\, \frac{w^2}{2g_w} + w\, \big[il^{\mathsf{T}}\phi(h)\big]\Big)\right\rangle_{h\sim\mathcal{N}(0,C^{(xx)})} + \mathrm{const.} \\ &\overset{P\propto N}{\simeq}\; iy^{\mathsf{T}} l + \frac{\kappa_0}{2}\, l^{\mathsf{T}} l \\ &- \mathrm{extr}_w \Big\{ - P\, \frac{w^2}{2g_w} + \ln \Big\langle \exp\Big(w\, \big[il^{\mathsf{T}}\phi\big]\Big)\Big\rangle_{h\sim\mathcal{N}(0,C^{(xx)})} \Big\} + \mathrm{const.} \end{aligned} \tag{144}$$

The extremum condition may have multiple degenerate values w_γ that appear with the relative log probability given by the entropy

$$\ln\, p_\gamma = -P\, \frac{w_\gamma^2}{2g_w} + \ln\, \Big\langle \exp\Big(w_\gamma\, \big[il^{*\mathsf{T}}\phi\big]\Big)\Big\rangle_{h\sim\mathcal{N}(0,C^{(xx)})} + \mathrm{const.}, \tag{145}$$

which corresponds to Eq. (B10) in (van Meegen & Sompolinsky, 2024) and the constant is determined such that $\sum_\gamma p_\gamma = 1$. The distribution of w hence approximates the posterior $p(w) \simeq \sum_\gamma p_\gamma\, \delta(w - w_\gamma^*)$, which is used in (144) to obtain the equation of state (54) for $\bar{\bar{f}}$ by taking a partial derivative of by l_α, which yields

$$y_\alpha = \sum_\gamma p_\gamma\, w_\gamma^*\, \big[\phi_\alpha\big]_{l^*,w^*} + \kappa_0 i l^*\,, \tag{146}$$

which corresponds to Eq. (B17) in (van Meegen & Sompolinsky, 2024), where the expectation value $\big[\ldots\big]_{l^*,w^*}$ is given by

$$[\ldots]_{l^*,w_\gamma^*} := \frac{\int d^P h\, \ldots \exp\Big(w_\gamma^*\, \big[il^{*\mathsf{T}}\phi(h)\big] - \frac{1}{2} h^{\mathsf{T}} C^{(xx)} h\Big)}{\int d^P h\, \exp\Big(w_\gamma^*\, \big[il^{*\mathsf{T}}\phi(h)\big] - \frac{1}{2} h^{\mathsf{T}} C^{(xx)} h\Big)}\,. \tag{147}$$

The extremum condition appearing in (144) leads to stationary values of the weights:

$$w_\gamma^* = \frac{g_w}{P}\, il^{*\mathsf{T}}\, \big[\phi(h)\big]_{l^*,w_\gamma^*}\,, \tag{148}$$

which corresponds to eq. B16 of (van Meegen & Sompolinsky, 2024).

In summary, the difference to our work is that in the scaling considered in the work by van Meegen & Sompolinsky (2024), $w \propto 1/(N\sqrt{P})$, the readout weights w tend to concentrate to non-fluctuating values w_γ^*, while in mean-field scaling $w \propto 1/N$, which we treat in the main part of our work, only the network outputs concentrate. Compared to the standard scaling $w \propto 1/N$ both scalings only require a tree-level approximation to describe the mean predictor whereas the fluctuations in the standard scaling require the inclusion of one-loop corrections.

C. Details of experiments

C.1. Self-consistency equations for numerics

In App. A, we derive train and test statistics in a framework involving imaginary variables $\bar{\bar{f}}$. To solve the resulting self-consistency equations, we need to account for their imaginary nature and substitute in all of the results above $\bar{\bar{f}} \to i\bar{\bar{f}}$, changing various signs in the process. The final expressions read as follows: In tree-level approximation, we have

$$\bar{\bar{f}} = N^{\gamma-1}\Big(\kappa_0\mathbb{I} + Q_{\mathrm{TL}}\big(\bar{\bar{f}}_\alpha\big)C^{(xx)}\Big)^{-1}y, \tag{149}$$

$$Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big) = \frac{g_w}{1-\frac{g_w}{N^\gamma}\bar{\bar{f}}^\intercal C^{(xx)}\bar{\bar{f}}}, \tag{150}$$

$$\langle\Delta\rangle_{\mathrm{TL}} = \kappa_0 N^{1-\gamma}\bar{\bar{f}} = \kappa_0\Big(\kappa_0\mathbb{I} + Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)}\Big)^{-1}y, \tag{151}$$

$$\langle f_*\rangle_{\mathrm{TL}} = Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)}_{*\alpha}\Big(\kappa_0\mathbb{I} + Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)}\Big)^{-1}_{\alpha\beta}y_\beta. \tag{152}$$

In one-loop approximation, we have for the train discrepancies

$$\bar{\bar{f}}_\delta = \big[A\big(\bar{\bar{f}}\big)\big]^{-1}_{\delta\epsilon}\Bigg[y_\epsilon + \frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S})}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial\bar{\bar{f}}_\epsilon}\Bigg], \tag{153}$$

$$A\big(\bar{\bar{f}}\big) = \kappa_0\mathbb{I} + Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)}, \tag{154}$$

$$\frac{\partial^2(-\mathcal{S})}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta}\bigg|_{j^*=0} = A\big(\bar{\bar{f}}\big) + \frac{2}{N}Q^2_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)}\bar{\bar{f}}\bar{\bar{f}}^\intercal C^{(xx)}, \tag{155}$$

$$\begin{aligned}\frac{\partial^3(-\mathcal{S})}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial\bar{\bar{f}}_\delta}\bigg|_{j^*=0} &= \frac{2}{N}Q^2_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,\Big(C^{(xx)}_{\alpha\beta}\Big[C^{(xx)}\bar{\bar{f}}\Big]_\delta + C^{(xx)}_{\alpha\delta}\Big[C^{(xx)}\bar{\bar{f}}\Big]_\beta + C^{(xx)}_{\beta\delta}\Big[C^{(xx)}\bar{\bar{f}}\Big]_\alpha\Big)\\ &\quad+\frac{8}{N^2}Q^3_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,\Big[C^{(xx)}\bar{\bar{f}}\Big]_\alpha\Big[C^{(xx)}\bar{\bar{f}}\Big]_\beta\Big[C^{(xx)}\bar{\bar{f}}\Big]_\delta,\end{aligned} \tag{156}$$

$$\langle\Delta_\alpha\rangle_{\text{1-Loop}} = \kappa_0 N^{1-\gamma}\bar{\bar{f}}_\alpha, \tag{157}$$

and for the test predictors

$$\langle f^*\rangle_{\text{1-Loop}} = Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)}_{*\alpha}\bar{\bar{f}}_\alpha - \frac{1}{2}\sum_{\alpha\beta}(-\mathcal{S}^{(2)})^{-1}_{\beta\alpha}\frac{\partial^3(-\mathcal{S})}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial j_*}\bigg|_{j^*=0}. \tag{158}$$

$$\begin{aligned}\frac{\partial^3(-\mathcal{S})}{\partial\bar{\bar{f}}_\alpha\partial\bar{\bar{f}}_\beta\partial j_*}\bigg|_{j^*=0} &= \frac{2}{N}Q^2_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,\Big(C^{(xx)}_{\alpha\beta}\Big[C^{(xx)}\bar{\bar{f}}\Big]_* + C^{(xx)}_{\alpha *}\Big[C^{(xx)}\bar{\bar{f}}\Big]_\beta + C^{(xx)}_{\beta *}\Big[C^{(xx)}\bar{\bar{f}}\Big]_\alpha\Big)\\ &\quad+\frac{8}{N^2}Q^3_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,\Big[C^{(xx)}\bar{\bar{f}}\Big]_\alpha\Big[C^{(xx)}\bar{\bar{f}}\Big]_\beta\Big[C^{(xx)}\bar{\bar{f}}\Big]_*,\end{aligned} \tag{159}$$

Finally, in the proportional limit $P \propto N \to \infty$ this reduces to

$$\bar{\bar{f}} = \Big(\kappa_0\mathbb{I} + Q_{\text{1-Loop}}\big(\bar{\bar{f}}\big)\,C^{(xx)}\Big)^{-1}y, \tag{160}$$

$$Q_{\text{1-Loop}}\big(\bar{\bar{f}}\big) = Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big) - \frac{1}{N}Q^2_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,\mathrm{Tr}\big[(\kappa_0\mathbb{I} + Q_{\mathrm{TL}}\big(\bar{\bar{f}}\big)\,C^{(xx)})^{-1}C^{(xx)}\big], \tag{161}$$

$$\langle\Delta\rangle = \kappa_0 N^{1-\gamma}\Big(\kappa_0\mathbb{I} + Q_{\text{1-Loop}}\big(\bar{\bar{f}}\big)\,C^{(xx)}\Big)^{-1}y, \tag{162}$$

$$\langle f^*\rangle_{\text{1-Loop, rescaling}} = \big[Q_{\text{1-Loop}}\big(\bar{\bar{f}}\big)\,C^{(xx)}_{*\alpha}\big]\,\big(\kappa_0\mathbb{I} + Q_{\text{1-Loop}}\big(\bar{\bar{f}}\big)\,C^{(xx)}\big)^{-1}_{\alpha\beta}y_\beta. \tag{163}$$

C.2. Numerical stability and computational complexity

To improve the numerical stability when solving these self-consistency equations, we use the following scheme: In standard scaling, we use the train predictors given by the NNGP as starting values for solving the tree-level equations. Then, we use the tree-level solution as the initial value for the one-loop equations since the latter contain additional fluctuation corrections to the tree-level result. Further, we anneal solutions from the standard to the mean-field regime since the solutions are more unstable in the mean-field regime (see pseudo code in 1).

The computational complexity for solving these equations is $\mathcal{O}(P^3)$ with a different pre-factor for tree-level and one-loop. Note that neither the input dimension D nor the network width N but only the number of training samples P affect the computational complexity.

Algorithm 1 Annealing of solutions across scaling regimes

Input: data X, labels Y, scales $\{\chi_i\}_i$
Compute NNGP train predictors f_α^{NNGP} from data X and labels Y.
Set initial value to NNGP predictor f_α^{NNGP}.
for χ **in** $\{\chi_i\}_i$ **do**
  Set $g_w \mapsto g_w/\chi$.
  Solve self-consistency solution for tree-level approximation $\bar{\bar{f}}_\alpha^{\text{TL}}$ with initial value $\bar{\bar{f}}_\alpha^{\text{NNGP}}$.
  Solve self-consistency solution for one-loop approximation $\bar{\bar{f}}_\alpha^{\text{1-Loop}}$ with initial value $\bar{\bar{f}}_\alpha^{\text{TL}}$.
end for

C.3. Network tasks and training

Ising task We use a linearly separable Ising task: Each pattern x_α in the Ising task is D-dimensional and $x_{\alpha i} \in \{\pm 1\}$. If the pattern belongs to class -1, each $x_{\alpha i}$ realizes $x_{\alpha i} = +1$ with a probability of $p_1 = 0.5 - \Delta p$ and the value $x_{\alpha i} = -1$ with $p_2 = 0.5 + \Delta p$. The value for each pattern element $x_{\alpha i}$ is drawn independently. If the pattern belongs to class $+1$, the probabilities for $x_{\alpha i} = 1$ and $x_{\alpha i} = -1$ are inverted. The task complexity decreases with larger Δp. We use $\Delta p = 0.1$ throughout, corresponding to an oracle accuracy on the classification task of $P_{\text{oracle}} = 99,78\%$.

Teacher-student task In this setting, the target is given by a $y_\alpha = w_* \cdot x_\alpha$, where $x_\alpha \in \mathbb{R}^D$ is standard normally distributed $x_{\alpha i} \sim \mathcal{N}(0, \mathbb{I})$. The teacher direction $w_* \in \mathbb{R}^D$ is chosen to be $\hat{e}_1$ in the standard basis.

Network training We train networks using Langevin stochastic gradient descent (LSGD) as detailed in (Naveh et al., 2021) so that the trained networks are effectively sampled from the posterior distribution (29). Here evolving network parameters Θ such as weights V, w with the stochastic differential equation

$$
\begin{aligned}
\partial_t \Theta(t) &= -\rho\Theta(t) - \nabla_\Theta \mathcal{L}(\Theta(t); y) + \sqrt{2T}\zeta(t), \\
\big\langle \zeta_i(t)\zeta_j(s) \big\rangle &= \delta_{ij}\delta(t-s),
\end{aligned} \tag{164}
$$

with the squared error loss $\mathcal{L}(\Theta; y) = \sum_{\alpha=1}^P (f_\alpha(\Theta) - y_\alpha)^2$, ζ a unit variance Gaussian white noise, and $f_\alpha(\Theta)$ denoting the network output for sample $\alpha = 1, \ldots, P$, leads to sampling from the equilibrium distribution for Θ for large times t which reads

$$
\lim_{t\to\infty} p\left(\Theta(t)\right) \sim \exp\left(-\frac{\rho}{2T}\|\Theta\|^2 - \frac{1}{T}\mathcal{L}(\Theta; y)\right). \tag{165}
$$

Using the Fokker-Planck equation (Risken, 1996) one can derive this density for Θ. Further, this implies a distribution on the network output

$$
\begin{aligned}
p(Y|X) &\propto \int d\Theta\, \exp\big(-\frac{\rho}{2T}\,\|\Theta\|^2 - \frac{1}{T}\,\|f-y\|^2\big) \\
&\propto \Big\langle \exp\big(-\frac{1}{T}\,\|f-y\|^2\big)\Big\rangle_{\Theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, T/\rho)} \\
&\propto \mathcal{N}(y|f, T/2)\,\langle \delta\big[f - f(\Theta)\big]\rangle_{\Theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, T/\rho)},
\end{aligned} \tag{166}
$$

In fact, $p(f|X) \equiv \langle \delta\big[f - f(\Theta)\big]\rangle_{\Theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, T/\rho)}$, leads to the posterior in (5) if one identifies $\kappa_0 = T/2$ with the regularization noise and $T/\rho = g/N$ with the variance of the parameter Θ_k. Implementing the sampling in practice this corresponds to requiring different weight decay ρ for each parameter, as weight variances can differ in the input and output layer.

The time discrete version of (164) is implemented in our PyTorch code as

$$
\begin{aligned}
\Theta_t &= \Theta_{t-1} - \eta\left(\rho\Theta_{t-1} + \nabla_\Theta \mathcal{L}(\Theta_{t-1}; y)\right) + \sqrt{2T\eta}\,\zeta_t, \\
\langle \zeta_t \zeta_s \rangle &= \delta_{ts},
\end{aligned} \tag{167}
$$

with standard normal ζ_t and finite time step η, which can also be interpreted as a learning rate. To accurately reflect the time evolution according to (164) the learning rate η needs to be small enough.

Hence the LSGD we implement corresponds to full-batch gradient descent with the addition of i.i.d. distributed standard normal noise and weight decay regularization (Krogh & Hertz, 1991). The value for κ_0 corresponds to a tradeoff in the optimization between the weight priors and the likelihood in terms of the loss $\mathcal{L}$. Choosing large κ_0 corresponds to large $T = 2\kappa_0$ and hence a large noise in the LSGD and therefore putting more emphasis on the Gaussian parameter priors. Small regularization values κ_0 favor the training data in terms of the loss in the exponent of (165).

To faithfully compare the numerical results with our theoretical results, the LSGD needs to sample from the equilibrium distribution. For this it needs to be ensured that the distribution is equilibrated by evolving the networks for 10.000 steps . We ensure uncorrelated network samples by initializing different networks with different random seeds.

For the Ising task, we average over $N_{\text{networks}} = 100$ with different initial weights to obtain the training and test predictors. For the teacher-student task, we average over $N_{\text{networks}} = 5.000$ with different initial weights to obtain the covariance of the network output projected onto different directions.

D. Additional figures

D.1. MNIST

Since the presented approach does not make any assumption on the data, it is applicable to arbitrary data sets. We here show results for binary classification on MNIST across scaling regimes.

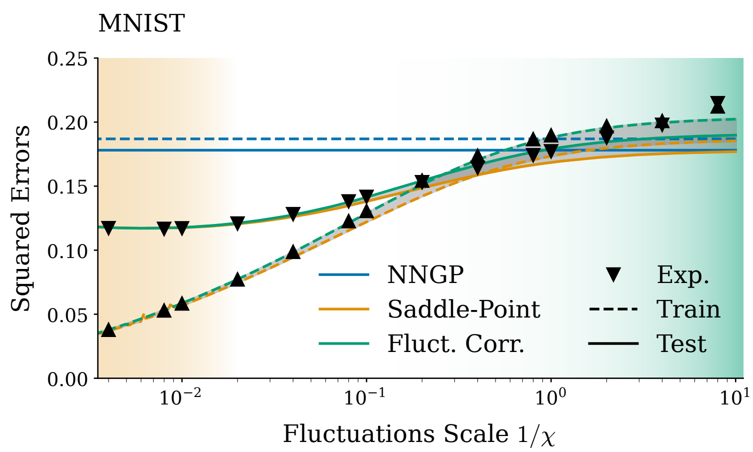

Figure 6. Binary classification on MNIST: training (solid line) and test errors (dashed line) across scaling regimes for different approaches. While standard scaling (green shaded area) requires a one-loop approximation with fluctuation corrections (Fluct. Corr.), a saddle-point or tree-level approximation (Saddle-Point) is sufficient in mean-field scaling (orange shaded area). Parameters: $P_{\text{train}} = 80$, $N = 100$, $D = 784$, $\kappa_0 = 1$, $P_{\text{test}} = 10^3$, $g_v = g_w = 2$.

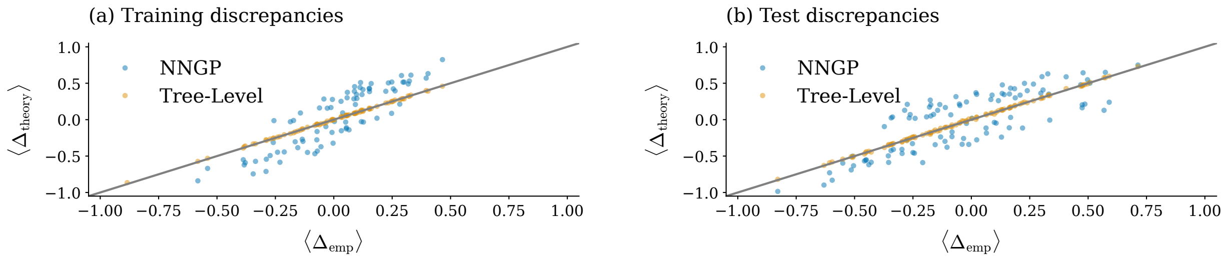

Figure 7. (a) Training discrepancies $\langle\Delta\rangle = y - \langle f_{\mathcal{D}}\rangle$ and (b) test discrepancies $\langle\Delta_*\rangle = y_* - \langle f_*\rangle$ for binary classification on MNIST in mean-field scaling. We show theoretical values for both NNGP and tree-level against empirical results, where the gray line marks the identity. In contrast to the NNGP, the tree-level approximation accurately matches the empirical values. Parameters: $\gamma = 2$, $P_{\text{train}} = 80$, $N = 100$, $D = 784$, $\kappa_0 = 1$, $P_{\text{test}} = 10^3$, $g_v = g_w = 2$.

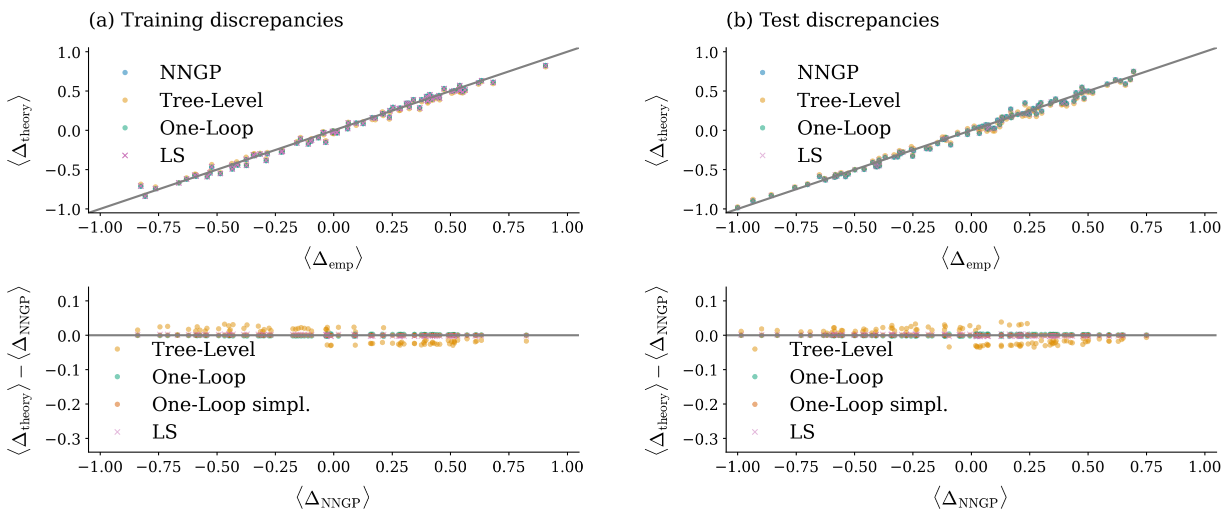

Figure 8. (a) Training discrepancies $\langle\Delta\rangle = y - \langle f_{\mathcal{D}}\rangle$ and (b) test discrepancies $\langle\Delta_*\rangle = y_* - \langle f_*\rangle$ for binary classification on MNIST in standard scaling. Upper row: theoretical values for different theories against empirical results; gray line marks the identity. Lower row: difference of theoretical values to the NNGP as a baseline against NNGP predictions, indicating small-scale differences between the different approaches. Results of the kernel approach by Li & Sompolinsky (2021) shown as reference (LS). Parameters: $\gamma = 1$, $P_{\text{train}} = 80$, $N = 100$, $D = 784$, $\kappa_0 = 1$, $P_{\text{test}} = 10^3$, $g_v = g_w = 2$.

D.2. Ising task in mean-field scaling

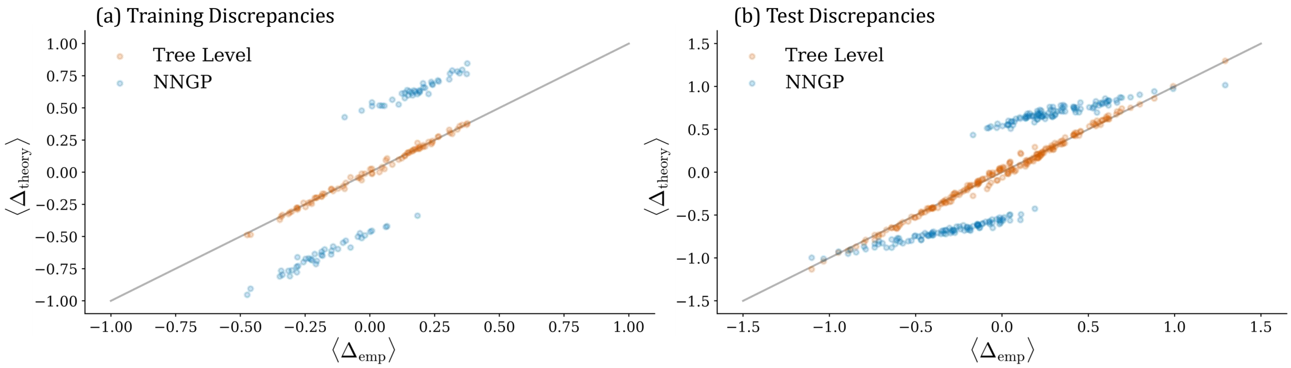

Figure 9. (a) Training discrepancies $\langle\Delta\rangle = y - \langle f_{\mathcal{D}}\rangle$ and (b) test discrepancies $\langle\Delta_*\rangle = y_* - \langle f_*\rangle$ on an Ising task in mean-field scaling. We show theoretical values for both NNGP and tree-level against empirical results, where the gray line marks the identity. In contrast to the NNGP, the tree-level approximation accurately matches the empirical values. We here use a non-linear activation function $\phi = \text{erf}$. Parameters: $\gamma = 2$, $P_{\text{train}} = 80$, $N = 100$, $D = 200$, $\kappa_0 = 1$, $P_{\text{test}} = 10^3$, $g_v = g_w = 0.5$, $\Delta p = 0.1$.

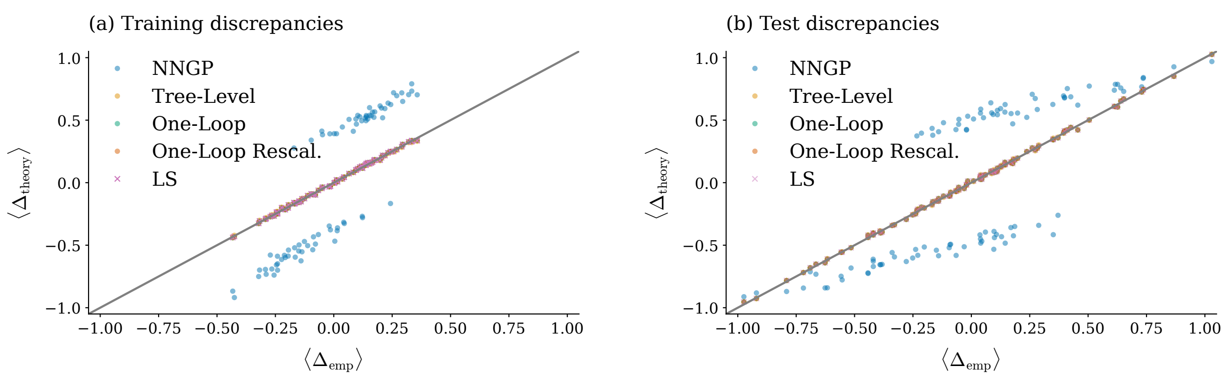

Figure 10. Scatter plots of (a) training discrepancies $\langle\Delta_\alpha\rangle = y_\alpha - \langle f_\alpha\rangle$ and (b) test discrepancies $\langle\Delta_*\rangle = y_* - \langle f_*\rangle$ on an Ising task in mean-field scaling. We show theoretical values for NNGP and different feature learning theories against empirical results, where the gray line marks the identity. In contrast to the NNGP, the tree-level approximation accurately matches the empirical values. Further, the different feature learning theories lie on top of one another in mean-field scaling. Parameters: $\gamma = 2$, $P_{\text{train}} = 80$, $N = 100$, $D = 200$, $\kappa_0 = 0.4$, $P_{\text{test}} = 10^3$, $g_v = 0.5$, $g_w = 0.2$, $\Delta p = 0.1$.

D.3. Coherent amplification of low-rank kernel structures

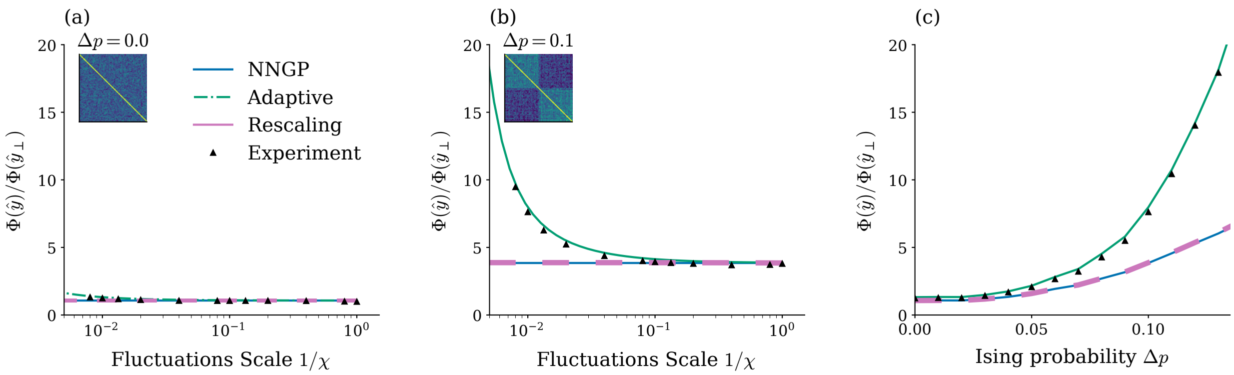

Figure 11. Relative directional feature learning on the Ising task as a function of the fluctuation scale $1/\chi$ for (a) an input kernel without structure and (b) an input kernel with block structure (input kernels shown as insets). Both NNGP and rescaling theory fail to capture directional feature learning, while the multi-scale adaptive theory accurately predicts network behavior. (c) An increase in structure in the input kernel increases with the Ising probability Δp and leads to a significantly higher directional feature learning in the adaptive theory than in both NNGP and rescaling, matching the experiments. Parameters: $P_{\text{train}} = 80$, $N = 100$, $D = 200$, $\kappa_0 = 2$, $g_v = 0.01$, $g_w = 0.5$.

D.4. Teacher-student task

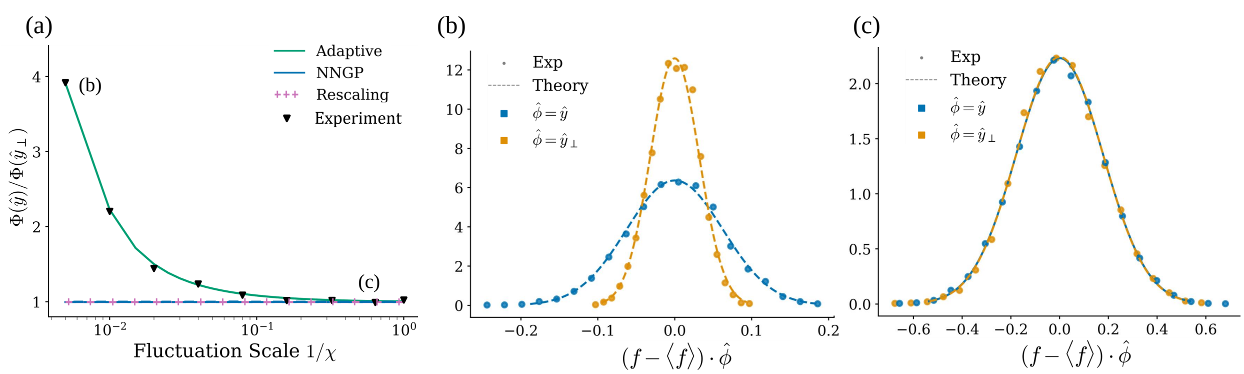

Figure 12. (a) Directional feature learning in a teacher-student setting as a function of the fluctuation scale $1/\chi$. Both NNGP and rescaling theory fail to capture directional feature learning, while the multi-scale adaptive theory accurately predicts network behavior. Output distribution in different directions (b) in mean field scaling ($\chi = N$) and (c) in standard scaling ($\chi = 128$) . Parameters: $P_{\text{train}} = 80$, $N = 200$, $D = 50$, $\kappa_0 = 2$, $g_v = 0.01$, $g_w = 2$.