Field theory for optimal signal propagation in residual networks

Kirsten Fischer,[1,2,*] David Dahmen,[1] and Moritz Helias[1,3]
[1]*Institute for Advanced Simulation (IAS-6), Jülich Research Centre, Jülich, Germany*
[2]*RWTH Aachen University, Aachen, Germany*
[3]*Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany*

(Received 27 August 2024; accepted 17 October 2025; published 1 December 2025)

Residual networks have significantly better trainability and thus performance than feed-forward networks at large depth. Introducing skip connections facilitates signal propagation to deeper layers. In addition, previous works found that adding a scaling parameter for the residual branch further improves generalization performance. While they empirically identified a particularly beneficial range of values for this scaling parameter, the mechanism for the resulting performance improvement and its universality across network hyperparameters remain an open question. For feed-forward networks, finite-size theories have led to important insights with regard to signal propagation and hyperparameter tuning. We here derive a systematic finite-size field theory for residual networks to study signal propagation and its dependence on the scaling for the residual branch. We derive analytical expressions for the response function, a measure for the network's sensitivity to inputs, and show that for deep networks the empirically found values for the scaling parameter lie within the range of maximal sensitivity. Furthermore, we obtain an analytical expression for the optimal scaling parameter that depends only weakly on other network hyperparameters, such as the weight variance, thereby explaining its universality across hyperparameters. Overall, this work provides a theoretical framework to study ResNets at finite size.

DOI: 10.1103/5lgz-4t7h

I. INTRODUCTION

While feed-forward neural networks (FFNets) have proven successful at learning a multitude of tasks [1,2], they become difficult to train at great depths [3]. As a result, very deep FFNets yield worse performance than their shallow counterparts. This empirical result is counterintuitive: Since feed-forward layers in principle can implement identity mappings, which can be added to a shallow network to increase its depth while yielding the same performance, such a performance degradation should not be present. This finding implies that identity mappings are difficult to learn; in consequence [3,4] introduced residual networks (ResNets) that contain skip connections between adjacent layers which implement identity mappings. Networks such as ResNet-50 [3] or ResNet-1001 [4] yield state-of-the-art performance on common benchmark data sets such as CIFAR-10 [5].

A scaling of the residual branch, i.e., of the nonidentity mapping in each residual layer, was first introduced by Szegedy *et al.* [6] who found that for networks with large numbers of convolutional filters training becomes unstable and leads to inactive neurons. While this effect could not be mitigated by additional batch normalization [7], downscaling the residual branch by a value ρ between 0.1 and 0.3 proved to be a reliable solution. Finding the optimal residual scaling and a mechanistic explanation for its effectiveness remains an open question.

One line of research studies how such an optimal residual scaling depends on the network depth. In the limit of infinite width, ResNets behave as a Gaussian process. The covariance of this process is also termed "Neural Network Gaussian Process (NNGP) kernel" and has been derived in [8]. While the NNGP corresponds to training only the readout layer, training all network layers in the limit of vanishing learning rates leads to a different Gaussian process, the "Neural Tangent Kernel" (NTK) [9]. In this kernel perspective, Huang *et al.* [8] argue that the NTK in the double limit of infinite width and depth becomes degenerate for FFNets but not for ResNets, suggesting a polynomial scaling of the residual branch with the inverse depth for better kernel stability at great depth. Bachlechner *et al.* [10] include the residual scaling as a trainable parameter instead of it being a hyperparameter and find that networks with the residual scaling being initialized at zero learn an inverse depth scaling.

According to Tirer *et al.* [11], smaller residual scalings lead to a smoother NTK and thereby to better interpolation properties between data points. Studying the spectral properties of the NTK, Barzilai *et al.* [12] find a bias of convolutional ResNets towards learning functions with low frequency or localized over few pixels. Further, they show that the scaling proposed by Huang *et al.* [8] leads to a less expressive dot-product kernel for convolutional ResNets, therefore

*Contact author: ki.fischer@fz-juelich.de

arguing for a depth-independent constant residual scaling. By performing a grid search, Zhang *et al.* [13] find a value near 0.1 to yield best generalization performance for deep ResNets.

References [14–17] argue for a residual scaling by the square root of the inverse depth: Arpit *et al.* [14] use a mean field analysis to argue that this scaling avoids exploding or vanishing gradients in the forward and backward pass. While Refs. [15,16] show that the resulting NTK is universal and can express any function, Zhang *et al.* [17] find that this scaling stabilizes forward and backward propagation of the signal in terms of its norm. On the practical side, Bordelon *et al.* [18] show that scaling the residual branch by the square root of the inverse depth allows them to extend μP scaling [19], a particular form of initialization that promotes trainability, to residual architectures.

We here tackle the problem of optimal scaling from a signal propagation perspective. We derive a field-theoretic description of residual networks to study their response function. This function describes the networks' sensitivity to varying inputs. As the network needs to be able to distinguish between different data samples, the overall range of output responses is a relevant indicator for both trainability and generalization. While a stronger signal generally ensures that two data samples can be better distinguished, this effect may be counteracted by saturation effects of the nonlinearity in the residual branch of the network. The residual scaling parameter determines how strongly differences across data samples are amplified and propagated through the network.

Our main contributions are as follows:

(1) We derive a field-theoretic description of the Bayesian network prior for residual networks that allows one to systematically account for finite-size properties of networks.

(2) We obtain the response function of residual networks as a finite-size effect that describes the networks' sensitivity to varying inputs.

(3) We show that the response function of the network output as a function of the residual scaling parameter has a distinct maximum and that the corresponding optimal residual scaling lies precisely within the value range empirically found by Szegedy *et al.* [6].

(4) We derive the dependence of the optimal residual scaling on network hyperparameters and find a strong dependence on the network depth and weak dependence on all other hyperparameters, explaining the universality of a $1/\sqrt{\text{depth}}$ scaling for deep residual networks.

The field-theoretic framework for the Bayesian network prior can be generalized beyond ResNets and generally used to systematically take into account finite-size properties of neural networks. Field-theoretic descriptions of feed-forward and recurrent networks have been derived in [20–23]. Building on a field-theoretic formulation of the Bayesian network prior, [24–30] study feature learning in finite-size networks and Lindner *et al.* [31] study the effect of data variability. Bordelon and Pehlevan [32] derive a dynamical field theory to study gradient dynamics in feed-forward networks in different scaling regimes. The presented theoretical framework thus has applications beyond understanding residual scaling, the main goal of the current work.

The main part is structured into two parts: In Sec. II we first derive a field-theoretic formulation of residual networks. In this field-theoretic framework, we recover the NNGP as a saddle point at infinite width and obtain the response function of residual networks as a finite-width correction to this saddle point. In Sec. III we then study the behavior of the response function as a function of the residual scaling and find a unique maximum of the response close to the empirically found optimal values. Finally, we relate this scaling to optimal signal propagation that is bounded by saturation effects of the nonlinearity and study its dependence on hyperparameters of the network. The code for theory and experiments can be found in [33].

II. FIELD THEORY OF RESIDUAL NETWORKS

We here study the following residual architecture:

$$\begin{aligned} h^{(0)} &= W^{\text{in}} x + b^{\text{in}}, \\ h^{(l)} &= h^{(l-1)} + \rho[W^{(l)} \phi(h^{(l-1)}) + b^{(l)}] \quad l = 1, \dots, L, \\ y &= W^{\text{out}} \phi(h^{(L)}) + b^{\text{out}}, \end{aligned} \tag{1}$$

yielding a mapping from the input $x \in \mathbb{R}^{d_{\text{in}}}$ to the output $y \in \mathbb{R}^{d_{\text{out}}}$ as $x \mapsto f(x;\theta) = y$ with trainable network parameters $\theta = \{W^{\text{in}}, b^{\text{in}}, W^{(l)}, b^{(l)}, W^{\text{out}}, b^{\text{out}}\}$. Similar to state-of-the-art models such as ResNet-50 [3], the model contains a linear readin and a fully connected readout layer. Thereby the input $x \in \mathbb{R}^{d_{\text{in}}}$, the signal $h^{(l)} \in \mathbb{R}^N$, and the output $y \in \mathbb{R}^{d_{\text{out}}}$ can have different dimensions. We refer to the residual branch

$$\mathcal{F}(h^{(l-1)}) \coloneqq \rho[W^{(l)} \phi(h^{(l-1)}) + b^{(l)}] \tag{2}$$

together with the skip connection $h^{(l-1)}$ in (1) as a network layer with index l [see Fig. 1(a)]. The total number of layers is given by L. We assume the nonlinear activation function ϕ to be saturating and twice differentiable almost everywhere; two common choices satisfying both conditions are the logistic function and the error function. In the following, we use $\phi = \text{erf}$. The residual branch is multiplied by a scaling factor ρ, which is referred to as the residual scaling parameter in the following. We study networks at initialization, which is equivalent to determining the network prior in a setting of Bayesian inference. To this end, we assume that the network parameters are Gaussian distributed: For the input layer $W_{ij}^{\text{in}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{w,\text{in}}^2/d_{\text{in}})$, $b_i^{\text{in}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{b,\text{in}}^2)$, for residual layers $W_{ij}^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2/N)$, $b_i^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_b^2)$, and for the readout layer $W_{ij}^{\text{out}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{w,\text{out}}^2/N)$, $b_i^{\text{out}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{b,\text{out}}^2)$. The tilde symbol is here and in the following employed as "is distributed as."

A. Network prior in field-theoretic framework

We derive analytic expressions of the network prior for the residual network architecture defined in (1) in terms of the network kernels. This derivation uses the field-theoretic approach employed in Segadlo *et al.* [22] to study deep feed-forward and recurrent networks and extends it to residual networks. Given a set of inputs $X = (x_\alpha)_{\alpha=1,\dots,P}$, the network prior defined by

$$p(Y|X) \coloneqq \int \mathrm{d}\theta \; p(Y|X,\theta)\, p(\theta) \tag{3}$$

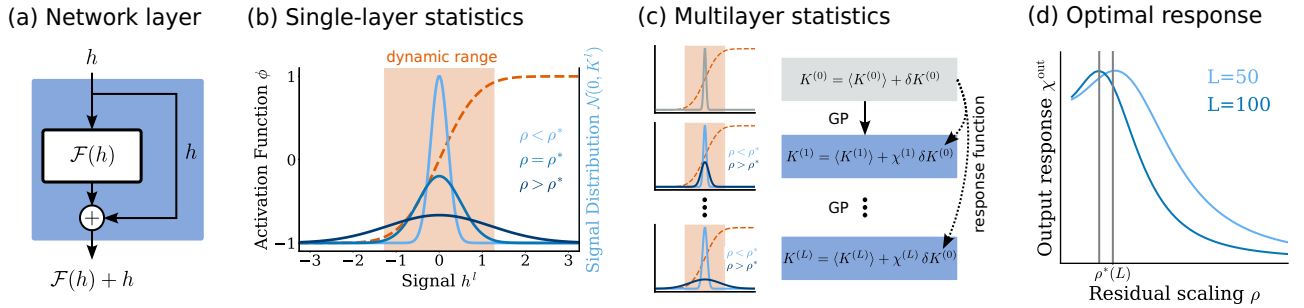

FIG. 1. Signal distribution in residual network. (a) Network layer with residual branch and skip connection. The residual branch returns $h \mapsto \mathcal{F}(h)$, the layer passes on $\mathcal{F}(h) + h$ to the next layer. (b) Distribution of the signal $h^{(l)}$ after layer l (solid curves) relative to the dynamic range $\mathcal{V}$ (shaded orange area) of the activation function $\phi = \mathrm{erf}$ (dashed curve). The signal is Gaussian distributed $h^{(l)} \sim \mathcal{N}(0, K^{(l)})$ with variance given by $K^{(l)}$, which depends on the residual scaling parameter ρ. For values larger than the optimal scaling $\rho > \rho^*$, part of the signal is lost in the saturation of the activation function ϕ (dark blue). For values smaller than the optimal scaling $\rho < \rho^*$, the signal is restricted to a small fraction of the dynamic range (light blue) in which the activation function is typically linear. For optimal scaling $\rho = \rho^*$, the signal optimally utilizes the whole dynamic range $\mathcal{V}$ of the activation function ϕ (blue). (c) The response function $\chi^{(l)}$ describes how the variance $K^{(l)}$, corresponding to the diagonal element of the GP kernel, changes to linear order in the perturbation of the input kernel $\delta K^{(0)}$ around its data mean $\langle K^{(0)} \rangle$. The kernel $K^{(l)}$ of the signal distribution can only increase across layers due to the skip connections; its rate of increase is governed by the residual scaling parameter ρ. If the signal goes into saturation ($\rho > \rho^*$) or remains close to zero ($\rho < \rho^*$), then the overall response of the network output to a change of the input kernel is limited. (d) The output response χ^{out} as a function of the residual scaling ρ exhibits a unique maximum that depends on the network depth L, yielding a scaling $\rho^*(L)$ that promotes optimal signal propagation in the network.

describes the joint distribution of all outputs $Y = (y_\alpha)_{\alpha=1,\dots,P}$ averaged over all possible initializations of the network parameters θ; each output y_α corresponds to one input x_α. Calculating the network prior jointly for all inputs x_α is analogous to the replica calculation in physics [34,35]: for each input x_α one considers a copy of the network with the same network parameters θ shared across all of these replica. Here we showcase the derivation of $p(x_\alpha|y_\beta)$ for a single input x_α, dropping the sample index α for simplicity. The general case of P inputs follows the same arguments and is given in Appendix A.

The network prior is given by the probability of an output y given an input x marginalized over the distribution of network parameters

$$p(y|x) = \int \mathrm{d}\theta \; p(y|x,\theta)\, p(\theta). \tag{4}$$

Given fixed network parameters θ, the probability $p(y|x,\theta)$ is given by enforcing the network model with Dirac δ distributions as

$$\begin{aligned} p(y|x,\theta) = \int \mathrm{d}h^{(0)} \cdots \int \mathrm{d}h^{(L)}\, \delta(h^{(0)} - W^{\mathrm{in}} x - b^{\mathrm{in}}) \\ \times \prod_{l=1}^{L} \delta(h^{(l)} - h^{(l-1)} - \rho W^{(l)} \phi(h^{(l-1)}) - \rho b^{(l)}) \\ \times \delta(y - W^{\mathrm{out}} \phi(h^{(L)}) - b^{\mathrm{out}}). \end{aligned} \tag{5}$$

1. *Marginalization over network parameters*

We marginalize over the network parameters as

$$\begin{aligned} p(y|x) = \int \mathrm{d}h^{(0)} \cdots \int \mathrm{d}h^{(L)}\, \langle \delta(h^{(0)} - W^{\mathrm{in}} x - b^{\mathrm{in}}) \rangle_{\{W^{\mathrm{in}},\, b^{\mathrm{in}}\}} \prod_{l=1}^{L} \langle \delta(h^{(l)} - h^{(l-1)} - \rho W^{(l)} \phi^{(l-1)} - \rho b^{(l)}) \rangle_{\{W^{(l)},\, b^{(l)}\}} \\ \times \langle \delta(y - W^{\mathrm{out}} \phi^{(L)} - b^{\mathrm{out}}) \rangle_{\{W^{\mathrm{out}},\, b^{\mathrm{out}}\}}, \end{aligned} \tag{6}$$

where $\langle \dots \rangle_{\{W,b\}}$ refers to the expectation value over the statistics of weights W and biases b and we use the shorthand $\phi^{(l)} = \phi(h^{(l)})$. We rewrite the Dirac δ distributions using its Fourier representation

$$\delta(z) = \prod_{k=1}^{N} \left\{ \int_{-\infty}^{\infty} \frac{d\omega_k}{2\pi} \right\} e^{i \sum_{k=1}^{N} \omega_k z_k} \tag{7}$$

$$= \int \mathrm{d}\tilde{z}\, \exp(\tilde{z}^{\mathsf{T}} z) \tag{8}$$

with scalar product $\tilde{z}^{\mathsf{T}} z = \sum_{i=1}^{N} \tilde{z}_i z_i$, where we integrate along the imaginary axis $\int \mathrm{d}\,\tilde{z} = \prod_k \int_{i\mathbb{R}} \frac{\mathrm{d}\,\tilde{z}_k}{2\pi i}$. Here $\tilde{z}$ is referred to as the conjugate variable to z. Writing the Fourier transform in complex variables is equivalent and commonly used in related works

[22,27,32]. This yields

$$p(y|x) = \int \mathrm{d}\,\tilde{y} \int \mathcal{D}\tilde{h} \int \mathcal{D}h \, \langle \exp((\tilde{h}^{(0)})^\mathsf{T}(h^{(0)} - W^{\mathrm{in}}x - b^{\mathrm{in}}))\rangle_{\{W^{\mathrm{in}},\, b^{\mathrm{in}}\}} \times \prod_{l=1}^{L} \langle \exp((\tilde{h}^{(l)})^\mathsf{T}(h^{(l)} - h^{(l-1)} - \rho W^{(l)}\phi^{(l-1)} - \rho b^{(l)}))\rangle_{\{W^{(l)},\, b^{(l)}\}} \times \langle \exp(\tilde{y}^\mathsf{T}(y - W^{\mathrm{out}}\phi^{(L)} - b^{\mathrm{out}}))\rangle_{\{W^{\mathrm{out}},\, b^{\mathrm{out}}\}}, \tag{9}$$

where $\int \mathcal{D}h = \prod_{l=0}^{L} \int \mathrm{d}h^{(l)}$ and $\int \mathcal{D}\tilde{h} = \prod_{l=0}^{L} \int \mathrm{d}\,\tilde{h}^{(l)}$ for brevity.

Since the network parameters θ_k are independently distributed, the integrals decouple and only integrals of the form $\int \mathrm{d}\theta_k \, p(\theta_k) \exp(z\theta_k)$ appear, which can be solved exactly for $\theta_k \sim \mathcal{N}(0, \sigma^2)$ yielding $\exp(\frac{1}{2}\sigma^2 z^2)$. As an example, we have

$$\left\langle \exp\left(-\sum_{i,j} W_{ij}^{(l)} \, \rho \tilde{h}_i^{(l)} \phi_j^{(l-1)} \right) \right\rangle_{W^{(l)}} = \exp\left(\frac{1}{2} \frac{\sigma_w^2}{N} \sum_{i,j} \left[\rho \tilde{h}_i^{(l)} \phi_j^{(l-1)} \right]^2 \right). \tag{10}$$

Rewriting the resulting terms as $\sum_{mn} [\tilde{h}_m \phi_n^{(l-1)}]^2 = \tilde{h}^\mathsf{T}\tilde{h}\,[\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)}$, we obtain the action $\mathcal{S}$ of the network prior

$$p(y|x) = \int \mathrm{d}\,\tilde{y} \int \mathcal{D}\tilde{h} \int \mathcal{D}h \, \exp[\mathcal{S}(y, \tilde{y}, h, \tilde{h}|x)], \tag{11}$$

where we distinguish between the contributions of the readin layer, the hidden layers of the network with residual connectivity, and the readout layer

$$\mathcal{S}(y, \tilde{y}, h, \tilde{h}|x) = \mathcal{S}_{\mathrm{in}}(h^{(0)}, \tilde{h}^{(0)}|x) + \mathcal{S}_{\mathrm{net}}(h, \tilde{h}) + \mathcal{S}_{\mathrm{out}}(y, \tilde{y}|h^{(L)}). \tag{12}$$

These contributions are given by

$$\mathcal{S}_{\mathrm{in}}(h^{(0)}, \tilde{h}^{(0)}|x) := [\tilde{h}^{(0)}]^\mathsf{T} h^{(0)} + \frac{1}{2} \frac{\sigma_{w,\,\mathrm{in}}^2}{d_{\mathrm{in}}} [\tilde{h}^{(0)}]^\mathsf{T}\tilde{h}^{(0)} \, x^\mathsf{T}x + \frac{1}{2}\sigma_{b,\,\mathrm{in}}^2 [\tilde{h}^{(0)}]^\mathsf{T}\tilde{h}^{(0)}, \tag{13}$$

$$\mathcal{S}_{\mathrm{net}}(h, \tilde{h}) := \sum_{l=1}^{L} [\tilde{h}^{(l)}]^\mathsf{T}[h^{(l)} - h^{(l-1)}] + \frac{1}{2}\rho^2 \frac{\sigma_w^2}{N} [\tilde{h}^{(l)}]^\mathsf{T}\tilde{h}^{(l)} \, [\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)} + \frac{1}{2}\rho^2\sigma_b^2 [\tilde{h}^{(l)}]^\mathsf{T}\tilde{h}^{(l)}, \tag{14}$$

$$\mathcal{S}_{\mathrm{out}}(y, \tilde{y}|h^{(L)}) := \tilde{y}^\mathsf{T} y + \frac{1}{2} \frac{\sigma_{w,\,\mathrm{out}}^2}{N} \tilde{y}^\mathsf{T}\tilde{y} \, [\phi^{(L)}]^\mathsf{T}\phi^{(L)} + \frac{1}{2}\sigma_{b,\,\mathrm{out}}^2 \tilde{y}^\mathsf{T}\tilde{y}. \tag{15}$$

In contrast to feed-forward networks [22], the conjugate variable $\tilde{h}^{(l)}$ of layer l does not only couple to the signal $h^{(l)}$ of layer l, but also to the signal $h^{(l-1)}$ of the previous layer $l-1$. This coupling across layers results from the skip connections in residual networks. The interdependence between layers induced by the coupling prohibits the marginalization over the intermediate signals $h^{(l)}$ in a direct manner as in feed-forward networks.

2. Auxiliary variables

Quadratic terms in h and $\tilde{h}$ can be solved as Gaussian integrals. However, in (13)–(15) terms proportional to $\propto [\tilde{h}^{(l)}]^\mathsf{T}\tilde{h}^{(l)} \, [\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)}$ appear, which are at least quartic in h and $\tilde{h}$ for linear $\phi(h) = h$ and of even higher order for nonlinear ϕ. To treat these terms, we introduce auxiliary variables:

$$C^{(0)} := \frac{\sigma_{w,\,\mathrm{in}}^2}{d_{\mathrm{in}}} x^\mathsf{T}x + \sigma_{b,\,\mathrm{in}}^2, \tag{16}$$

$$C^{(l)} := \rho^2 \frac{\sigma_w^2}{N} [\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)} + \rho^2\sigma_b^2 \quad l = 1, \ldots, L, \tag{17}$$

$$C^{(L+1)} := \frac{\sigma_{w,\,\mathrm{out}}^2}{N} [\phi^{(L)}]^\mathsf{T}\phi^{(L)} + \sigma_{b,\,\mathrm{out}}^2. \tag{18}$$

For wide networks $N \gg 1$, we expect the average $\frac{1}{N}\sum_{i=1}^{N} [\phi_i^{(l-1)}]^2$ to concentrate around its mean value. Based on this intuition, we aim to rewrite the network prior $p(y|x)$ in terms of these scalar variables. Note that, while the auxiliary variables are scalar for a single input x, they become kernel matrices $C_{\alpha\beta}^{(l)}$ for multiple inputs with sample indices α, β (see Appendix A for details).

We enforce the definitions of the auxiliary variables with Dirac δ distributions as in (5), e.g.,

$$\delta\left(-NC^{(l)} + \rho^2\sigma_w^2 [\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)} + N\rho^2\sigma_b^2\right) = \int_{i\mathbb{R}} \frac{\mathrm{d}\,\tilde{C}^{(l)}}{2\pi i} \exp\left(-N\tilde{C}^{(l)}C^{(l)} + \rho^2\sigma_w^2 \, \tilde{C}^{(l)} [\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)} + N\rho^2\sigma_b^2 \, \tilde{C}^{(l)}\right) = \int_{i\mathbb{R}} \frac{\mathrm{d}\,\tilde{C}^{(l)}}{2\pi i} \exp\left(-N\tilde{C}^{(l)}C^{(l)} + N\rho^2\sigma_w^2 \, \tilde{C}^{(l)} \left[\phi_i^{(l-1)}\right]^2 + N\rho^2\sigma_b^2 \, \tilde{C}^{(l)}\right). \tag{19}$$

In the last line, we used that the scalar variables C and and its conjugate variables $\tilde{C}$ only couple to sums of $\tilde{h}$ and $\phi(h)$ over all neuron indices, so that all components of h, $\tilde{h}$, and thus also $\phi(h)$ are identically distributed. Thus, we can rewrite the expression in scalar variables h and $\tilde{h}$, pulling out a factor N in all terms. For the input layer, we denote the ratio $d_{\text{in}}/N = \nu_0$ and set $\nu_l = 1$ for $l > 0$; different network widths N_l across layers l can be considered by setting $\nu_l = N_l/N$. Moving all integrals over scalar variables h and $\tilde{h}$ to the exponent, we can write the network prior in terms of the kernels as

$$p(y|x) = \int \mathrm{d}\tilde{y} \left\langle \exp\left(\tilde{y}^{\mathsf{T}} y + \frac{1}{2} C^{(L+1)} \tilde{y}^{\mathsf{T}} \tilde{y}\right)\right\rangle_{C,\tilde{C}}, \tag{20}$$

where the statistics of $C, \tilde{C}$ are given by $p(C, \tilde{C}) \propto \exp\left[\mathcal{S}(C, \tilde{C})\right]$ with

$$\mathcal{S}(C, \tilde{C}) \coloneqq -N \sum_{l=1}^{L+1} \nu_l \, C^{(l)} \, \tilde{C}^{(l)} + N \mathcal{W}(\tilde{C}|C), \tag{21}$$

$$\begin{aligned} \mathcal{W}(\tilde{C}|C) \coloneqq \ln \prod_{l=1}^{L} \int \mathrm{d}h^{(l)} \int \mathrm{d}\tilde{h}^{(l)} \, \exp\left(\tilde{h}^{(l)}[h^{(l)} - h^{(l-1)}] + \frac{1}{2} C^{(l)} [\tilde{h}^{(l)}]^2\right) \exp\left(\tilde{C}^{(l)} \left[\rho^2 \sigma_w^2 \, \phi^{(l-1)} \phi^{(l-1)} + \rho^2 \sigma_b^2\right]\right) \\ \times \exp\left(\tilde{C}^{(L+1)} \left[\sigma_{w,\,\text{out}}^2 \, \phi^{(L)} \phi^{(L)} + \sigma_{b,\,\text{out}}^2\right]\right) \\ \times \int \mathrm{d}h^{(0)} \int \mathrm{d}\tilde{h}^{(0)} \exp\left(\tilde{h}^{(0)} h^{(0)} + \frac{1}{2} C^{(0)} [\tilde{h}^{(0)}]^2 + \tilde{C}^{(0)} \left[\frac{\sigma_{w,\,\text{in}}^2}{N} x^{\mathsf{T}} x + \nu_0 \, \sigma_{b,\,\text{in}}^2\right]\right). \end{aligned} \tag{22}$$

Note that the conjugate variables $\tilde{C}^{(l)}$ are not proper random variables, but will be integrated out later to obtain the statistics of the auxiliary variables $C^{(l)}$.

B. Saddle point approximation yields NNGP

We obtain the NNGP kernel for residual networks in the limit of infinite network width $N \to \infty$ as the saddle point of the action $\mathcal{S}$ describing the network prior. Since the action $\mathcal{S}$ scales with the network width N, we can perform such a saddle point approximation in the limit of infinite width $N \to \infty$ to evaluate integrals of the form

$$\int \mathcal{D}C \int \mathcal{D}\tilde{C} \, f(C, \tilde{C}) \, \exp[\mathcal{S}(C, \tilde{C})] \overset{N \to \infty}{=} f(C_*, \tilde{C}_*), \tag{23}$$

where C_* and $\tilde{C}_*$ are the saddle points of the action $\mathcal{S}$.

We compute the saddle points using the conditions

$$\frac{\partial \mathcal{S}}{\partial C} = 0, \quad \frac{\partial \mathcal{S}}{\partial \tilde{C}} = 0 \tag{24}$$

and get

$$C_*^{(l)} = \begin{cases} \frac{\sigma_{w,\,\text{in}}^2}{d_{\text{in}}} x^{\mathsf{T}} x + \sigma_{b,\,\text{in}}^2 & l = 0, \\ \rho^2 \sigma_w^2 \, \langle \phi^{(l-1)} \phi^{(l-1)} \rangle_p + \rho^2 \sigma_b^2 & 1 \leqslant l \leqslant L, \\ \sigma_{w,\,\text{out}}^2 \, \langle \phi^{(L)} \phi^{(L)} \rangle_p + \sigma_{b,\,\text{out}}^2 & l = L+1, \end{cases} \tag{25}$$

$$\tilde{C}_*^{(l)} = \begin{cases} \frac{1}{2} \langle [\tilde{h}^{(l)}]^2 \rangle_p = 0 & l = 0, \ldots, L, \\ \frac{1}{2} \langle \tilde{y}^2 \rangle_p = 0 & l = L+1, \end{cases} \tag{26}$$

where

$$\langle \cdots \rangle_p = \int \mathrm{d}h^{(0)} \int \mathrm{d}\tilde{h}^{(0)} \exp\left(\tilde{h}^{(0)} h^{(0)} + \frac{1}{2} C_*^{(0)} [\tilde{h}^{(0)}]^2\right) \prod_{l=1}^{L} \int \mathrm{d}h^{(l)} \int \mathrm{d}\tilde{h}^{(l)} \cdots \exp\left(\tilde{h}^{(l)}[h^{(l)} - h^{(l-1)}] + \frac{1}{2} C_*^{(l)} [\tilde{h}^{(l)}]^2\right). \tag{27}$$

For brevity, we also include the input kernel $C_*^{(0)} = C^{(0)}$ here, which is fixed by the data as $C^{(0)} = \sigma_{w,\,\text{in}}^2 / d_{\text{in}} \, x^{\mathsf{T}} x + \sigma_{b,\,\text{in}}^2$. The expectation value over the auxiliary variables $\tilde{h}$ and $\tilde{y}$ in (26) vanishes due to the normalization the probability distribution, which we explain in Appendix B in detail. In the following, we use the notation $\tilde{y} = \tilde{h}^{(L+1)}$ for brevity. The appearing expectation values are computed self-consistently with respect to $C_*^{(l)}$.

By defining the residual $f^{(l)} \coloneqq h^{(l)} - h^{(l-1)}$ for $1 \leqslant l \leqslant L$ and $f^{(0)} \coloneqq h^{(0)}$, we rewrite the appearing average as

$$\langle \cdots \rangle_p = \int \mathrm{d}f^{(0)} \int \mathrm{d}\tilde{f}^{(0)} \exp\left(\tilde{f}^{(0)} f^{(0)} + \frac{1}{2} C_*^{(0)} [\tilde{f}^{(0)}]^2\right) \prod_{l=1}^{L} \int \mathrm{d}f^{(l)} \int \mathrm{d}\tilde{h}^{(l)} \cdots \exp\left(\tilde{h}^{(l)} f^{(l)} + \frac{1}{2} C_*^{(l)} [\tilde{h}^{(l)}]^2\right). \tag{28}$$

We observe that the residuals $f^{(l)}$ for $0 \leqslant l \leqslant L$ are Gaussian distributed $f^{(l)} \sim \mathcal{N}(0, C_*^{(l)})$ with zero mean and covariance $C_*^{(l)}$. Since the expectation values in (25) have $h^{(l)}$ as an argument, we would like to rewrite $\langle \cdots \rangle_p$ with respect to $h^{(l)}$. Since $h^{(l)} = \sum_{k=0}^{l} f^{(k)}$ and the residuals $f^{(l)}$ are independent Gaussians, the signal $h^{(l)}$ is also Gaussian distributed with covariance $K^{(l)} = \sum_{k=0}^{l} C_*^{(k)}$.

Thus, we can rewrite (25) as

$$C_*^{(l)} = \rho^2 \sigma_w^2 \, \langle \phi^{(l-1)} \phi^{(l-1)} \rangle_{\mathcal{N}(0, K^{(l-1)})} + \rho^2 \sigma_b^2, \tag{29}$$

$$K^{(l)} = \begin{cases} \frac{\sigma_{w,\text{in}}^2}{d_{\text{in}}} x^\mathsf{T} x + \sigma_{b,\text{in}}^2 & l = 0, \\ \sum_{k=0}^{l} C_*^{(k)} & 1 \leqslant l \leqslant L, \\ \sigma_{w,\text{out}}^2 \, \langle \phi^{(L)} \phi^{(L)} \rangle_{\mathcal{N}(0, K^{(L)})} + \sigma_{b,\text{out}}^2 & l = L+1. \end{cases} \tag{30}$$

We recover the known NNGP result for the kernels as $K^{(l)} = K^{(l-1)} + C_*^{(l)}$ [8,11,12].

C. Next-to-leading-order correction yields response function

The strength of the here employed field-theoretic formalism is that finite-size corrections to the NNGP result in (30) can be systematically calculated by expanding the action $\mathcal{S}$ in (12) around its saddle point. In particular, for finite-size networks, the residual kernels $C^{(l)}$ do not exactly concentrate to the saddle point value but fluctuate around it. To lowest order, we describe these fluctuations as Gaussian. The Gaussian fluctuations correspond to physical quantities such as the response function: The response function of the network is defined as the linear response of the covariance measured in layer l to a perturbation of the covariance in layer m. To define the perturbation, we introduce the perturbed auxiliary variable (16) in the mth layer as $\bar{C}^{(m)}(\epsilon) := C^{(m)} + \epsilon$ for small ϵ, which obey a perturbed probability measure $(\bar{C}, \tilde{\bar{C}}) \sim p_m(\epsilon)$. Then the response function is given by

$$\Delta_{12}^{lm} := \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\langle \bar{C}^{(l)} \rangle|_{p_m(\epsilon)} - \langle \bar{C}^{(l)} \rangle|_{p_m(0)} \right). \tag{31}$$

Since the perturbation term ϵ appears in the constraint (19) in the form

$$\cdots \exp\left(N \tilde{C}^{(m)} \left[-\bar{C}^{(m)} + \epsilon + N \rho^2 \sigma_w^2 \left[\phi_i^{(l-1)} \right]^2 + N \rho^2 \sigma_b^2 \right] \right),$$

this definition is equivalent to the correlator

$$\Delta_{12}^{lm} = N \, \langle C^{(l)} \tilde{C}^{(m)} \rangle, \tag{32}$$

as one may regard ϵ as a source term for $N \tilde{C}^{(m)}$. The response function measures the network's sensitivity to different signals and thus is a measure of signal propagation. In turn, signal propagation in networks is a key indicator of network trainability and performance [36].

We obtain the Gaussian fluctuations of the kernels by computing the Hessian $S^{(2)}|_{(C_*, \tilde{C}_*)}$ of the action $\mathcal{S}$ at the saddle point

$$\mathcal{S}^{(2)}|_{(C_*, \tilde{C}_*)} = \begin{pmatrix} \frac{\partial^2}{\partial C^2} \mathcal{S} & \frac{\partial^2}{\partial C \, \partial \tilde{C}} \mathcal{S} \\ \frac{\partial^2}{\partial \tilde{C} \, \partial C} \mathcal{S} & \frac{\partial^2}{\partial \tilde{C}^2} \mathcal{S} \end{pmatrix} \Bigg|_{(C_*, \tilde{C}_*)} =: \begin{pmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} \\ \mathcal{S}_{21} & \mathcal{S}_{22} \end{pmatrix}. \tag{33}$$

Due to the evaluation at the saddle point, all following expectation values are with respect to the measure $\langle \cdots \rangle_p$ defined in (28). In the following, we drop $|_{(C_*, \tilde{C}_*)}$ for notational brevity. The diagonal blocks of the Hessian are given by

$$\frac{\partial^2}{\partial C^{(l)} \partial C^{(k)}} \mathcal{S} = \frac{1}{4} \langle \tilde{h}^{(l)} \tilde{h}^{(l)}, \tilde{h}^{(k)} \tilde{h}^{(k)} \rangle_p^c = 0, \tag{34}$$

$$\frac{\partial^2}{\partial \tilde{C}^{(l)} \partial \tilde{C}^{(k)}} \mathcal{S} = N \sigma_w^4 \, 1_{l>0} 1_{k>0} \, \langle \phi^{(l-1)} \phi^{(l-1)}, \phi^{(k-1)} \phi^{(k-1)} \rangle_p^c \times \begin{cases} \rho^4 & k, l \neq L+1, \\ \rho^2 & k \neq l = L+1 \vee l \neq k = L+1, \\ 1 & \text{else}, \end{cases} \tag{35}$$

where $1_{l>0}$ denotes the indicator function. We write $\langle \cdots \rangle^c$ for connected correlations defined as

$$\langle z^{(l)} z^{(l)}, z^{(k)} z^{(k)} \rangle_p^c = \langle z^{(l)} z^{(l)} z^{(k)} z^{(k)} \rangle_p - \langle z^{(l)} z^{(l)} \rangle_p \, \langle z^{(k)} z^{(k)} \rangle_p. \tag{36}$$

The average over the auxiliary variables $\tilde{h}$ in (34) vanishes due to the normalization of the probability distribution, which we explain in Appendix B in detail. For the off-diagonal terms, we have

$$\frac{\partial^2}{\partial C^{(l)} \partial \tilde{C}^{(k)}} \mathcal{S} = -N \nu_l \delta_{kl} + N \, 1_{k>0} \, \sigma_w^2 \langle [\phi'^{,(k-1)}]^2 + \phi''^{,(k-1)} \phi^{(k-1)} \rangle_{\mathcal{N}(0, K^{(k-1)})} \, 1_{k>l} \times \begin{cases} \rho^2 & k \leqslant L \\ 1 & k = L+1 \end{cases}, \tag{37}$$

where we used Price's theorem for a one-dimensional variable [37]

$$\partial_A \langle \phi^2(h) \rangle_{h \sim \mathcal{N}(0,A)} = \langle \partial_h^2 \phi^2(h) \rangle_{h \sim \mathcal{N}(0,A)} \tag{38}$$

to compute the derivative of the expectation value by the covariance (see Appendix D for details). The condition $k > l$ enforced by the indicator function $1_{k>l}$ results from a term $\partial_{C^{(l)}} K^{(k-1)}$ appearing in the derivative, because the network kernel $K^{(k-1)}$ only depends on the residual kernels $C^{(l)}$ with $l < k$.

We obtain the Gaussian fluctuations of the variables $C^{(l)}$ and $\tilde{C}^{(l)}$ by taking the negative inverse of the Hessian, also called the propagator $\mathcal{G}$ in field theory

$$\mathcal{G} := -(\mathcal{S}^{(2)})^{-1} =: \begin{pmatrix} \mathcal{G}_{11} & \mathcal{G}_{12} \\ \mathcal{G}_{21} & \mathcal{G}_{22} \end{pmatrix}. \tag{39}$$

By using the block structure and the fact that $\mathcal{S}_{11} = 0$, we have

$$\mathcal{G}_{11} = \mathcal{G}_{12}\,\mathcal{S}_{22}\,\mathcal{G}_{21}, \tag{40}$$

$$\mathcal{G}_{12} = -\mathcal{S}_{21}^{-1}, \tag{41}$$

$$\mathcal{G}_{22} = 0. \tag{42}$$

Since the off-diagonal block matrix $\mathcal{S}_{21}$ is a lower triangular matrix, its inverse can be computed using forward propagation

$$\begin{aligned} \mathcal{G}_{12}^{lm} = {} & N^{-1} v_l^{-1} \delta_{lm} + 1_{L \geqslant l > 0}\, \rho^2 \sigma_w^2 \langle [\phi'^{,(l-1)}]^2 + \phi''^{,(l-1)} \phi^{(l-1)} \rangle_{\mathcal{N}(0, K^{(l-1)})} \sum_{k=0}^{l-1} \mathcal{G}_{12}^{km} \\ & + \delta_{l,L+1}\, \sigma_w^2 \langle [\phi'^{,(l-1)}]^2 + \phi''^{,(l-1)} \phi^{(l-1)} \rangle_{\mathcal{N}(0, K^{(l-1)})} \sum_{k=0}^{l-1} \mathcal{G}_{12}^{km}, \end{aligned} \tag{43}$$

yielding the response function $\Delta_{12}^{lm} = N \langle C^{(l)} \tilde{C}^{(m)} \rangle = N \mathcal{G}_{12}^{lm}$ as a $O(N^{-1})$ correction to the NNGP result. Due to the residual architecture, any response can only propagate forward in the network, which is reflected in the term $1_{k>l}$ in (37). The kernel $K^{(l)}$ in layer l depends explicitly not only on the kernel of the previous layer as for FFNets but on all preceding layers. This property directly results from the skip connections in residual networks. Thereby, the response function in layer l contains information from all preceding layers, such that information can propagate to deeper network layers.

We are ultimately interested in the response with respect to the network input, hence we define

$$\eta^{(l)} := \Delta_{12}^{l\,0}. \tag{44}$$

For the response of the residual kernels $C^{(l)}$ in all intermediate layers $0 < l < L + 1$, we then have

$$\eta^{(l)} = \rho^2 \sigma_w^2 \left\langle [\phi'^{,(k-1)}]^2 + \phi''^{,(k-1)} \phi^{(k-1)} \right\rangle_{\mathcal{N}(0, K^{(l-1)})} \sum_{k=0}^{l-1} \eta^{(k)}. \tag{45}$$

In particular, it is $\eta^{(0)} = N/d_{\text{in}}$. Due to their additive nature, the response function $\chi^{(l)}$ of the kernels $K^{(l)}$ is given by

$$\chi^{(l)} := \langle K^{(l)}\, \tilde{C}^{(0)} \rangle \tag{46}$$

$$= \sum_{k=0}^{l} \eta^{(k)}. \tag{47}$$

Finally, the output response χ^{out} is given by

$$\chi^{\text{out}} := \langle K^{(L+1)}\, \tilde{C}^{(0)} \rangle \tag{48}$$

$$= \sigma_{w,\,\text{out}}^2 \left\langle [\phi^{(L)}]' [\phi^{(L)}]' + [\phi^{(L)}]'' [\phi^{(L)}] \right\rangle_{\mathcal{N}(0, K^{(L)})} \sum_{k=0}^{L} \eta^{(k)}. \tag{49}$$

In addition, one can also compute the fluctuation corrections from the Hessian via (40). For details, see Appendix C.

In Fig. 2 we compare the behavior of the residual kernels $C_*^{(l)}$ and the response function $\eta^{(l)}$ between FFNets and ResNets. While the residual kernels $C_*^{(l)}$ in FFNets decay to zero as a function of depth, they approach a value larger than zero in ResNets due to the accumulation of variance across layers. Similarly, while the response function in FFNets decays exponentially to zero, it decays much slower in ResNets and approaches zero only asymptotically (see Fig. 8 in Appendix F 1). The latter observation matches previous results by [38] based on the convergence rate of the kernels; however, we here derive the response function that explicitly measures the dependence on the input kernel.

D. Relation to linear response theory

In the previous section, we derived the response function of the network as the first-order approximation in $O(N^{-1})$ from a systematic field-theoretic calculation. The field-theoretic

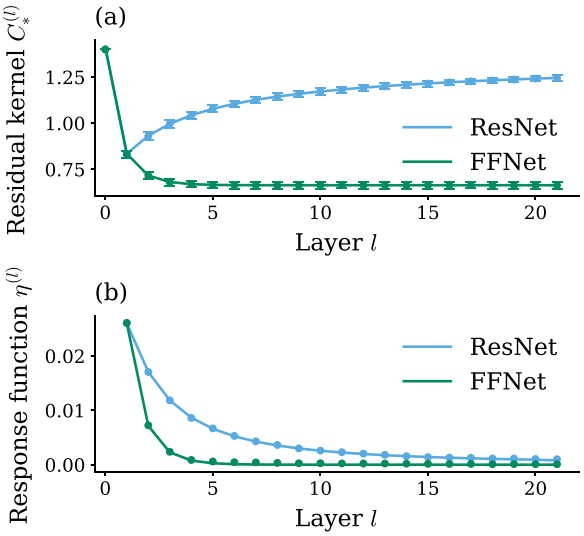

FIG. 2. Residual kernels $C_*^{(l)}$ (a) and the respective response function $\eta^{(l)}$ (b) in ResNets (blue) compared to FFNets (green). In (a) error bars indicate standard error of the mean obtained from simulation over 10^3 network initializations, solid curves show theory values from (29). In (b) dots represent simulations over 10^2 input samples and 10^3 network initializations, solid curves show theory values from (45). Errors are of order 10^{-5} and therefore not shown. ResNets exhibit a slower decay over layers l compared to FFNets. Other parameters: $\sigma_{w,\,\text{in}}^2 = \sigma_w^2 = \sigma_{w,\,\text{out}}^2 = 1.2$, $\sigma_{w,\text{in}}^2 = \sigma_w^2 = \sigma_{w,\text{out}}^2 = 1.2$, $\sigma_{b,\,\text{in}}^2 = \sigma_b^2 = \sigma_{b,\,\text{out}}^2 = 0.2$, $d_{\text{in}} = d_{\text{out}} = 100$, $N = 500$, $\rho = 1$, $\phi = \text{erf}$.

approach has the advantage of providing a framework in which we can systematically compute such correction terms to different orders; to provide more intuition on the response function we here discuss its relation to linear response theory.

Consider a change of the input kernel given by $K^{(0)} = \langle K^{(0)} \rangle + \delta K^{(0)}$. For small perturbations $\delta K^{(0)}$, we can ask how kernels in subsequent layers are affected by this perturbation. To linear order in $\delta K^{(0)}$, we expand the residual kernel as

$$C_*^{(l)} = C_*^{(l)}|_{\langle K^{(0)} \rangle} + \left.\frac{\partial C^{(l)}}{\partial K^{(0)}}\right|_{\langle K^{(0)} \rangle} \delta K^{(0)} + O(\delta^2). \tag{50}$$

The first-order Taylor term accounts for the effect of the perturbation and corresponds to the response function of the residual kernels $C_*^{(l)}$ as

$$\eta^{(l)} = \left.\frac{\partial C^{(l)}}{\partial K^{(0)}}\right|_{\langle K^{(0)} \rangle} \tag{51}$$

$$= \rho^2 \sigma_w^2 \frac{\partial}{\partial K^{(l-1)}} \langle \phi^{(l-1)} \phi^{(l-1)} \rangle_{\mathcal{N}(0, K^{(l-1)})} \left.\frac{\partial K^{(l-1)}}{\partial K^{(0)}}\right|_{\langle K^{(0)} \rangle} \tag{52}$$

$$= \rho^2 \sigma_w^2 \langle [\phi'^{,(l-1)}]^2 + \phi''^{,(l-1)} \phi^{(l-1)} \rangle_{\mathcal{N}(0, K^{(l-1)})} \sum_{k=0}^{l-1} \eta^{(k)}, \tag{53}$$

where we used Price's theorem [37] as in (38) from the second to the third line, and in the last line that

$$\left.\frac{\partial K^{(l-1)}}{\partial K^{(0)}}\right|_{\langle K^{(0)} \rangle} = \sum_{k=0}^{l-1} \left.\frac{\partial C^{(k)}}{\partial K^{(0)}}\right|_{\langle K^{(0)} \rangle} = \sum_{k=0}^{l-1} \eta^{(k)}. \tag{54}$$

The expectation value of the derivatives $\langle [\phi'^{,(l-1)}]^2 + \phi''^{,(l-1)} \phi^{(l-1)} \rangle_{\mathcal{N}(0, K^{(l-1)})}$ measures how the perturbation of the kernel $K^{(l-1)}$ affects the residual kernel $C_*^{(l)}$ in the subsequent layer l. It gets multiplied by the accumulated perturbations of all previous layers, as one expects intuitively due to the skip connections in residual networks. The expression for the linear response of the kernels $K^{(l)}$ follows directly from its definition

$$\chi^{(l)} = \left.\frac{\partial K^{(l)}}{\partial K^{(0)}}\right|_{\langle K^{(0)} \rangle} = \frac{\partial}{\partial K^{(0)}} \sum_{k=0}^{l} C_*^{(k)}|_{\langle K^{(0)} \rangle} = \sum_{k=0}^{l} \eta^{(l)}. \tag{55}$$

While expressions for the response function can be computed using linear response theory, the field-theoretic formalism in Sec. II C formally shows that the response function is the leading order $O(N^{-1})$ finite-size correction to the NNGP result.

III. SIGNAL PROPAGATION AND OPTIMAL SCALING IN RESIDUAL NETWORKS

Next we apply the field-theoretic framework derived above to study signal propagation in ResNets. The sensitivity of signal propagation to different inputs can be measured by the response function [36]. Good signal propagation is linked to improved trainability and thus higher generalization performance of trained networks [36,38]. We first focus on the behavior of the diagonal elements $K_{\alpha\alpha}^{(l)}$ of the network kernels on the same sample x_α to obtain an intuition for the signal propagation in residual networks and then discuss how their behavior extends to off-diagonal elements $K_{\alpha\beta}^{(l)}$ of the network kernels for different samples $x_\alpha \neq x_\beta$. We drop the sample index where only a single sample is considered and write explicitly where multiple samples are considered.

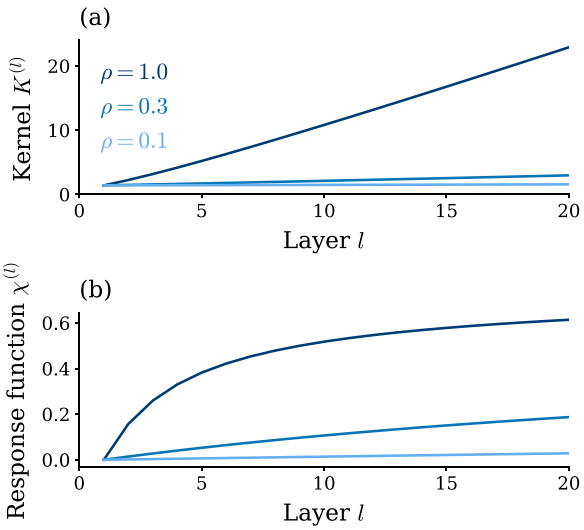

FIG. 3. Dependence of (a) kernels $K^{(l)}$ and (b) the respective response function $\chi^{(l)}$ on the residual scaling parameter ρ. The residual scaling takes values $\rho \in [1.0,\ 0.3,\ 0.1]$ (from dark to light). The residual scaling parameter ρ governs the rate of increase in both quantities. Other parameters: $\sigma_{w,\text{in}}^2 = \sigma_w^2 = \sigma_{w,\text{out}}^2 = 1.2$, $\sigma_{b,\text{in}}^2 = \sigma_b^2 = \sigma_{b,\text{out}}^2 = 0.2$, $d_{\text{in}} = d_{\text{out}} = 100$, $N = 500$, $\phi = \text{erf}$.

A. Optimal scaling of the residual branch

We start by studying the effect of the residual scaling parameter ρ on the kernels $K^{(l)}$ and response function $\chi^{(l)}$ of layer l as these quantities describe the distribution of the signal $h^{(l)}$. Since ρ^2 scales the residual kernels $C_*^{(l)}$ in (29), which are being summed to obtain $K^{(l)}$, the residual scaling governs the rate of increase of $K^{(l)}$ across layers [see Fig. 3(a)]. The response function $\chi^{(l)}$ exhibits the same scaling and thus behavior [see Fig. 3(b)].

The dependence of the residual kernels $C_*^{(l)}$ on the residual scaling ρ transfers to the response function; due to the output layer being a nonlinear feed-forward layer the output response shown in Figs. 4(a) and 4(b) exhibits a complex nonlinear behavior with a unique maximum ρ^*. The shape of the response function and thus the optimal value ρ^* depend on the network depth L, shifting to smaller values ρ^* with larger depth. However, we observe an antagonistic effect: the depth dependence becomes weaker for deeper networks. The optimal value ρ^* lies between $[0.1,\ 0.3]$, as found empirically in previous works [6,13].

Due to the recursive nature of the nonlinear Eqs. (45)–(48) for the response function, we cannot determine the optimal value ρ^* analytically. However, for the variance $K^{(l)}$ we can make an intuitive argument regarding the signal propagation in the network: For deeper networks, the kernels $K^{(l)}$ in (30) grow continuously, so that the signal $h^{(l)}$ leaves the dynamic range $\mathcal{V}$ of the activation function ϕ [see Fig. 1(b)]. In con-

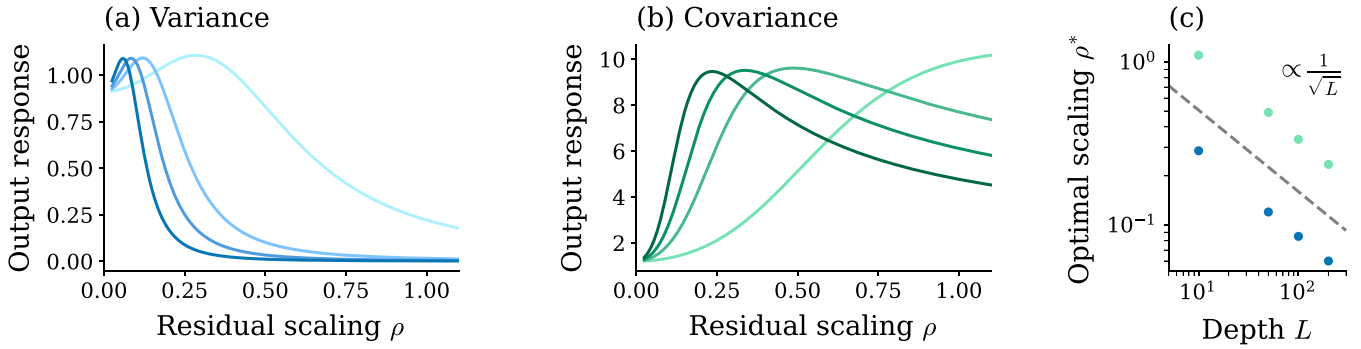

FIG. 4. Optimal scaling of the residual branch. Output response χ^{out} for (a) diagonal and (b) off-diagonal elements of the network kernel $K^{(l)}_{\alpha\beta}$. Different curves correspond to different network depths $L \in [10,\ 50,\ 100,\ 200]$ (light to dark). All curves exhibit a unique maximum; the residual scaling values ρ^* with largest response concentrate with increasing depth. (c) Optimal residual scaling $\rho^* = \text{argmax}(\chi^{\text{out}})$ for diagonal (blue) and off-diagonal (green) elements of the network kernel $K^{(l)}_{\alpha\beta}$. In both cases, these scale with $1/\sqrt{L}$ (gray). Other parameters: input kernel $K^{(0)} = \begin{pmatrix} 0.05 & 0.03 \\ 0.03 & 0.05 \end{pmatrix}$, $\sigma_w^2 = 1.25$, $\sigma_b^2 = 0.05$, $d_{\text{in}} = d_{\text{out}} = 100$, $N = 500$, $\phi = \text{erf}$.

sequence, part of the signal $h^{(L)}$ is lost in the readout layer, reducing the output response χ^{out} to changing inputs. The magnitude of the network kernels $K^{(l)}$ depends on ρ^2, so that a smaller residual scaling leads to a less rapid growth of the kernels $K^{(l)}$ and the signal $h^{(L)}$ stays in the dynamic range $\mathcal{V}$. For very small scalings ρ, the residual branch is suppressed and the network reduces to a single hidden-layer perceptron [see Fig. 1(c)].

Based on this intuition, we obtain an approximate expression for the optimal scaling ρ^*: We assume that the signal $h^{(l)}$ stays in the dynamic range $\mathcal{V}$ of the activation function so that $\phi(h^{(l)}) \approx \phi'(0)\, h^{(l)}$, where $\phi'(0)$ accounts for the slope of the activation function at its origin. The residual kernel then simplifies to

$$C_*^{(l)} = \rho^2 \sigma_w^2\, \phi'(0)^2 \sum_{k=0}^{l-1} C_*^{(k)} + \rho^2 \sigma_b^2, \tag{56}$$

and hence we get

$$C_*^{(l)} = C_*^{(l-1)} + \rho^2 \sigma_w^2\, \phi'(0)^2\, C_*^{(l-1)}. \tag{57}$$

Solving this recursion, we obtain

$$C_*^{(l)} = \left[1 + \rho^2 \sigma_w^2\, \phi'(0)^2\right]^{l-1} \left[\rho^2 \sigma_w^2\, \phi'(0)^2\, K^{(0)} + \rho^2 \sigma_b^2\right]. \tag{58}$$

Using the sum of the first $L+1$ terms of the geometric series and $C^{(0)} = K^{(0)}$ per definition yields

$$\begin{aligned} K^{(L)} &= \sum_{k=0}^{L} C^{(k)} \qquad (59) \\ &= \left[1 + \rho^2 \sigma_w^2\, \phi'(0)^2\right]^L K^{(0)} \\ &\quad + \frac{\sigma_b^2}{\phi'(0)^2 \sigma_w^2} \left\{\left[1 + \rho^2 \sigma_w^2\, \phi'(0)^2\right]^L - 1\right\}. \qquad (60) \end{aligned}$$

Assuming the 1σ range of the distribution to stay within the dynamic range $\mathcal{V}$ for a point-symmetric activation function ϕ, we set $\mathcal{V}/2 = \sqrt{K^{(L)}}$ and obtain for the optimal scaling parameter

$$\rho^* \approx \frac{1}{\sigma_w\, \phi'(0)} \sqrt{\left(\frac{\sigma_w^2\, \phi'(0)^2 \left(\frac{\mathcal{V}}{2}\right)^2 + \sigma_b^2}{\sigma_w^2\, \phi'(0)^2\, K^{(0)} + \sigma_b^2}\right)^{\frac{1}{L}} - 1}. \tag{61}$$

The assumption $\mathcal{V}/2 = \sqrt{K^{(L)}}$ is only an estimate; multiple σ ranges could be required for optimal signal propagation. Alternatively, we derive this condition from a maximum-entropy argument for the signal distribution (see Appendix E). A similar-maximum entropy argument has been used by Bukva *et al.* [39] to study trainability of feed-forward networks. Note that due to the assumptions made in deriving this expression, it cannot fully capture the behavior of the signal when it reaches the nonlinear part of the activation function. However, this is solely a limitation of this particular ansatz to obtain an analytically tractable expression for ρ^*; the response function itself depends in a nonlinear manner on the hyperparameters of the network and on the nonlinear activation function of the network.

B. Depth scaling dominates optimal scaling

At large depth L, the expression for optimal scaling in (61) is dominated by the appearing Lth root and can be written as

$$\rho^* \approx \sqrt{\frac{1}{L}} \sqrt{\frac{1}{\sigma_w^2\, \phi'(0)^2} \log\left(\frac{\sigma_w^2\, \phi'(0)^2 \left(\frac{\mathcal{V}}{2}\right)^2 + \sigma_b^2}{\sigma_w^2\, \phi'(0)^2\, K^{(0)} + \sigma_b^2}\right)}, \tag{62}$$

neglecting terms of order $O(L^{-1})$. Thus, we obtain the proportionality of the optimal scaling value with $\propto 1/\sqrt{L}$ in Fig. 4(c); this scaling has been reported based on different approaches in earlier works [15–17]. In contrast to those works, we here also obtain the dependence on other hyperparameters of the network. While the scaling with $1/\sqrt{L}$ dominates, there is a weak dependence on other hyperparameters due to the appearing logarithm as shown in Fig. 5. This weak dependence on network hyperparameters but strong dependence on network depth offers an explanation for the widespread success of the $1/\sqrt{L}$ scaling across different architectures [18].

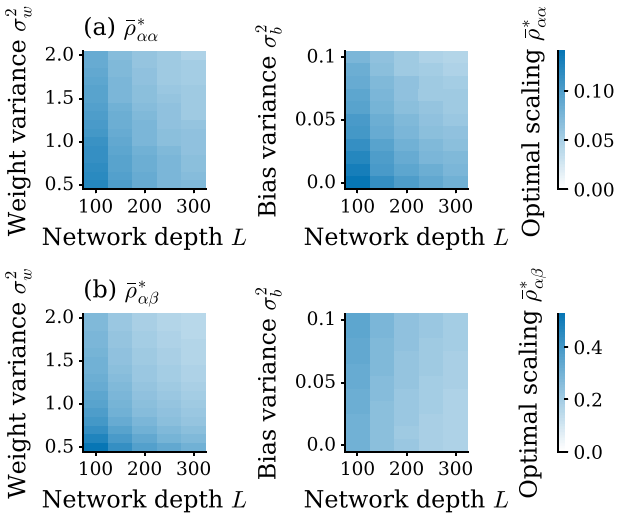

FIG. 5. Optimal scalings depend strongly on network depth but weakly on other hyperparameters. We illustrate the weak dependence on the weight variance σ_w^2 and bias variance σ_b^2 relative to the network depth L for CIFAR-10 for both (a) variances and (b) covariances; samples are either dogs or airplanes. We measure the scaling with maximal output response averaged over all diagonal or all off-diagonal elements of the covariance, $\bar{\rho}^*_{\alpha\alpha} = \frac{1}{N}\sum_\alpha \operatorname{argmax}(\chi^{\text{out}}_{\alpha\alpha})$ or $\bar{\rho}^*_{\alpha\beta} = \frac{1}{N(N-1)}\sum_{\alpha\neq\beta} \operatorname{argmax}(\chi^{\text{out}}_{\alpha\beta})$. Other parameters: data set size $P=20$, input scale $K^{(0)}=0.05$, $\sigma_w^2=1.25$, $\sigma_b^2=0.05$, $d_{\text{in}}=d_{\text{out}}=100$, $N=500$, $\phi=\text{erf}$.

C. Behavior across full dataset

While the above considerations apply to the diagonal elements $K^{(l)}_{\alpha\alpha}$ of the network kernels, i.e., for the statistics of a single sample x_α, we require efficient signal propagation for all samples of a data set of size P. The joint statistics of multiple data samples cannot be analyzed with the simplified intuition of saturation arguments and variances. The non-approximated field-theoretic result, however, still yields the behavior of the full covariances. The off-diagonal elements do not get a contribution from the term involving the second derivative in (45), which is negative for $\phi=\text{erf}$ and other saturating activation functions. In consequence, the response function for off-diagonal elements is larger and admits a larger residual scaling ρ^*.

To investigate the dependence of the optimal scaling with maximal output response $\rho^*(K_{\alpha\beta}) = \operatorname{argmax}(\chi^{\text{out}}_{\alpha\beta})$ on differences in data samples, we generate samples of unit length and encode sample differences by angles that have an equal spacing in the range of $[0, 2\pi]$ by steps of $2\pi/P$. In Fig. 6(a) we observe a noticeable angular dependence of the optimal scaling. For common data sets such as MNIST and CIFAR-10 [see Figs. 6(b)–6(c)], we find that optimal scalings $\rho^*(K_{\alpha\beta})$ for off-diagonal elements of the kernels behave rather homogeneously, even though samples stem from two different classes. Thus, the average scalings $\bar{\rho}^* = \frac{1}{N(N-1)}\sum_{\alpha\neq\beta} \operatorname{argmax}(\chi^{\text{out}}_{\alpha\beta})$ for off-diagonal elements of each data set provide a good indicator for suitable values.

Further, we find the same $1/\sqrt{L}$ scaling of variances for the covariances in Fig. 4(c). Apart from this strong dependence on network depth, there is again only a weak dependence of the optimal scaling on other hyperparameters [see Fig. 5(b)].

In summary, the response function, which naturally arises in a field-theoretic formulation of residual networks, predicts the dependence of optimal scaling ρ^* on all network hyperparameters and data statistics and thus provides a theoretical explanation for the empirically well-tested $1/\sqrt{L}$ scaling.

D. Behavior in trained networks

We obtain the response function as a next-to-leading-order term to the NNGP from the network prior. Thus, its form strictly only holds for networks at initialization. Schoenholz *et al.* [36] show for feed-forward networks that signal propagation at initialization is a useful indicator of network performance at training time. We therefore study in this section whether the properties of the response function in residual networks are robust at training time.

For the chosen scaling of weights $w \propto 1/\sqrt{N}$, the prior distribution of weights is known to change only very mildly due to training, a result often termed "lazy learning" [40]. One therefore expects that predictions of the theory maintain their validity, at least approximately. To check this, we train networks on a binary classification between digits 0 and 3 of MNIST using Langevin gradient descent [21,41,42] until convergence. The stationary weight distribution can alternatively be regarded as drawn from the Bayesian posterior.

As shown in the top panels of Fig. 7, we find that for $\rho=1$ the response is indeed robust with regard to network changes during training. Comparing the empirically measured network kernels at initialization [Fig. 7(c), left] and after training [Fig. 7(c), right], we observe that these change only mildly despite the network having learned to solve the task [Fig. 7(b)], explaining the robustness of the response. Choosing optimal residual scaling $\rho \approx 1/\sqrt{L}$ (Fig. 7 bottom), however, leads to a noticeable adaptation of the kernels to the data after training [Fig. 7(c)]. As a result, the response across layers increases [Fig. 7(a)] and, while network learning is slower across training epochs, the network generalizes well during the entire training process [Fig. 7(b)]. Thus, utilizing a residual scaling $\rho = 1/\sqrt{L}$ that promotes better signal propagation across layers results in stronger network adaptation and generalization properties of the network. These results are in line with the response functions driving gradient-based learning [36], which is also predicted by the theory of feature learning in feed-forward networks [27,29].

We hypothesize that, while the response function increases towards the optimal residual scaling $\rho = 1/\sqrt{L}$, its fundamental dependence on network hyperparameters such as the depth L does not change significantly, explaining the success of this scaling choice in trained networks. Studying this hypothesis analytically requires determining the Bayesian posterior of the network, which we leave for future work.

IV. DISCUSSION

Understanding signal propagation in neural networks is essential for a theory of trainability and generalization. Regarding these points, residual networks have shown to be superior to feed-forward network [3,4]; scaling the residual

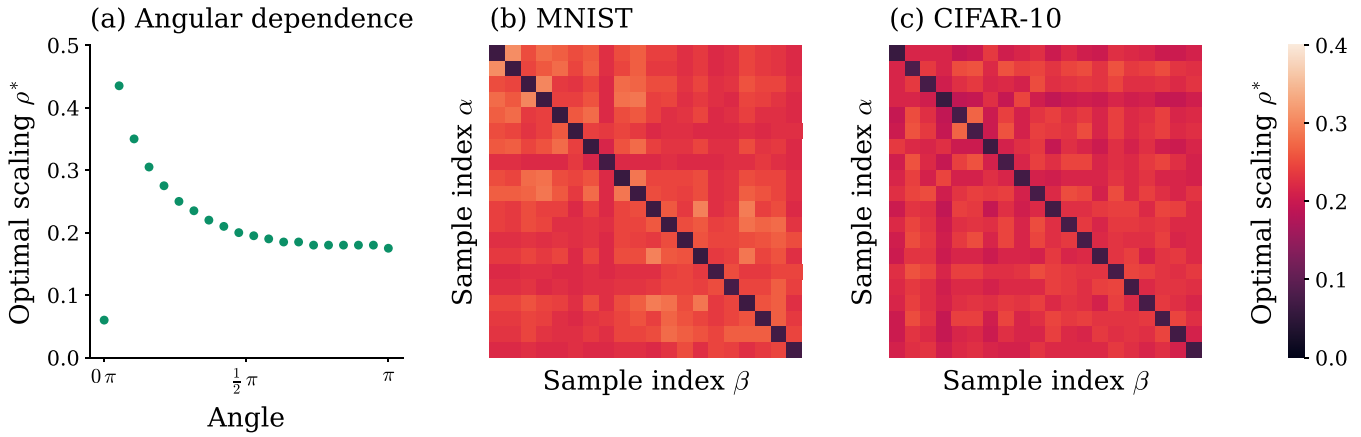

FIG. 6. Dependence of optimal scaling ρ^* on sample variability. (a) Dependence of optimal scaling on the angle between normalized synthetic data samples. On common, real-world data sets such as (b) MNIST and (c) CIFAR-10, optimal scalings ρ^* for off-diagonal elements of the kernels (samples sorted by classes only) yield similar values across data samples. For MNIST we use samples from digit 0 and digit 3, and for CIFAR-10 samples are either dogs or airplanes. Other parameters: data set size $P = 20$, input scale $K^{(0)} = 0.05$, $\sigma_w^2 = 1.25$, $\sigma_b^2 = 0.05$, $d_{\text{in}} = d_{\text{out}} = 100$, $N = 500$, $L = 200$, $\phi = \text{erf}$.

branches in ResNets further amplifies this effect [6]. We here derive a field-theoretic formulation of residual networks that allows us to determine finite-size corrections in a systematic way. The response function of residual networks, a measure for the network's sensitivity to variability in the input and thus network trainability, naturally appears as the leading-order correction to the NNGP. We show that, in contrast to feed-forward networks, the response function decays to zero only asymptotically, consequently allowing information to propagate to very deep layers, in line with Yang and Schoenholz [38]. Further, we show that signal propagation in ResNets is optimal when the signal distribution utilizes the whole dynamic range of the activation function. Beyond this range, information is lost due to saturation effects. Notably, we are able to explain the universality of empirically found optimal values due to a weak dependence on all network hyperparameters but the network depth. Thereby, this work sheds light on the interplay between signal propagation, saturation effects and signal scales in residual networks. Finally, we show for trained networks that optimal residual scaling does not only improve signal propagation but also generalization capabilities.

A. Limitations

Since we study the network prior of residual networks, the obtained analytical results strictly only apply to networks at initialization. We provide empirical results for trained networks, however, indicating that the hyperparameter dependence of the response function may be robust at training time. Studying the properties of the response after training analytically requires determining the full Bayesian posterior, which we leave for future work.

The presented field-theoretic formalism allows the systematic computation of finite-size corrections from an expansion of the exact action. We here truncated this expansion at first order, which yields the response function. In general, higher orders may appear and are computationally more costly to obtain. By definition, these, however, do not alter the response function, which was studied in this work.

B. Related works

Understanding the effect of skip connections in residual networks has been studied using different quantities: Building on the empirical observation that connections skipping a certain number of fully connected layers lead to smaller training errors, Li *et al.* [43] show that the condition number of the Hessian of the loss function does not grow with network depth but is depth-invariant. Further, ResNets achieve improved data separability compared to FFNets as they preserve angles between samples and thus exhibit less degradation of the ratio between within-class distance and between-class distance [44].

While we here explicitly focus on finite-size networks and the implications of the finite size, other works examine the scaling behavior in the infinite-size limit. Yang and Schoenholz [38] explicitly study signal propagation in residual networks in the infinite-width limit by determining the decay rate of sample correlations to their fixed point values. They find a subexponential or even polynomial decay, indicating that, in contrast to feed-forward networks [36,45,46], residual networks are always close to the edge of chaos, leading to better trainability also at great depth. We recover their result and go beyond it by investigating the effect of scaling the residual branch on signal propagation. The closeness to the edge of chaos they describe can be linked to the long depth scale and increased amplitude of the response function we demonstrate here.

Neural ordinary differential equations (neural ODEs) [47] consider the limit of infinite depth for residual networks, yielding a set of differential equations describing the network dynamics. Cohen *et al.* [48] find three different scaling regimes of neural ODEs depending on the choice of architecture and activation function. Barboni *et al.* provide global convergence guarantees for gradient descent and optimal transport training [49,50]. Chen *et al.* [51] derive generaliza-

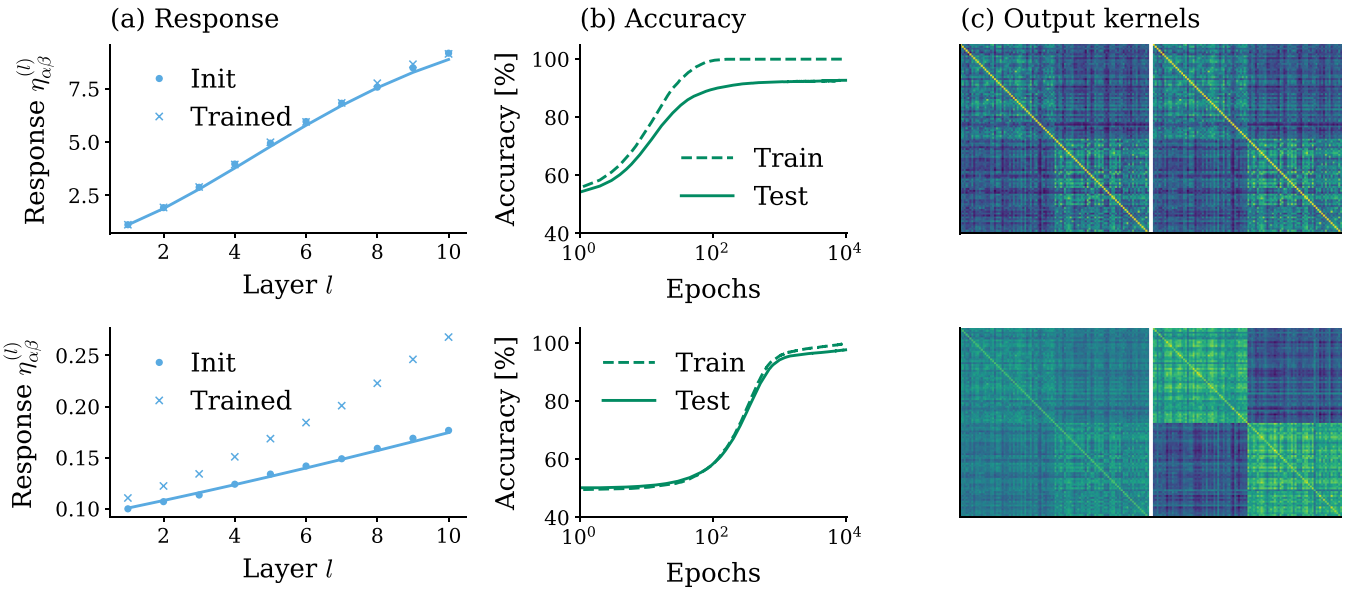

FIG. 7. Behavior of the response function in trained networks. We show results for $\rho = 1.0$ (upper row) and close to optimal scaling $\rho = 0.3 \approx 1/\sqrt{L}$ (lower row). (a) For the residual response $\eta^{(l)}_{\alpha\beta}$ of the covariances, we show empirical results at initialization (dots) and after training (crosses) as well as theoretical predictions from the network prior (solid lines). While the response is almost unchanged after training for $\rho = 1.0$, the response is increased after training for close to optimal residual scaling $\rho = 0.3 \approx 1/\sqrt{L}$. The empirical values are measured by finite differences of (31) with $\epsilon = 10^{-6}$ and rescaling the variances to their original values so that only the covariances are perturbed. (b) The increased response coincides with slower learning but improved generalization throughout training time. (c) The empirically measured output kernels before (left) and after training (right) indicate significantly stronger network adaptation to the data close to the optimal residual scaling $\rho = 0.3 \approx 1/\sqrt{L}$. Other parameters: binary classification on MNIST between digits 0 and 3, training data set size $P = 100$, $N_{\text{epochs}} = 10^4$, $\sigma_w^2 = 1.25$, $\sigma_b^2 = 0$, $d_{\text{in}} = 784$, $d_{\text{out}} = 1$, $N = 500$, $L = 10$, $\phi = \text{erf}$.

tion bounds beyond the NTK regime. Li and Nica [52] find a connection between the infinite-depth and -width limits for residual networks with a scaling by square root of the inverse depth on the residual branches and feed-forward networks with activation functions scaled by the same factor, building on the $1/\sqrt{\text{depth}}$ scaling for which we here derive the theoretical basis. Marion *et al.* [53] find a strong relationship between residual scaling and the regularity of the weights for obtaining nontrivial dynamics in residual networks, impacting performance for initialized and trained networks.

The residual scaling parameter affects the behavior of gradients in ResNets: Smaller scaling values reduce the whitening of gradients with increasing depth [54]. Zaeemzadeh *et al.* [55] show that skip connections lead to norm preservation of the gradients during backpropagation by shifting the singular values closer to one; norm preservation in turn improves trainability and generalization. Regarding the problem of vanishing or exploding gradients, Ling and Qiu [56] require the singular values of the input-output Jacobian to be of order one, suggesting a scaling of the weight variance by the square root of the inverse depth. While their result holds only in the double limit of infinite width and depth, our theoretical predictions apply to finite-size networks.

Another line of research studies the scaling of the skip connection instead of the residual branch: According to Zhang *et al.* [57], scaling the skip connection improves semantic feature learning; suggesting an inverse scaling with depth while observing that the particular decay scheme is less relevant than the total decay over all layers. Doshi *et al.* [58] investigate critical points in residual networks, observing only a weak dependence on other hyperparameters such as weight and bias variance at initialization similar as the weak dependence explained by the here developed theory; they empirically compute partial Jacobians while we study the response function that naturally appears in the presented field-theoretic framework. Scaling the skip connections can straightforwardly be included into this framework; we leave this for future work.

C. Outlook

The field-theoretic formulation in Sec. II allows the systematic computation of finite-size corrections to the NNGP result. Fischer *et al.* [29] compute the Bayesian posterior network kernel that results from fluctuation corrections to the NNGP kernels in feed-forward networks. In the future, we plan to investigate the effect of skip connections on the network posterior in this framework. Further, Lindner *et al.* [31] employ a related framework for studying the effect of data variability on the network posterior, which is also an interesting avenue of future research for residual networks.

ACKNOWLEDGMENTS

We thank Peter Bouss and Claudia Merger for helpful discussions. This work was partly supported by the German Federal Ministry for Education and Research (BMBF Grant No. 01IS19077A to Jülich and BMBF Grant No. 01IS19077B to Aachen) and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Founda-

tion) No. 368482240/GRK2416, the Excellence Initiative of the German federal and state governments (ERS PF-JARA-SDS005), and the Helmholtz Association Initiative and Networking Fund under Project No. SO-092 (Advanced Computing Architectures, ACA). 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 No. JINB33). Open access publication funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), No. 491111487.

DATA AVAILABILITY

The data that support the findings of this article are openly available [33].

APPENDIX A: NETWORK PRIOR FOR MULTIPLE INPUTS

We here calculate an analytical expression for the network prior $p(Y|X)$, derived in Sec. II for a single input, for multiple inputs $X = (x_\alpha)_{\alpha=1,\dots,P}$ and network outputs $Y = (y_\alpha)_{\alpha=1,\dots,P}$:

$$p(Y|X) = \int \mathrm{d}\theta \prod_\alpha p(y_\alpha|x_\alpha,\theta)\, p(\theta). \tag{A1}$$

For fixed network parameters θ, the probability $p(Y|X,\theta)$ is given by enforcing the network model with Dirac δ distributions as

$$\begin{aligned} p(Y|X,\theta) = \prod_\alpha \int \mathrm{d}h_\alpha^{(0)} \cdots \int \mathrm{d}h_\alpha^{(L)}\, \delta\big(h_\alpha^{(0)} - W^{\mathrm{in}} x_\alpha - b^{\mathrm{in}}\big) \prod_{l=1}^{L} \delta\big(h_\alpha^{(l)} - h_\alpha^{(l-1)} - \rho W^{(l)} \phi\big(h_\alpha^{(l-1)}\big) - \rho b^{(l)}\big) \\ \times\, \delta\big(y_\alpha - W^{\mathrm{out}} \phi\big(h_\alpha^{(L)}\big) - b^{\mathrm{out}}\big). \end{aligned} \tag{A2}$$

In contrast to the case of a single input x_α in the main text, we here need to multiply across all P inputs $X = (x_\alpha)_{\alpha=1,\dots,P}$.

1. Marginalization over network parameters

We write the marginalization over network parameters

$$\begin{aligned} p(Y|X) = \prod_\alpha \int \mathrm{d}h_\alpha^{(0)} \cdots \int \mathrm{d}h_\alpha^{(L)} \left\langle \delta(h_\alpha^{(0)} - W^{\mathrm{in}} x_\alpha - b^{\mathrm{in}}) \right\rangle_{\{W^{\mathrm{in}},\, b^{\mathrm{in}}\}} \prod_{l=1}^{L} \left\langle \delta\big(h_\alpha^{(l)} - h_\alpha^{(l-1)} - \rho W^{(l)} \phi\big(h_\alpha^{(l-1)}\big) - \rho b^{(l)}\big) \right\rangle_{\{W^{(l)},\, b^{(l)}\}} \\ \times \left\langle \delta\big(y_\alpha - W^{\mathrm{out}} \phi\big(h_\alpha^{(L)}\big) - b^{\mathrm{out}}\big) \right\rangle_{\{W^{\mathrm{out}},\, b^{\mathrm{out}}\}}, \end{aligned} \tag{A3}$$

where $\langle \dots \rangle_{\{W,b\}}$ refers to the Gaussian average over weights W and biases b according to the priors on the network parameters: For the input layer $W_{ij}^{\mathrm{in}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{w,\,\mathrm{in}}^2/d_{\mathrm{in}})$, $b_i^{\mathrm{in}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{b,\,\mathrm{in}}^2)$, for residual layers $W_{ij}^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2/N)$, $b_i^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_b^2)$, and for the readout layer $W_{ij}^{\mathrm{out}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{w,\,\mathrm{out}}^2/N)$, $b_i^{\mathrm{out}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{b,\,\mathrm{out}}^2)$. We use the Fourier representation of the Dirac δ distribution

$$\delta(z) = \int \mathrm{d}\tilde{z}\, \exp(\tilde{z}^{\mathsf{T}} z) \tag{A4}$$

with $\tilde{z}$ the conjugate variable to z and $\tilde{z}^{\mathsf{T}} z = \sum_{i=1}^{N} \tilde{z}_i z_i$, where we integrate along the imaginary axis $\int \mathrm{d}\tilde{z} = \prod_k \int_{i\mathbb{R}} \frac{\mathrm{d}\tilde{z}_k}{2\pi i}$ for notational convenience, and rewrite

$$\begin{aligned} p(Y|X) = \prod_\alpha \left\{ \int \mathcal{D}\tilde{y}_\alpha \int \mathcal{D}\tilde{h}_\alpha \int \mathcal{D}h_\alpha \right\} \left\langle \exp\left(\sum_{\alpha,i} \tilde{h}_{i,\alpha}^{(0)} \left(h_{i,\alpha}^{(0)} - \sum_j W_{ij}^{\mathrm{in}} x_{j,\alpha} - b_i^{\mathrm{in}} \right) \right) \right\rangle_{\{W^{\mathrm{in}},\, b^{\mathrm{in}}\}} \\ \times \prod_{l=1}^{L} \left\langle \exp\left(\sum_{\alpha,i} \tilde{h}_{i,\alpha}^{(l)} \left(h_{i,\alpha}^{(l)} - h_{i,\alpha}^{(l-1)} - \rho \sum_j W_{ij}^{(l)} \phi_{j,\alpha}^{(l-1)} - \rho b_i^{(l)} \right) \right) \right\rangle_{\{W^{(l)},\, b^{(l)}\}} \\ \times \left\langle \exp\left(\sum_{\alpha,i} \tilde{y}_{i,\alpha} \left(y_{i,\alpha} - \sum_j W_{ij}^{\mathrm{out}} \phi_{j,\alpha}^{(L)} - b_i^{\mathrm{out}} \right) \right) \right\rangle_{\{W^{\mathrm{out}},\, b^{\mathrm{out}}\}}, \end{aligned} \tag{A5}$$

where $\int \mathcal{D}h_\alpha = \prod_{l=0}^{L} \int \mathrm{d}h_\alpha^{(l)}$ and $\int \mathcal{D}\tilde{h}_\alpha = \prod_{l=0}^{L} \int \mathrm{d}\tilde{h}_\alpha^{(l)}$. We introduced the shorthand $\phi_{j,\alpha}^{(l-1)} = \phi(h_{j,\alpha}^{(l-1)})$ for brevity. Since the network parameters θ_k are independently distributed, we can compute the averages for each parameter separately $\int \mathrm{d}\theta_k\, p(\theta_k) \exp(z\theta_k)$, which yields the moment-generating function of the distribution $p(\theta_k)$. For Gaussian distributed

parameters $\theta_k \sim \mathcal{N}(0, \sigma^2)$, this gives

$$\int \mathrm{d}\theta_k\, p(\theta_k)\, \exp(z\theta_k) = \int \mathrm{d}\theta_k\, \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\theta_k^2\right) \exp(z\theta_k) \tag{A6}$$

$$= \int \mathrm{d}\theta_k\, \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\theta_k - \sigma z)^2\right) \exp\left(\frac{1}{2}\sigma^2 z^2\right) \tag{A7}$$

$$= \exp\left(\frac{1}{2}\sigma^2 z^2\right). \tag{A8}$$

For the individual terms in (A5), we then get

$$\left\langle \exp\left(-\sum_{i,j} W^{\mathrm{in}}_{ij} \sum_{\alpha} \tilde{h}^{(0)}_{i,\alpha} x_{j,\alpha}\right)\right\rangle_{W^{\mathrm{in}}} = \exp\left(\frac{1}{2}\frac{\sigma^2_{w,\,\mathrm{in}}}{d_{\mathrm{in}}} \sum_{i,j}\left[\sum_{\alpha} \tilde{h}^{(0)}_{i,\alpha} x_{j,\alpha}\right]^2\right) = \exp\left(\frac{1}{2}\frac{\sigma^2_{w,\,\mathrm{in}}}{d_{\mathrm{in}}} \sum_{\alpha,\beta}\sum_{i,j} \tilde{h}^{(0)}_{i,\alpha} x_{j,\alpha}\, \tilde{h}^{(0)}_{i,\beta} x_{j,\beta}\right), \tag{A9}$$

$$\left\langle \exp\left(-\sum_{i} b^{\mathrm{in}}_{i} \sum_{\alpha} \tilde{h}^{(0)}_{i,\alpha}\right)\right\rangle_{b^{\mathrm{in}}} = \exp\left(\frac{1}{2}\sigma^2_{b,\,\mathrm{in}} \sum_{i}\left[\sum_{\alpha} \tilde{h}^{(0)}_{i,\alpha}\right]^2\right) = \exp\left(\frac{1}{2}\sigma^2_{b,\,\mathrm{in}} \sum_{\alpha,\beta}\sum_{i} \tilde{h}^{(0)}_{i,\alpha}\, \tilde{h}^{(0)}_{i,\beta}\right), \tag{A10}$$

$$\left\langle \exp\left(-\sum_{i,j} W^{(l)}_{ij}\, \rho \sum_{\alpha} \tilde{h}^{(l)}_{i,\alpha} \phi^{(l-1)}_{j,\alpha}\right)\right\rangle_{W^{(l)}} = \exp\left(\frac{1}{2}\frac{\sigma^2_w}{N} \sum_{i,j}\left[\rho \sum_{\alpha} \tilde{h}^{(l)}_{i,\alpha} \phi^{(l-1)}_{j,\alpha}\right]^2\right) = \exp\left(\frac{1}{2}\rho^2\frac{\sigma^2_w}{N} \sum_{\alpha,\beta}\sum_{i,j} \tilde{h}^{(l)}_{i,\alpha} \phi^{(l-1)}_{j,\alpha}\, \tilde{h}^{(l)}_{i,\beta} \phi^{(l-1)}_{j,\beta}\right), \tag{A11}$$

$$\left\langle \exp\left(-\sum_{i} b^{(l)}_{i}\, \rho \sum_{\alpha} \tilde{h}^{(l)}_{i,\alpha}\right)\right\rangle_{b^{(l)}} = \exp\left(\frac{1}{2}\sigma^2_b \sum_{i}\left[\rho \sum_{\alpha} \tilde{h}^{(l)}_{i,\alpha}\right]^2\right) = \exp\left(\frac{1}{2}\rho^2\sigma^2_b \sum_{\alpha,\beta}\sum_{i} \tilde{h}^{(l)}_{i,\alpha}\, \tilde{h}^{(l)}_{i,\beta}\right), \tag{A12}$$

$$\left\langle \exp\left(-\sum_{i,j} W^{\mathrm{out}}_{ij} \sum_{\alpha} \tilde{y}_{i,\alpha} \phi^{(L)}_{j,\alpha}\right)\right\rangle_{W^{\mathrm{out}}} = \exp\left(\frac{1}{2}\frac{\sigma^2_{w,\,\mathrm{out}}}{N} \sum_{i,j}\left[\sum_{\alpha} \tilde{y}_{i,\alpha} \phi^{(L)}_{j,\alpha}\right]^2\right) = \exp\left(\frac{1}{2}\frac{\sigma^2_{w,\,\mathrm{out}}}{N} \sum_{\alpha,\beta}\sum_{i,j} \tilde{y}_{i,\alpha} \phi^{(L)}_{j,\alpha}\, \tilde{y}_{i,\beta} \phi^{(L)}_{j,\beta}\right), \tag{A13}$$

$$\left\langle \exp\left(-\sum_{i} b^{\mathrm{out}}_{i} \sum_{\alpha} \tilde{y}_{i,\alpha}\right)\right\rangle_{b^{\mathrm{out}}} = \exp\left(\frac{1}{2}\sigma^2_{b,\,\mathrm{out}} \sum_{i}\left[\sum_{\alpha} \tilde{y}_{i,\alpha}\right]^2\right) = \exp\left(\frac{1}{2}\sigma^2_{b,\,\mathrm{out}} \sum_{\alpha,\beta}\sum_{i} \tilde{y}_{i,\alpha} \tilde{y}_{i,\beta}\right). \tag{A14}$$

To ease notation, we use an implicit summation convention in the following for lower indices that appear twice as $\sum_\alpha \sum_i \tilde{h}^{(l)}_{i,\alpha}\, \tilde{h}^{(l)}_{i,\alpha} = \tilde{h}^{(l)}_{i,\alpha}\, \tilde{h}^{(l)}_{i,\alpha}$. Further, we write $\int \mathcal{D}\tilde{h} = \prod_\alpha \int \mathcal{D}\tilde{h}_\alpha$. Rewriting the sums over neuron indices as $\sum_{mn}[\tilde{h}_m \phi^{(l-1)}_n]^2 = \tilde{h}^\mathsf{T}\tilde{h}\,[\phi^{(l-1)}]^\mathsf{T}\phi^{(l-1)}$, we overall get for the prior

$$\begin{aligned} p(Y|X) = \int \mathcal{D}\tilde{y} \int \mathcal{D}\tilde{h} \int \mathcal{D}h\, \exp&\left(\tilde{y}^\mathsf{T}_\alpha y_\alpha + \frac{1}{2}\frac{\sigma^2_{w,\,\mathrm{out}}}{N} \tilde{y}^\mathsf{T}_\alpha \tilde{y}_\beta \left[\phi^{(L)}_\alpha\right]^\mathsf{T} \phi^{(L)}_\beta + \frac{1}{2}\sigma^2_{b,\,\mathrm{out}} \sum_{\alpha,\beta} \tilde{y}^\mathsf{T}_\alpha \tilde{y}_\beta + \sum_{l=1}^{L} \left[\tilde{h}^{(l)}_\alpha\right]^\mathsf{T}\left[h^{(l)}_\alpha - h^{(l-1)}_\alpha\right]\right) \\ &\times \exp\left(\sum_{l=1}^{L}\left(\frac{1}{2}\rho^2\frac{\sigma^2_w}{N}\left[\tilde{h}^{(l)}_\alpha\right]^\mathsf{T}\tilde{h}^{(l)}_\beta \left[\phi^{(l-1)}_\alpha\right]^\mathsf{T}\phi^{(l-1)}_\beta + \frac{1}{2}\rho^2\sigma^2_b \sum_{\alpha,\beta}\left[\tilde{h}^{(l)}_\alpha\right]^\mathsf{T}\tilde{h}^{(l)}_\beta\right)\right) \\ &\times \exp\left(\left[\tilde{h}^{(0)}_\alpha\right]^\mathsf{T} h^{(0)}_\alpha + \frac{1}{2}\frac{\sigma^2_{w,\,\mathrm{in}}}{d_{\mathrm{in}}}\left[\tilde{h}^{(0)}_\alpha\right]^\mathsf{T}\tilde{h}^{(0)}_\beta\, x_\alpha{}^\mathsf{T} x_\beta + \frac{1}{2}\sigma^2_{b,\,\mathrm{in}} \sum_{\alpha,\beta}\left[\tilde{h}^{(0)}_\alpha\right]^\mathsf{T}\tilde{h}^{(0)}_\beta\right) \end{aligned} \tag{A15}$$

$$=: \int \mathcal{D}\tilde{y} \int \mathcal{D}\tilde{h} \int \mathcal{D}h\, \exp(\mathcal{S}(Y, \tilde{Y}, H, \tilde{H}|X)). \tag{A16}$$

The action $\mathcal{S}$ consists of three contributions,

$$\mathcal{S}(Y, \tilde{Y}, H, \tilde{H}|X) = \mathcal{S}_{\mathrm{in}}(H^{(0)}, \tilde{H}^{(0)}|X) + \mathcal{S}_{\mathrm{net}}(H, \tilde{H}) + \mathcal{S}_{\mathrm{out}}(Y, \tilde{Y}|H^{(L)}), \tag{A17}$$

one term from the readin layer containing the dependence on the inputs X as

$$\mathcal{S}_{\text{in}}(H^{(0)},\tilde{H}^{(0)}|X) \coloneqq \left[\tilde{h}_\alpha^{(0)}\right]^\intercal h_\alpha^{(0)} + \frac{1}{2}\frac{\sigma_{w,\text{in}}^2}{d_{\text{in}}}\left[\tilde{h}_\alpha^{(0)}\right]^\intercal \tilde{h}_\beta^{(0)}\, x_\alpha{}^\intercal x_\beta + \frac{1}{2}\sigma_{b,\text{in}}^2 \sum_{\alpha,\beta}\left[\tilde{h}_\alpha^{(0)}\right]^\intercal \tilde{h}_\beta^{(0)}, \tag{A18}$$

one term from the intermediate layers containing the skip connections

$$\mathcal{S}_{\text{net}}(H,\tilde{H}) \coloneqq \sum_{l=1}^{L}\left[\tilde{h}_\alpha^{(l)}\right]^\intercal\left[h_\alpha^{(l)} - h_\alpha^{(l-1)}\right] + \frac{1}{2}\rho^2\frac{\sigma_w^2}{N}\left[\tilde{h}_\alpha^{(l)}\right]^\intercal \tilde{h}_\beta^{(l)}\left[\phi_\alpha^{(l-1)}\right]^\intercal \phi_\beta^{(l-1)} + \frac{1}{2}\rho^2\sigma_b^2\sum_{\alpha,\beta}\left[\tilde{h}_\alpha^{(l)}\right]^\intercal \tilde{h}_\beta^{(l)}, \tag{A19}$$

and one term for the readout layer containing the network outputs Y as

$$\mathcal{S}_{\text{out}}(Y,\tilde{Y}|H^{(L)}) \coloneqq \tilde{y}_\alpha^\intercal y_\alpha + \frac{1}{2}\frac{\sigma_{w,\text{out}}^2}{N}\tilde{y}_\alpha^\intercal\tilde{y}_\beta\left[\phi_\alpha^{(L)}\right]^\intercal\phi_\beta^{(L)} + \frac{1}{2}\sigma_{b,\text{out}}^2\tilde{y}_\alpha^\intercal\tilde{y}_\beta. \tag{A20}$$

2. Auxiliary variables

In the above integrals, we can only solve Gaussian terms exactly, which is limited to interaction terms between h and $\tilde{h}$ up to quadratic terms. In the action $\mathcal{S}$ in (A17) we have terms $\propto [\tilde{h}^{(l)}]^\intercal\tilde{h}^{(l)}\,[\phi^{(l-1)}]^\intercal\phi^{(l-1)}$ appearing, which are quartic in h and $\tilde{h}$ for linear activations $\phi(h) = h$ and of higher order for nonlinear activations ϕ. To decouple these interaction terms into lower-order ones, we introduce auxiliary variables

$$C_{\alpha\beta}^{(l)} \coloneqq \begin{cases} \frac{\sigma_{w,\text{in}}^2}{d_{\text{in}}}(XX^\intercal)_{\alpha\beta} + \sigma_{b,\text{in}}^2 & l = 0, \\ \left[\phi_\alpha^{(l-1)}\right]^\intercal\phi_\beta^{(l-1)} + \rho^2\sigma_b^2 & 1 \leqslant l \leqslant L, \\ \left[\phi_\alpha^{(L)}\right]^\intercal\phi_\beta^{(L)} + \sigma_{b,\text{out}}^2 & l = L+1. \end{cases} \tag{A21}$$

To account for the original higher-order interactions between $\tilde{h}$ and $\phi^{(l-1)}$, we enforce the definition of the auxiliary variables by Dirac δ distributions and obtain for the network prior

$$\begin{aligned} p(Y|X) = \int\mathcal{D}\tilde{y}\int\mathcal{D}\tilde{h}\int\mathcal{D}h\prod_{\alpha,\beta}\int\mathcal{D}C_{\alpha\beta}\,\exp\left(\tilde{y}_\alpha^\intercal y_\alpha + \frac{1}{2}C_{\alpha\beta}^{(L+1)}\tilde{y}_\alpha^\intercal\tilde{y}_\beta\right)\delta\left(C_{\alpha\beta}^{(L+1)} - \frac{\sigma_{w,\text{out}}^2}{N}\left[\phi_\alpha^{(L)}\right]^\intercal\phi_\beta^{(L)} - \sigma_{b,\text{out}}^2\right) \\ \times\exp\left(\sum_{l=1}^{L}\left[\left[\tilde{h}_\alpha^{(l)}\right]^\intercal\left[h_\alpha^{(l)} - h_\alpha^{(l-1)}\right] + \frac{1}{2}C_{\alpha\beta}^{(l)}\left[\tilde{h}_\alpha^{(l)}\right]^\intercal\tilde{h}_\beta^{(l)}\right]\right)\delta\left(C_{\alpha\beta}^{(l)} - \rho^2\frac{\sigma_w^2}{N}\left[\phi_\alpha^{(l-1)}\right]^\intercal\phi_\beta^{(l-1)} - \rho^2\sigma_b^2\right) \\ \times\exp\left(\left[\tilde{h}_\alpha^{(0)}\right]^\intercal h_\alpha^{(0)} + \frac{1}{2}C_{\alpha\beta}^{(0)}\left[\tilde{h}_\alpha^{(0)}\right]^\intercal\tilde{h}_\beta^{(0)}\right)\delta\left(C_{\alpha\beta}^{(0)} - \frac{\sigma_{w,\text{in}}^2}{d_{\text{in}}}x_\alpha^\intercal x_\beta - \sigma_{b,\text{in}}^2\right), \end{aligned} \tag{A22}$$

with $\int\mathcal{D}C_{\alpha\beta} = \prod_{l=0}^{L+1}\int\mathcal{D}C_{\alpha\beta}^{(l)}$. We rewrite the Dirac δ distributions using their Fourier representation

$$\delta\left(C_{\alpha\beta}^{(l)}\right) = \int_{i\mathbb{R}}\frac{\mathrm{d}\tilde{C}_{\alpha\beta}^{(l)}}{2\pi i}\,\exp\left(\tilde{C}_{\alpha\beta}^{(l)}\,C_{\alpha\beta}^{(l)}\right), \tag{A23}$$

introducing conjugate variables $\tilde{C}_{\alpha\beta}^{(l)}$ to the auxiliary variables $C_{\alpha\beta}^{(l)}$, as

$$\delta\left(d_{\text{in}}C_{\alpha\beta}^{(0)} - \sigma_{w,\text{in}}^2 x_\alpha^\intercal x_\beta - d_{\text{in}}\sigma_{b,\text{in}}^2\right) = \int_{i\mathbb{R}}\frac{\mathrm{d}\tilde{C}_{\alpha\beta}^{(0)}}{2\pi i}\,\exp\left(d_{\text{in}}\tilde{C}_{\alpha\beta}^{(0)}C_{\alpha\beta}^{(0)} - \sigma_{w,\text{in}}^2\,\tilde{C}_{\alpha\beta}^{(0)}\,x_\alpha^\intercal x_\beta - d_{\text{in}}\sigma_{b,\text{in}}^2\,\tilde{C}_{\alpha\beta}^{(0)}\right), \tag{A24}$$

$$\delta\left(NC_{\alpha\beta}^{(l)} - \rho^2\sigma_w^2\left[\phi_\alpha^{(l-1)}\right]^\intercal\phi_\beta^{(l-1)} - N\rho^2\sigma_b^2\right) = \int_{i\mathbb{R}}\frac{\mathrm{d}\tilde{C}_{\alpha\beta}^{(l)}}{2\pi i}\,\exp\left(N\tilde{C}_{\alpha\beta}^{(l)}C_{\alpha\beta}^{(l)} - \rho^2\sigma_w^2\,\tilde{C}_{\alpha\beta}^{(l)}\left[\phi_\alpha^{(l-1)}\right]^\intercal\phi_\beta^{(l-1)} - N\rho^2\sigma_b^2\,\tilde{C}_{\alpha\beta}^{(l)}\right), \tag{A25}$$

$$\delta\left(NC_{\alpha\beta}^{(L+1)} - \sigma_{w,\text{out}}^2\left[\phi_\alpha^{(L)}\right]^\intercal\phi_\beta^{(L)} - N\sigma_{b,\text{out}}^2\right) = \int_{i\mathbb{R}}\frac{\mathrm{d}\tilde{C}_{\alpha\beta}^{(L+1)}}{2\pi i}\,\exp\left(N\tilde{C}_{\alpha\beta}^{(L+1)}C_{\alpha\beta}^{(L+1)} - \sigma_{w,\text{out}}^2\,\tilde{C}_{\alpha\beta}^{(L+1)}\left[\phi_\alpha^{(L)}\right]^\intercal\phi_\beta^{(L)} - N\sigma_{b,\text{out}}^2\,\tilde{C}_{\alpha\beta}^{(L+1)}\right). \tag{A26}$$

Reordering the above terms, we obtain

$$\begin{aligned} p(Y|X) = \int\mathcal{D}\tilde{y}\int\mathcal{D}\tilde{h}\int\mathcal{D}h\int\mathcal{D}C\int\mathcal{D}\tilde{C}\,\exp\left(\tilde{y}_\alpha^\intercal y_\alpha + \frac{1}{2}C_{\alpha\beta}^{(L+1)}\tilde{y}_\alpha^\intercal\tilde{y}_\beta\right) \\ \times\exp\left(\sum_{l=1}^{L}\left[\left[\tilde{h}_\alpha^{(l)}\right]^\intercal\left[h_\alpha^{(l)} - h_\alpha^{(l-1)}\right] + \frac{1}{2}C_{\alpha\beta}^{(l)}\left[\tilde{h}_\alpha^{(l)}\right]^\intercal\tilde{h}_\beta^{(l)}\right]\right) \end{aligned}$$

$$\times \exp\left(\left[\tilde{h}_\alpha^{(0)}\right]^\mathsf{T} h_\alpha^{(0)} + \frac{1}{2} C_{\alpha\beta}^{(0)} \left[\tilde{h}_\alpha^{(0)}\right]^\mathsf{T} \tilde{h}_\beta^{(0)}\right)$$
$$\times \exp\left(-N \sum_{l=0}^{L+1} \nu_l \, C_{\alpha\beta}^{(l)} \tilde{C}_{\alpha\beta}^{(l)} + \rho^2 \sigma_w^2 \sum_{l=1}^{L} \tilde{C}_{\alpha\beta}^{(l)} \left[\phi_\alpha^{(l-1)}\right]^\mathsf{T} \phi_\beta^{(l-1)} + N \rho^2 \sigma_b^2 \sum_{l=1}^{L} \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(l)}\right)$$
$$\times \exp\left(\sigma_{w,\,\mathrm{out}}^2 \tilde{C}_{\alpha\beta}^{(L+1)} \left[\phi_\alpha^{(L)}\right]^\mathsf{T} \phi_\beta^{(L)} + N \sigma_{b,\,\mathrm{out}}^2 \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(L+1)}\right)$$
$$\times \exp\left(\sigma_{w,\,\mathrm{in}}^2 \tilde{C}_{\alpha\beta}^{(0)} \left(XX^\mathsf{T}\right)_{\alpha\beta} + d_{\mathrm{in}} \sigma_{b,\,\mathrm{in}}^2 \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(0)}\right), \tag{A27}$$

where we write $\int \mathcal{D}C = \prod_{\alpha,\beta} \{\int \mathcal{D}C_{\alpha\beta}\}$ and $\int \mathcal{D}\tilde{C} = \prod_{\alpha,\beta} \prod_{l=0}^{L+1} \int_{i\mathbb{R}} \frac{\mathrm{d}\tilde{C}_{\alpha\beta}^{(l)}}{2\pi i}$ and $\nu_l = 1 + \delta_{0l}\,(d_{\mathrm{in}}/N - 1)$. The variables $C_{\alpha\beta}^{(l)}$ and $\tilde{C}_{\alpha\beta}^{(l)}$ only couple to sums of $\tilde{h}_{i,\alpha}^{(l)}$ and $\phi_{i,\alpha}^{(l)}$ over all neuron indices i, so all components of $h_{i,\alpha}^{(l)}$ and $\tilde{h}_{i,\alpha}^{(l)}$ are identically distributed. We can thus rewrite the network prior in terms of scalar variables, pulling out a factor N in place of all previous scalar products so that

$$p(Y|X) = \int \mathcal{D}\tilde{y} \int \mathcal{D}\tilde{h} \int \mathcal{D}h \int \mathcal{D}C \int \mathcal{D}\tilde{C} \, \exp\left(\tilde{y}_\alpha^\mathsf{T} y_\alpha + \frac{1}{2} C_{\alpha\beta}^{(L+1)} \tilde{y}_\alpha^\mathsf{T} \tilde{y}_\beta\right)$$
$$\times \exp\left(N \sum_{l=1}^{L} \left[\tilde{h}_\alpha^{(l)} \left[h_\alpha^{(l)} - h_\alpha^{(l-1)}\right] + \frac{1}{2} \tilde{h}_\alpha^{(l)} C_{\alpha\beta}^{(l)} \tilde{h}_\beta^{(l)}\right]\right)$$
$$\times \exp\left(N \sum_{\alpha} \tilde{h}_\alpha^{(0)} h_\alpha^{(0)} + N \frac{1}{2} \sum_{\alpha,\beta} \tilde{h}_\alpha^{(0)} C_{\alpha\beta}^{(0)} \tilde{h}_\beta^{(0)}\right)$$
$$\times \exp\left(-N \sum_{l=0}^{L+1} \nu_l \, C_{\alpha\beta}^{(l)} \tilde{C}_{\alpha\beta}^{(l)} + N \rho^2 \sigma_w^2 \sum_{l=1}^{L} \phi_\alpha^{(l-1)} \tilde{C}_{\alpha\beta}^{(l)} \phi_\beta^{(l-1)} + N \rho^2 \sigma_b^2 \sum_{l=1}^{L} \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(l)}\right)$$
$$\times \exp\left(N \sigma_{w,\,\mathrm{out}}^2 \phi_\alpha^{(L)} \tilde{C}_{\alpha\beta}^{(L+1)} \phi_\beta^{(L)} + N \sigma_{b,\,\mathrm{out}}^2 \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(L+1)}\right)$$
$$\times \exp\left(\sigma_{w,\,\mathrm{in}}^2 \tilde{C}_{\alpha\beta}^{(0)} \left(XX^\mathsf{T}\right)_{\alpha\beta} + d_{\mathrm{in}} \sigma_{b,\,\mathrm{in}}^2 \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(0)}\right). \tag{A28}$$

We again write $h^{(l)}$ and $\tilde{h}^{(l)}$ for their scalar variables.

We aim to write the network prior $p(Y|X)$ in terms of the order parameters $C_{\alpha\beta}^{(l)}$ and its conjugate variables $\tilde{C}_{\alpha\beta}^{(l)}$. To this end, we move the integrals over $h^{(l)}$ and $\tilde{h}^{(l)}$ into the exponent, yielding

$$p(Y|X) = \int \mathcal{D}\tilde{y} \int \mathcal{D}C \int \mathcal{D}\tilde{C} \, \exp\left(\tilde{y}_\alpha^\mathsf{T} y_\alpha + \frac{1}{2} C_{\alpha\beta}^{(L+1)} \tilde{y}_\alpha^\mathsf{T} \tilde{y}_\beta - N \sum_{l=0}^{L+1} \nu_l \, C_{\alpha\beta}^{(l)} \tilde{C}_{\alpha\beta}^{(l)}\right)$$
$$\times \exp\left[N \ln \prod_{l=1}^{L} \int \mathcal{D}\tilde{h}^{(l)} \int \mathcal{D}h^{(l)} \, \exp\left(\tilde{h}_\alpha^{(l)} \left[h_\alpha^{(l)} - h_\alpha^{(l-1)}\right] + \frac{1}{2} \tilde{h}_\alpha^{(l)} C_{\alpha\beta}^{(l)} \tilde{h}_\beta^{(l)}\right)\right.$$
$$\times \exp\left(\rho^2 \sigma_w^2 \sum_{l=1}^{L} \phi_\alpha^{(l-1)} \tilde{C}_{\alpha\beta}^{(l)} \phi_\beta^{(l-1)} + \rho^2 \sigma_b^2 \sum_{l=1}^{L} \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(l)}\right)$$
$$\times \exp\left(\sigma_{w,\,\mathrm{out}}^2 \phi_\alpha^{(L)} \tilde{C}_{\alpha\beta}^{(L+1)} \phi_\beta^{(L)} + \sigma_{b,\,\mathrm{out}}^2 \sum_{\alpha,\beta} \tilde{C}_{\alpha\beta}^{(L+1)}\right)$$
$$\left. \times \int \mathcal{D}\tilde{h}^{(0)} \int \mathcal{D}h^{(0)} \exp\left(\tilde{h}_\alpha^{(0)} h_\alpha^{(0)} + \frac{1}{2} \tilde{h}_\alpha^{(0)} C_{\alpha\beta}^{(0)} \tilde{h}_\beta^{(0)}\right)\right.$$

$$\times \exp\left(\sigma^2_{w,\,\mathrm{in}}\,\tilde{C}^{(0)}_{\alpha\beta}\,(XX^\mathsf{T})_{\alpha\beta} + d_{\mathrm{in}}\sigma^2_{b,\,\mathrm{in}}\sum_{\alpha,\beta}\tilde{C}^{(0)}_{\alpha\beta}\right)\Bigg]$$

$$= \int \mathcal{D}\tilde{y}\left\langle \exp\left(\tilde{y}^\mathsf{T}_\alpha y_\alpha + \frac{1}{2}\tilde{y}^\mathsf{T}_\alpha C^{(L+1)}_{\alpha\beta}\tilde{y}_\beta\right)\right\rangle_{C,\tilde{C}}. \tag{A29}$$

The expectation value appearing in the last line is with respect to the auxiliary variable $C^{(l)}_{\alpha\beta}$ and its conjugate variable $\tilde{C}^{(l)}_{\alpha\beta}$, which are distributed according to the auxiliary action $p(C,\tilde{C}) \propto \exp[S(C,\tilde{C})]$ given by

$$\mathcal{S}(C,\tilde{C}) \coloneqq -N\sum_{l=0}^{L+1}\nu_l\,C^{(l)}_{\alpha\beta}\,\tilde{C}^{(l)}_{\alpha\beta} + N\mathcal{W}(\tilde{C}|C), \tag{A30}$$

with

$$\begin{aligned}\mathcal{W}(\tilde{C}|C) \coloneqq \ln\prod_{l=1}^{L}\int\mathcal{D}\tilde{h}^{(l)}\int\mathcal{D}h^{(l)}\,&\exp\left(\tilde{h}^{(l)}_\alpha\left[h^{(l)}_\alpha - h^{(l-1)}_\alpha\right] + \frac{1}{2}\tilde{h}^{(l)}_\alpha C^{(l)}_{\alpha\beta}\tilde{h}^{(l)}_\beta\right)\\ &\times\exp\left(\rho^2\sigma^2_w\sum_{l=1}^{L}\phi^{(l-1)}_\alpha\tilde{C}^{(l)}_{\alpha\beta}\phi^{(l-1)}_\beta + \rho^2\sigma^2_b\sum_{l=1}^{L}\sum_{\alpha,\beta}\tilde{C}^{(l)}_{\alpha\beta}\right)\\ &\times\exp\left(\sigma^2_{w,\,\mathrm{out}}\,\phi^{(L)}_\alpha\tilde{C}^{(L+1)}_{\alpha\beta}\phi^{(L)}_\beta + \sigma^2_{b,\,\mathrm{out}}\sum_{\alpha,\beta}\tilde{C}^{(L+1)}_{\alpha\beta}\right)\\ &\times\int\mathcal{D}\tilde{h}^{(0)}\int\mathcal{D}h^{(0)}\exp\left(N\tilde{h}^{(0)}_\alpha h^{(0)}_\alpha + N\frac{1}{2}\tilde{h}^{(0)}_\alpha C^{(0)}_{\alpha\beta}\tilde{h}^{(0)}_\beta + \sigma^2_{w,\,\mathrm{in}}\,\tilde{C}^{(0)}_{\alpha\beta}\,(XX^\mathsf{T})_{\alpha\beta} + d_{\mathrm{in}}\sigma^2_{b,\,\mathrm{in}}\sum_{\alpha,\beta}\tilde{C}^{(0)}_{\alpha\beta}\right).\end{aligned} \tag{A31}$$

Note that the conjugate variables $\tilde{C}^{(l)}$ are not proper random variables, but will be integrated out later to obtain the statistics of the auxiliary variables $C^{(l)}$.

APPENDIX B: SADDLE POINT APPROXIMATION

Since the action $\mathcal{S}$ scales with the network width N, we can perform a saddle point approximation for infinite width $N \to \infty$. We compute the saddle points $C_*,\ \tilde{C}_*$ using the conditions

$$\frac{\partial\mathcal{S}}{\partial C^{(l)}_{\alpha\beta}} = 0, \quad \frac{\partial\mathcal{S}}{\partial\tilde{C}^{(l)}_{\alpha\beta}} = 0, \tag{B1}$$

and obtain

$$C^{(l)}_{\alpha\beta,*} = \begin{cases}\frac{\sigma^2_{w,\,\mathrm{in}}}{d_{\mathrm{in}}}(XX^\mathsf{T})_{\alpha\beta} + \sigma^2_{b,\,\mathrm{in}} & l = 0,\\ \rho^2\sigma^2_w\left\langle\phi^{(l-1)}_\alpha\,\phi^{(l-1)}_\beta\right\rangle_p + \rho^2\sigma^2_b & 1 \leqslant l \leqslant L,\\ \sigma^2_{w,\,\mathrm{out}}\left\langle\phi^{(L)}_\alpha\,\phi^{(L)}_\beta\right\rangle_p + \sigma^2_{b,\,\mathrm{out}} & l = L+1,\end{cases} \tag{B2}$$

$$\tilde{C}^{(l)}_{\alpha\beta,*} = \begin{cases}\frac{1}{2}\left\langle\tilde{h}^{(l)}_\alpha\tilde{h}^{(l)}_\beta\right\rangle_p = 0 & 0 \leqslant l \leqslant L,\\ \frac{1}{2}\left\langle\tilde{y}_\alpha\tilde{y}_\beta\right\rangle_p = 0 & l = L+1,\end{cases} \tag{B3}$$

with $\tilde{y} = \tilde{h}^{(L+1)}$ for notational brevity in the following and

$$\begin{aligned}\langle\cdots\rangle_p = \prod_{l=1}^{L}\int\mathcal{D}\tilde{h}^{(l)}\int\mathcal{D}h^{(l)}\,\cdots\,&\exp\left(\tilde{h}^{(l)}_\alpha\left[h^{(l)}_\alpha - h^{(l-1)}_\alpha\right] + \frac{1}{2}\tilde{h}^{(l)}_\alpha C^{(l)}_{\alpha\beta,*}\tilde{h}^{(l)}_\beta\right)\\ &\times\int\mathcal{D}\tilde{h}^{(0)}\int\mathcal{D}h^{(0)}\exp\left(\tilde{h}^{(0)}_\alpha h^{(0)}_\alpha + \frac{1}{2}\tilde{h}^{(0)}_\alpha C^{(0)}_{\alpha\beta,*}\tilde{h}^{(0)}_\beta\right).\end{aligned} \tag{B4}$$

We show that the second moment of the auxiliary variables $\tilde{h}$ in (B3) vanishes: We obtain the moments from the moment-generating function defined as

$$\mathcal{Z}[j,\tilde{j}] \coloneqq \int\mathcal{D}h\int\mathcal{D}\tilde{h}\,\exp[\mathcal{S}(h,\tilde{h}) + j^\mathsf{T}h + \tilde{j}^\mathsf{T}\tilde{h}] \tag{B5}$$

by taking derivatives with respect to the source term $\tilde{j}$

$$\langle \tilde{h}_{\alpha_1} \cdots \tilde{h}_{\alpha_m} \rangle = \frac{\partial^m \mathcal{Z}[j, \tilde{j}]}{\partial \tilde{j}_{\alpha_1} \cdots \partial \tilde{j}_{\alpha_m}}|_{j,\tilde{j}=0}, \tag{B6}$$

where $\mathcal{S}(h, \tilde{h})$ is given by (A17). Due to the normalization of the probability distribution, it holds that

$$\mathcal{Z}[0, \tilde{j}] = 1 \quad \forall \tilde{j}. \tag{B7}$$

Thus, all derivatives in (B6) vanish and consequently also all moments involving only auxiliary variables $\tilde{h}$.

While the input kernel $C^{(0)} = C^{(0)}_*$ in (B2) is fixed by the data $C^{(0)}_{\alpha\beta,*} = \sigma^2_{w,\text{in}}/d_{\text{in}}\,(XX^\mathsf{T})_{\alpha\beta} + \sigma^2_{b,\text{in}}$, all other residual kernels $C^{(l)}_*$ are computed self-consistently with respect to the expectation value involving $C^{(l)}_*$. We can rewrite this average by defining the residual $f^{(l)} \coloneqq h^{(l)} - h^{(l-1)}$ for $1 \leqslant l \leqslant L$ and $f^{(0)} \coloneqq h^{(0)}$, yielding

$$\langle \cdots \rangle_p = \prod_{l=0}^{L} \int \mathcal{D}f^{(l)} \int \mathcal{D}\tilde{h}^{(l)} \cdots \exp\left(\tilde{h}^{(l)}_\alpha f^{(l)}_\alpha + \frac{1}{2} \tilde{h}^{(l)}_\alpha C^{(l)}_{\alpha\beta,*} \tilde{h}^{(l)}_\beta \right). \tag{B8}$$

We identify the appearing terms with the moment-generating function of a Gaussian with zero mean and covariance matrix $C^{(l)}_{\alpha\beta,*}$, from which follows that the residuals $F^{(l)} = (f^{(l)}_\alpha)_{\alpha=1,\ldots,P} \sim \mathcal{N}(0, C^{(l)}_*)$ are Gaussian distributed with zero mean and covariance matrix $C^{(l)}_{\alpha\beta,*}$. Since the expectation values in (B2) depend on $h^{(l)}$, we would like to rewrite $\langle \cdots \rangle_p$ with respect to $h^{(l)}$. Since the signal $h^{(l)}$ decomposes into a sum of the residuals as $h^{(l)} = \sum_{k=1}^{l} f^{(l)}$, and the residuals $f^{(l)}$ are independent across l, and means and covariances for independent Gaussians sum up, the signal $h^{(l)} \sim \mathcal{N}(0, K^{(l)})$ is also Gaussian distributed with zero mean and covariance $K^{(l)} = \sum_{k=0}^{l} C^{(k)}_*$. Using the recursion relation $K^{(l)} = K^{(l-1)} + C^{(l-1)}_*$, we recover the NNGP result [8,11,12] for residual networks

$$C^{(l)}_{\alpha\beta,*} = \rho^2 \sigma_w^2 \left\langle \phi^{(l-1)}_\alpha \phi^{(l-1)}_\beta \right\rangle_{\mathcal{N}(0,K^{(l-1)})} + \rho^2 \sigma_b^2, \quad \text{for } 1 \leqslant l \leqslant L, \tag{B9}$$

$$K^{(l)}_{\alpha\beta} = \begin{cases} \frac{\sigma^2_{w,\text{in}}}{d_{\text{in}}} (XX^\mathsf{T})_{\alpha\beta} + \sigma^2_{b,\text{in}} & l = 0, \\ K^{(l-1)}_{\alpha\beta} + C^{(l-1)}_{\alpha\beta,*} & 1 \leqslant l \leqslant L, \\ \sigma^2_{w,\text{out}} \left\langle \phi^{(L)}_\alpha \phi^{(L)}_\beta \right\rangle_{\mathcal{N}(0,K^{(L)})} + \sigma^2_{b,\text{out}} & l = L+1. \end{cases} \tag{B10}$$

APPENDIX C: NEXT-TO-LEADING-ORDER CORRECTIONS

We compute the next-to-leading-order corrections to the saddle point in the previous section. While the auxiliary variables $C^{(l)}$ concentrate to the saddle point $C^{(l)}_*$ for infinite width $N \to \infty$, they fluctuate around this value for large but finite network width N. To lowest order, these fluctuations are Gaussian and can be obtained by computing the Hessian of the action $\mathcal{S}$ at the saddle point

$$p(Y|X) \simeq \int \mathcal{D}\delta C \int \mathcal{D}\delta\tilde{C} \exp\left(\frac{1}{2} (\delta C, \delta\tilde{C})^\mathsf{T} \mathcal{S}^{(2)} (\delta C, \delta\tilde{C}) \right) = \int \mathcal{D}\delta C \int \mathcal{D}\delta\tilde{C} \exp\left(-\frac{1}{2} (\delta C, \delta\tilde{C})^\mathsf{T} \mathcal{G}^{-1} (\delta C, \delta\tilde{C}) \right), \tag{C1}$$

where we write $\delta C = C - C^*$, $\delta\tilde{C} = \tilde{C} - \tilde{C}^*$ and the negative inverse of the Hessian corresponds to their covariance

$$\mathcal{G} \coloneqq -(\mathcal{S}^{(2)})^{-1} =: \begin{pmatrix} \langle \delta C\, \delta C \rangle & \langle \delta C\, \delta\tilde{C} \rangle \\ \langle \delta\tilde{C}\, \delta C \rangle & \langle \delta\tilde{C}\, \delta\tilde{C} \rangle \end{pmatrix} \tag{C2}$$

with $\mathcal{G}$ being the propagator. We evaluate all terms at the saddle point; thus there is no linear term in (C1) and all expectation values that appear in the following are with respect to the Gaussian measure $\langle \ldots \rangle_p$ in (B8).

The diagonal entries of the Hessian compute to

$$\frac{\partial^2}{\partial C^{(l)}_{\alpha\beta} \partial C^{(k)}_{\gamma\beta}} \mathcal{S}|_{(C_*,\tilde{C}_*)} = \frac{1}{4} \left\langle \tilde{h}^{(l)}_\alpha \tilde{h}^{(l)}_\beta, \tilde{h}^{(k)}_\gamma \tilde{h}^{(k)}_\delta \right\rangle^c_p = 0, \tag{C3}$$

$$\begin{aligned} \frac{\partial^2}{\partial \tilde{C}^{(l)}_{\alpha\beta} \partial \tilde{C}^{(k)}_{\gamma\delta}} \mathcal{S}|_{(C_*,\tilde{C}_*)} &= \frac{\partial}{\partial \tilde{C}^{(l)}_{\alpha\beta}} \left(N[\delta_{k,L+1} + (1 - \delta_{k,L+1})\rho^2]\, \sigma_w^2 \left\langle \phi^{(k-1)}_\gamma \phi^{(k-1)}_\delta \right\rangle_p + \text{const}(\tilde{C}) \right) \\ &= \delta_{L0} N \left[\sigma^2_{w,\text{in}} (XX^\mathsf{T})_{\alpha\beta} + d_{\text{in}} \right] [\delta_{k,L+1} + (1 - \delta_{k,L+1})\rho^2]\, \sigma_w^2 \left\langle \phi^{(k-1)}_\gamma \phi^{(k-1)}_\delta \right\rangle_p \\ &\quad + N \sigma_w^4\, 1_{l>0} 1_{k>0} \left\langle \phi^{(l-1)}_\alpha \phi^{(l-1)}_\beta, \phi^{(k-1)}_\gamma \phi^{(k-1)}_\delta \right\rangle^c_p \times \begin{cases} \rho^4 & k, l \neq L+1, \\ \rho^2 & k \neq l = L+1 \vee l \neq k = L+1, \\ 1 & \text{else}, \end{cases} \end{aligned} \tag{C4}$$

where $1_{l>0}$ denotes the indicator function. We write $\langle\cdots\rangle^c$ for connected correlations (cumulants) defined as

$$\langle z_\alpha z_\beta, z_\gamma z_\delta\rangle_p^c = \langle z_\alpha z_\beta z_\gamma z_\delta\rangle_p - \langle z_\alpha z_\beta\rangle_p \langle z_\gamma z_\delta\rangle_p. \tag{C5}$$

The average over the auxiliary variables $\tilde{h}$ in (C3), where we again use the notation $\tilde{y} = \tilde{h}^{(L+1)}$ for brevity, vanishes due to the normalization of the probability distribution as explained in detail in the previous section.

The off-diagonal terms compute to

$$\begin{aligned}
\frac{\partial^2}{\partial C^{(l)}_{\alpha\beta}\partial\tilde{C}^{(k)}_{\gamma\delta}}\mathcal{S}|_{(C_*,\tilde{C}_*)} &= -N\nu_l\delta_{kl} + N\,1_{k>0}\,\sigma_w^2\frac{\partial}{\partial C^{(l)}_{\alpha\beta}}\big\langle\phi^{(k-1)}_\gamma\phi^{(k-1)}_\delta\big\rangle_p \times \begin{cases}\rho^2 & k\leqslant L\\ 1 & k=L+1\end{cases}\\
&= -N\nu_l\delta_{kl} + N\,1_{k>0}\,\sigma_w^2\frac{\partial}{\partial K^{(k-1)}_{\alpha\beta}}\big\langle\phi^{(k-1)}_\gamma\phi^{(k-1)}_\delta\big\rangle_{\mathcal{N}(0,K^{(k-1)})}\frac{\partial}{\partial C^{(l)}_{\alpha\beta}}K^{(k-1)}_{\alpha\beta} \times \begin{cases}\rho^2 & k\leqslant L\\ 1 & k=L+1\end{cases}\\
&= -N\nu_l\delta_{kl} + N\,\delta_{(\alpha\beta),(\gamma\delta)}\,1_{k>0}\,\sigma_w^2\big\langle\big[\phi^{(k-1)}_\alpha\big]'\big[\phi^{(k-1)}_\beta\big]' + \delta_{\alpha\beta}\big[\phi^{(k-1)}_\alpha\big]''\big[\phi^{(k-1)}_\beta\big]\big\rangle_{\mathcal{N}(0,K^{(k-1)})}\,1_{k>l}\times\begin{cases}\rho^2 & k\leqslant L\\ 1 & k=L+1\end{cases},
\end{aligned} \tag{C6}$$

where we used Price's theorem [37,59]

$$\partial_{K_{\alpha\beta}}\langle\phi(h_\gamma)\phi(h_\delta)\rangle_{h\sim\mathcal{N}(0,K)} = \langle\partial_{h_\alpha}\partial_{h_\beta}\phi(h_\gamma)\phi(h_\delta)\rangle_{h\sim\mathcal{N}(0,K)} \tag{C7}$$

from the second to the third line (see Appendix D for details). The condition $k > l$ enforced by the indicator function $1_{k>l}$ results from the appearing factor $\partial_{C^{(l)}}K^{(k-1)}$ in the derivative, because the network kernel $K^{(k-1)}$ only depends on the residual kernels $C^{(l)}$ with $l < k$.

To calculate the negative inverse of the Hessian, we write

$$\mathcal{S}^{(2)} = \begin{pmatrix}\frac{\partial^2}{\partial C^2}\mathcal{S} & \frac{\partial^2}{\partial C\,\partial\tilde{C}}\mathcal{S}\\ \frac{\partial^2}{\partial\tilde{C}\,\partial C}\mathcal{S} & \frac{\partial^2}{\partial\tilde{C}^2}\mathcal{S}\end{pmatrix} =: \begin{pmatrix}\mathcal{S}_{11} & \mathcal{S}_{12}\\ \mathcal{S}_{21} & \mathcal{S}_{22}\end{pmatrix}. \tag{C8}$$

Using the block structure and the fact that $\mathcal{S}_{11} = 0$, we get

$$\mathcal{G}_{11} = \mathcal{G}_{12}\,\mathcal{S}_{22}\,\mathcal{G}_{21}, \tag{C9}$$

$$\mathcal{G}_{12} = -\mathcal{S}_{21}^{-1}, \tag{C10}$$

$$\mathcal{G}_{22} = 0. \tag{C11}$$

1. Response function

The general forward response function is given by

$$\Delta_{12}^{lm,(\alpha\beta),(\gamma\delta)} := N\langle C^{(l)}_{(\alpha\beta)}\tilde{C}^{(m)}_{(\gamma\delta)}\rangle = N\mathcal{G}_{12}^{(lm),(\alpha\beta),(\gamma\delta)}, \tag{C12}$$

which measures the response of the residual kernel $C^{(l)}_{(\alpha\beta)}$ in layer l to a perturbation in layer m mitigated by the conjugate variable $\tilde{C}^{(m)}_{(\gamma\delta)}$. We calculate it as the negative inverse of the off-diagonal block matrix $\mathcal{S}_{21}$. Due to the forward dependence of the kernels $K^{(l)}$, $\mathcal{S}_{21}$ is a lower triangular matrix and we get its inverse from forward propagation

$$\begin{aligned}
\mathcal{G}_{12}^{(lm),(\alpha\beta),(\gamma\delta)} &= N^{-1}\nu_l^{-1}\delta_{lm} + 1_{l>0}\,\delta_{(\alpha\beta),(\gamma\delta)}\,\sigma_w^2\big\langle\big[\phi^{(k-1)}_\alpha\big]'\big[\phi^{(k-1)}_\beta\big]'\\
&\quad + \delta_{\alpha\beta}\big[\phi^{(k-1)}_\alpha\big]''\big[\phi^{(k-1)}_\beta\big]\big\rangle_{\mathcal{N}(0,K^{(l-1)})}\sum_{k=0}^{l-1}\mathcal{G}_{12}^{(km),(\alpha\beta),(\gamma\delta)} \times \begin{cases}\rho^2 & k\leqslant L\\ 1 & k=L+1\end{cases}.
\end{aligned} \tag{C13}$$

Note that despite the double index $(\alpha\beta)$, $(\gamma\delta)$, the response is always a function of covariance entries $(\alpha\beta)$ due to the appearing $\delta_{(\alpha\beta),(\gamma\delta)}$ but depends on all other elements $(\gamma\delta)$ via the appearing expectation value.

We are ultimately interested in the response function with respect to the network input as a measure for network trainability. Since $\Delta_{12}^{(lm),(\alpha\beta),(\gamma\delta)}$ is with respect to the residual kernels $C^{(l)}$, we define the residual response for all intermediate network layers $1\leqslant l\leqslant L$ as

$$\eta^{(l)}_{\alpha\beta} := \Delta_{12}^{(l,0),(\alpha\beta),(\gamma\delta)} \tag{C14}$$

and obtain

$$\eta^{(l)}_{\alpha\beta} = \delta_{(\alpha\beta),(\gamma\delta)}\,\rho^2\sigma_w^2\big\langle\big[\phi^{(l-1)}_\alpha\big]'\big[\phi^{(l-1)}_\beta\big]' + \delta_{\alpha\beta}\big[\phi^{(l-1)}_\alpha\big]''\big[\phi^{(l-1)}_\beta\big]\big\rangle_{\mathcal{N}(0,K^{(l-1)})}\sum_{k=0}^{l-1}\eta^{(k)}_{\alpha\beta} \tag{C15}$$

with initial condition $\eta^{(0)}_{\alpha\beta} = N/d_{\text{in}}$. Due to their additive structure the response function of the kernels $K^{(l)}$ is given by

$$\chi^{(l)}_{\alpha\beta} \coloneqq \sum_{k=0}^{l} \eta^{(k)}_{\alpha\beta}. \tag{C16}$$

Finally, the output response $\chi^{\text{out}}_{\alpha\beta}$ is then given by

$$\chi^{\text{out}}_{\alpha\beta} = \sigma^2_{w,\,\text{out}} \left\langle \left[\phi^{(L)}_\alpha\right]' \left[\phi^{(L)}_\beta\right]' + \delta_{\alpha\beta} \left[\phi^{(L)}_\alpha\right]'' \left[\phi^{(L)}_\beta\right] \right\rangle_{\mathcal{N}(0,K^{(L)})} \sum_{k=0}^{L} \eta^{(k)}_{\alpha\beta}. \tag{C17}$$

2. Kernel fluctuations

Using (C9), we get for the diagonal term

$$\mathcal{G}^{(lm)}_{11} = \sum_{k,n} \mathcal{G}^{(lk)}_{12}\, \mathcal{S}^{(kn)}_{22}\, \mathcal{G}^{(nm)}_{21}. \tag{C18}$$

This quantity describes the Gaussian fluctuations of the residual kernels $C^{(l)}$ in networks of finite width N around the NNGP value:

$$C^{(l)} = C^{(l)}_* + \delta C^{(l)} \quad \text{with} \quad \delta C^{(l)} \sim \mathcal{N}\big(0, \mathcal{G}^{(ll)}_{11}\big). \tag{C19}$$

The fluctuations for the network kernels $K^{(l)}$ can be once again computed based on their additive nature as

$$K^{(l)} = \sum_{k=0}^{l} C^{(k)} = \sum_{k=0}^{l} C^{(k)}_* + \sum_{k=0}^{l} \delta C^{(k)}, \tag{C20}$$

yielding

$$\delta K^{(l)} \sim \mathcal{N}(0, \sum_{k=0}^{l} \mathcal{G}^{(kk)}_{11}).$$

Kernel fluctuations can be used to compute finite-width corrections to quantities such as the posterior kernels, generalization error, etc [29,31,60].

APPENDIX D: PRICE'S THEOREM

Consider an expectation value of $f : \mathbb{R}^N \to \mathbb{R}$ over centered jointly Gaussian distributed x_i with covariance C

$$\langle f(x) \rangle_{x \sim \mathcal{N}(0,C)}. \tag{D1}$$

We assume that $f(x)$ grows slower than $e^{x_i^2}$ for large x_i. Rewriting the Gaussian $\mathcal{N}(0, C)$ in terms of its Fourier transform $\mathcal{N}(0, C) = \left\{\prod_j \int_{-i\infty}^{i\infty} \frac{d\tilde{x}_j}{2\pi i}\right\} \exp\left(-x^\intercal \tilde{x} + \frac{1}{2}\tilde{x}^\intercal C \tilde{x}\right)$ one obtains

$$\langle f(x) \rangle_{x \sim \mathcal{N}(0,C)} = \prod_j \left\{ \int_{-\infty}^{\infty} dx_j \int_{-i\infty}^{i\infty} \frac{d\tilde{x}_j}{2\pi i} \right\} f(x) \exp\left(-x^\intercal \tilde{x} + \frac{1}{2}\tilde{x}^\intercal C \tilde{x}\right), \tag{D2}$$

which yields the property

$$\frac{\partial}{\partial C_{kl}} \langle f(x) \rangle_{x \sim \mathcal{N}(0,C)} = \prod_j \left\{ \int_{-\infty}^{\infty} dx_j \int_{-i\infty}^{i\infty} \frac{d\tilde{x}_j}{2\pi i} \right\} f(x) \frac{1}{2} \tilde{x}_k \tilde{x}_l \exp\left(-x^\intercal \tilde{x} + \frac{1}{2}\tilde{x}^\intercal C \tilde{x}\right). \tag{D3}$$

One notices that one may replace both occurrences of $\tilde{x}_i \to -\partial/\partial x_i$ under the integral: integrating by parts twice and using the assumption that f grows slower than $e^{x_i^2}$ for large x_i so that boundary terms vanish, one obtains

$$\frac{\partial}{\partial C_{kl}} \langle f(x) \rangle_{x \sim \mathcal{N}(0,C)} = \prod_j \left\{ \int_{-\infty}^{\infty} dx_j \int_{-i\infty}^{i\infty} \frac{d\tilde{x}_j}{2\pi i} \right\} \frac{1}{2} \left\{ \frac{\partial}{\partial x_k} \frac{\partial}{\partial x_l} f(x) \right\} \exp\left(x^\intercal \tilde{x} + \frac{1}{2}\tilde{x}^\intercal C \tilde{x}\right) = \frac{1}{2} \left\langle f^{(2)}_{kl} \right\rangle_{x \sim \mathcal{N}(0,C)}, \tag{D4}$$

where $f^{(2)}_{kl}$ is the Hessian of f. This expression is known as Price's theorem [37,59]. Note that sometimes the theorem is only stated for derivatives by $C_{k \neq l}$ only.

APPENDIX E: MAXIMUM ENTROPY CONDITION FOR OPTIMAL SCALING

We here derive an alternative condition for optimal signal variance, building on Bukva *et al.* [39] who proposed this method to study trainability in feed-forward networks. Their conjecture is that networks with signal distributions that are approximately uniform, or put differently maximally entropic, are more expressive.

For wide networks, the signal distribution of internal layers is approximately Gaussian

$$p(h;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}h^2\right), \tag{E1}$$

considering only a scalar component h here as all components h_i are independently and identically distributed.

We here focus on the readout layer. The distribution of the postactivation $y = \phi(h)$ is then

$$p(y;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}\phi'(\phi^{-1}(x))} \exp\left(-\frac{1}{2\sigma^2}\phi^{-1}(zy)\right). \tag{E2}$$

For $\phi = \mathrm{erf}$, the postactivation is approximately bounded by $y \in [-1, 1]$. Thus, we compute the Kullback-Leibler divergence between the distribution of the postactivation and a uniform distribution on that interval

$$\begin{aligned} D_{\mathrm{KL}}(p_{\mathrm{uni}}|p_\phi) &= \int_{-1}^{1} \mathrm{d}y\, p_{\mathrm{uni}}(y)\,[\ln p_{\mathrm{uni}}(y) - \ln p_\phi(y)] \qquad \text{(E3)} \\ &= \int_{-1}^{1} \mathrm{d}y\, \frac{1}{2}\ln\left(\frac{1}{2}\right) + \frac{1}{2}\frac{1}{2\sigma^2}\phi^{-1}(y) + \frac{1}{2}\ln(\sqrt{2\pi}\sigma\phi'[\phi^{-1}(y)]) \\ &= \ln\left(\frac{1}{2}\right) + \frac{1}{2}\int_{-1}^{1} \mathrm{d}y\, \frac{1}{2\sigma^2}\phi^{-1}(y)^2 + \ln\left(\sqrt{2\pi}\sigma\frac{2}{\sqrt{\pi}}\exp[-\phi^{-1}(y)^2]\right) \\ &= \ln\left(\frac{1}{2}\right) + \ln(\sqrt{8}\sigma) + \frac{1}{2}\int_{-1}^{1} \mathrm{d}y\left(\frac{1}{2\sigma^2} - 1\right)\phi^{-1}(y)^2 \\ &= \ln\left(\frac{\sqrt{8}}{2}\right) + \ln(\sigma) + \frac{1}{2}\left(\frac{1}{2\sigma^2} - 1\right)\int_{-\infty}^{\infty} \mathrm{d}h\,\phi^{-1}[\phi(h)]^2\,\phi'(h) \\ &= \ln(\sqrt{2}) + \frac{1}{2}\ln(\sigma^2) + \frac{1}{2}\left(\frac{1}{2\sigma^2} - 1\right)\int_{-\infty}^{\infty} \mathrm{d}h\,h^2\,\frac{2}{\sqrt{\pi}}\exp(-h^2) \\ &= \ln(\sqrt{2}) + \frac{1}{2}\ln(\sigma^2) + \frac{1}{2}\left(\frac{1}{2\sigma^2} - 1\right). \qquad \text{(E4)} \end{aligned}$$

Maximizing the Kullback-Leibler divergence amounts to

$$0 = \frac{\partial}{\partial\sigma^2} D_{\mathrm{KL}}(p_{\mathrm{uni}}|p_\phi) = \frac{1}{\sigma^2} - \frac{1}{4}\frac{1}{\sigma^4}, \tag{E5}$$

yielding as the condition for the signal variance before the readout layer $\sigma^2 = 1/4$. This condition is equivalent to the one in the Sec. III under the assumption that the dynamic range of the error function is given by $\mathcal{V} = 1$.

APPENDIX F: ADDITIONAL PLOTS

1. Decay of response function

Matching the observations by Yang and Schoenholz [38], the response function in ResNets decays subexponentially as shown in Fig. 8.

2. Input kernels for different tasks

In Sec. III we study the signal propagation and scaling behavior in residual networks for different tasks. In Fig. 9 we show the normalized overlap kernels $\frac{1}{\max_{\alpha\beta} x_\alpha \cdot x_\beta} X^\top X$ with P data samples for the studied tasks. For MNIST, we consider binary classification between 0 and 3 with equal number of samples from both classes $P_0 = P_3 = \frac{1}{2}P$. For CIFAR-10, we investigate binary classification between airplanes and dogs with equal number of samples from both classes $P_{\mathrm{airplane}} = P_{\mathrm{dog}} = \frac{1}{2}P$.

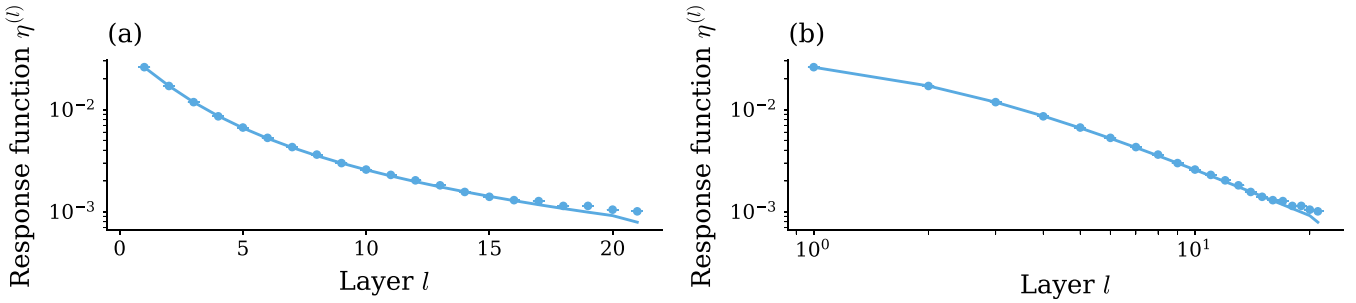

FIG. 8. Log-plot (a) and log-log-plot (b) of the response function $\eta^{(l)}$ for a residual network of depth $L = 20$. Dots represent simulations over 10^2 input samples and 10^3 network initializations, solid curves show theory values. (a) The decay of the response function is subexponential. (b) In later layers, the decay follows a power law. Other parameters: $\sigma^2_{w,\text{in}} = \sigma^2_w = \sigma^2_{w,\text{out}} = 1.2$, $\sigma^2_{b,\text{in}} = \sigma^2_b = \sigma^2_{b,\text{out}} = 0.2$, $d_{\text{in}} = d_{\text{out}} = 100$, $N = 500$, $\rho = 1$, $\phi = \text{erf}$.

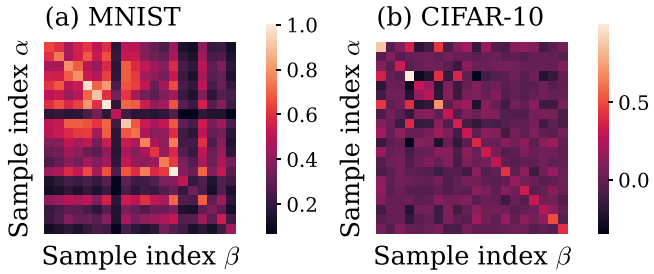

FIG. 9. Normalized overlap kernels of different tasks for $P = 20$ samples. Details on the tasks can be found in the text.

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2012), Vol. 25, pp. 1097–1105.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, Mastering the game of go with deep neural networks and tree search, Nature (London) **529**, 484 (2016).

[3] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA* (IEEE, Piscataway, NJ, 2016), pp. 770–778.

[4] K. He, X. Zhang, S. Ren, and J. Sun, Identity mappings in deep residual networks, in *European Conference on Computer Vision*, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling, Lecture Notes in Computer Science Vol. 9908 (Springer, Cham, 2016), pp. 630–645.

[5] A. Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, Technical Report, Department of Computer Science, University of Toronto, 2009.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in *Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence*, AAAI'17 (AAAI Press, Reston, VA, 2017), pp. 4278–4284.

[7] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, Proceedings of Machine Learning Research Vol. 37 (JMLR, Lille, France, 2015), pp. 448–456.

[8] K. Huang, Y. Wang, M. Tao, and T. Zhao, Why do deep residual networks generalize better than deep feedforward networks? A neural tangent kernel perspective, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Red Hook, NY, 2020), Vol. 33, pp. 2698–2709.

[9] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2018), Vol. 31, pp. 8580–8589.

[10] T. Bachlechner, B. P. Majumder, H. Mao, G. Cottrell, and J. McAuley, Rezero is all you need: Fast convergence at large depth, in *Proceedings of the Thirty-seventh Conference on Uncertainty in Artificial Intelligence*, edited by C. de Campos and M. H. Maathuis, Proceedings of Machine Learning Research Vol. 161 (JMLR, Cambridge, MA, 2021), pp. 1352–1361.

[11] T. Tirer, J. Bruna, and R. Giryes, Kernel-based smoothness analysis of residual networks, in *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, edited by J. Bruna, J. Hesthaven, and L. Zdeborova, Proceedings of Machine Learning Research Vol. 145 (JMLR, Cambridge, MA, 2022), pp. 921–954.

[12] D. Barzilai, A. Geifman, M. Galun, and R. Basri, A kernel perspective of skip connections in convolutional networks, in *The Eleventh International Conference on Learning Representations* (ICLR, 2023).

[13] J. Zhang, B. Han, L. Wynter, B. K. H. Low, and M. Kankanhalli, Towards robust ResNet: A small step but a giant leap, in *Proceedings of the Twenty-eighth International Joint Confer-*

ence on Artificial Intelligence, IJCAI-19 (International Joint Conferences on Artificial Intelligence Organization, 2019), pp. 4285–4291.

[14] D. Arpit, V. Campos, and Y. Bengio, How to initialize your network? Robust initialization for weightnorm & ResNets, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Red Hook, NY, 2019), Vol. 32.

[15] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau, Stable ResNet, in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, edited by A. Banerjee and K. Fukumizu, Proceedings of Machine Learning Research Vol. 130 (JMLR, Cambridge, MA, Cambridge, MA, 2021), pp. 1324–1332.

[16] S. Hayou, J.-F. Ton, A. Doucet, and Y. W. Teh, Robust pruning at initialization, in *International Conference on Learning Representations* (ICLR, 2021).

[17] H. Zhang, D. Yu, M. Yi, W. Chen, and T.-Y. Liu, Stabilize deep ResNet with a sharp scaling factor τ, Mach. Learn. **111**, 3359 (2022).

[18] B. Bordelon, L. Noci, M. B. Li, B. Hanin, and C. Pehlevan, Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit, in *The Twelfth International Conference on Learning Representations* (ICLR, 2024).

[19] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao, Tuning large neural networks via zero-shot hyperparameter transfer, in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Red Hook, NY, 2021).

[20] J. Halverson, A. Maiti, and K. Stoner, Neural networks and quantum field theory, Mach. Learn.: Sci. Technol. **2**, 035002 (2021).

[21] G. Naveh, O. Ben David, H. Sompolinsky, and Z. Ringel, Predicting the outputs of finite deep neural networks trained with noisy gradients, Phys. Rev. E **104**, 064301 (2021).

[22] K. Segadlo, B. Epping, A. van Meegen, D. Dahmen, M. Krämer, and M. Helias, Unified field theoretical approach to deep and recurrent neuronal networks, J. Stat. Mech. (2022) 103401.

[23] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, Cambridge, 2022).

[24] G. Naveh and Z. Ringel, A self consistent theory of gaussian processes captures feature learning effects in finite CNNS, in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Red Hook, NY, 2021).

[25] J. A. Zavatone-Veth and C. Pehlevan, Exact marginal prior distributions of finite Bayesian neural networks, in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Red Hook, NY, 2021).

[26] J. A. Zavatone-Veth, W. L. Tong, and C. Pehlevan, Contrasting random and learned features in deep Bayesian linear regression, Phys. Rev. E **105**, 064118 (2022).

[27] I. Seroussi, G. Naveh, and Z. Ringel, Separation of scales and a thermodynamic description of feature learning in some CNNs, Nat. Commun. **14**, 908 (2023).

[28] A. X. Yang, M. Robeyns, E. Milsom, B. Anson, N. Schoots, and L. Aitchison, A theory of representation learning gives a deep generalisation of kernel methods, in *Proceedings of the 40th International Conference on Machine Learning*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Proceedings of Machine Learning Research Vol. 202 (JMLR, Cambridge, MA, 2023), pp. 39380–39415.

[29] K. Fischer, J. Lindner, D. Dahmen, Z. Ringel, M. Krämer, and M. Helias, Critical feature learning in deep neural networks, in *Proceedings of the 41st International Conference on Machine Learning*, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Proceedings of Machine Learning Research Vol. 235 (JMLR, Cambridge, MA, 2024), pp. 13660–13690.

[30] N. Rubin, I. Seroussi, and Z. Ringel, Grokking as a first order phase transition in two layer networks, in *The Twelfth International Conference on Learning Representations* (ICLR, 2024).

[31] J. Lindner, D. Dahmen, M. Krämer, and M. Helias, A theory of data variability in neural network Bayesian inference, arXiv:2307.16695.

[32] B. Bordelon and C. Pehlevan, Self-consistent dynamical field theory of kernel evolution in wide neural networks, J. Stat. Mech. (2023) 114009.

[33] K. Fischer, D. Dahmen, and M. Helias, Field theory for optimal signal propagation in ResNets, Zenodo (2023), 10.5281/zenodo.17395886.

[34] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Clarendon Press, Oxford, 1996).

[35] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Westview Press Inc, 1991).

[36] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, *International Conference on Learning Representations* (OpenReview.net, 2017).

[37] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. (McGraw-Hill, Boston, 2002).

[38] G. Yang and S. Schoenholz, Mean field residual networks: On the edge of chaos, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017), Vol. 30.

[39] A. Bukva, J. de Gier, K. T. Grosvenor, R. Jefferson, K. Schalm, and E. Schwander, Criticality versus uniformity in deep neural networks, arXiv:2304.04784.

[40] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2019), Vol. 32.

[41] M. Welling and Y. W. Teh, Bayesian learning via stochastic gradient langevin dynamics, in *Proceedings of the 28th International Conference on International Conference on Machine Learning* (ICLR, 2011), pp. 681–688.

[42] S. Mandit, M. D. Hoffman, and D. M. Blei, Stochastic gradient descent as approximate Bayesian inference, J. Mach. Learn. Res. **18**, 1 (2017).

[43] S. Li, J. Jiao, Y. Han, and T. Weissman, Demystifying ResNet, arXiv:1611.01186.

[44] Y. Furusho and K. Ikeda, Resnet and batch-normalization improve data separability, in *Proceedings of the Eleventh Asian Conference on Machine Learning*, edited by W. S. Lee and T.

Suzuki, Proceedings of Machine Learning Research Vol. 101 (JMLR, 2019), pp. 94–108.

[45] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Red Hook, NY, 2016), pp. 3360–3368.

[46] B. Hanin, Which neural net architectures give rise to exploding and vanishing gradients?, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Red Hook, NY, 2018), Vol. 31.

[47] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Red Hook, NY, 2018), pp. 6571–6583.

[48] A.-S. Cohen, R. Cont, A. Rossier, and R. Xu, Scaling properties of deep residual networks, in *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, Proceedings of Machine Learning Research Vol. 139 (ICLR, 2021), pp. 2039–2048.

[49] R. Barboni, G. Peyré, and F.-X. Vialard, On global convergence of resnets: From finite to infinite width using linear parameterization, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Red Hook, NY, 2022), Vol. 35, pp. 16385–16397.

[50] R. Barboni, G. Peyré, and F.-X. Vialard, Understanding the training of infinitely deep and wide ResNets with conditional optimal transport, Comm. Pure Appl. Math. **78**, 2149 (2025).

[51] Y. Chen, F. Liu, Y. Lu, G. Chrysos, and V. Cevher, Generalization of scaled deep ResNets in the mean-field regime, in *The Twelfth International Conference on Learning Representations* (ICLR, 2024).

[52] M. B. Li and M. Nica, Differential equation scaling limits of shaped and unshaped neural networks, *Transactions on Machine Learning Research* (OpenReview.net, 2024).

[53] P. Marion, A. Fermanian, G. Biau, and J.-P. Vert, Scaling ResNets in the large-depth regime, J. Mach. Learn. Res. **26**, 1 (2025).

[54] D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams, The shattered gradients problem: If ResNets are the answer, then what is the question?, in *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh, Proceedings of Machine Learning Research Vol. 70 (JMLR, Cambridge, MA, 2017), pp. 342–350.

[55] A. Zaeemzadeh, N. Rahnavard, and M. Shah, Norm-preservation: Why residual networks can become extremely deep? IEEE Trans. Pattern Anal. Mach. Intell. **43**, 3980 (2021).

[56] Z. Ling and R. C. Qiu, Spectrum concentration in deep residual learning: A free probability approach, IEEE Access, **7**, 105212 (2019).

[57] X. Zhang, R. Jiang, W. Gao, R. Willett, and M. Maire, Residual connections harm abstract feature learning in masked autoencoders, arXiv:2404.10947.

[58] D. Doshi, T. He, and A. Gromov, Critical initialization of wide and deep neural networks using partial Jacobians: General theory and applications, in *Advances in Neural Information Processing Systems 36*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Curran Associates, Red Hook, NY, 2023), pp. 40054–40095.

[59] R. Price, A useful theorem for nonlinear devices having Gaussian inputs, IRE Trans. Inf. Theory **4**, 69 (1958).

[60] N. Rubin, K. Fischer, J. Lindner, I. Seroussi, Z. Ringel, M. Krämer, and M. Helias, From kernels to features: A multi-scale adaptive theory of feature learning, in *Proceedings of the 42nd International Conference on Machine Learning*, edited by A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Proceedings of Machine Learning Research Vol. 267 (PMLR, 2025), pp. 52225–52257.