Proceedings of Machine Learning Research vol 107:165–192, 2020

Non-Gaussian processes and neural networks at finite widths

Sho Yaida SHOYAIDA@FB.COM
Facebook AI Research
Facebook Inc.
Menlo Park, California 94025, USA

Abstract

Gaussian processes are ubiquitous in nature and engineering. A case in point is a class of neural networks in the infinite-width limit, whose priors correspond to Gaussian processes. Here we perturbatively extend this correspondence to finite-width neural networks, yielding non-Gaussian processes as priors. The methodology developed herein allows us to track the flow of preactivation distributions by progressively integrating out random variables from lower to higher layers, reminiscent of renormalization-group flow. We further develop a perturbative procedure to perform Bayesian inference with weakly non-Gaussian priors.

1. Inception

Gaussian processes model many phenomena in the physical world. A prime example is Brownian motion (Brown, 1828), modeled as the integral of Gaussian-distributed bumps exerted on a point-like solute (Einstein, 1905). The theory of elementary particles (Weinberg, 1995) also becomes a Gaussian process in the free limit where interactions between particles are turned off, and many-body systems as complex as glasses come to be Gaussian in the infinite-dimensional, mean-field, limit (Parisi and Zamponi, 2010). In the context of machine learning, Neal (1996) pointed out that a class of neural networks give rise to Gaussian processes in the infinite-width limit, which can perform exact Bayesian inference from training to test data (Williams, 1997). They occupy a corner of theoretical playground wherein the *karakuri* of neural networks is scrutinized (Lee et al., 2018; Matthews et al., 2018; Jacot et al., 2018; Chizat et al., 2018; Lee et al., 2019; Geiger et al., 2019).

In reality, Gaussian processes are but mere idealizations. Brownian particles have finite-size structure, elementary particles interact, and many-body systems respond nonlinearly. In order to understand rich phenomena exhibited by these real systems, Gaussian processes rather serve as starting points to be perturbed around. Indeed many edifices in theoretical physics are built upon the successful treatment of non-Gaussianity, with a notable example being renormalization-group flow (Kadanoff, 1966; Wilson, 1971; Weinberg, 1996; Goldenfeld, 2018). In the quest to elucidate behaviors of real neural networks away from the infinite-width limit, it is thus natural to wonder if the similar treatment of non-Gaussianity yields equally elegant and powerful machinery.

Here we set out on this program, perturbatively treating finite-width corrections to neural networks. Prior distributions of outputs are obtained through progressively integrating out preactivation of neurons layer by layer, yielding non-Gaussian priors. The whole procedure closely resembles renormalization-group flow (Goldenfeld, 2018; Mehta and Schwab, 2014): it bridges probability distributions at different scales through coarse-graining of random variables at microscopic scales; the flow of distributions is traced through running couplings, which in particular capture the de-

© 2020 S. Yaida.

gree of non-Gaussianity in these distributions; resulting recursive equations (R1,R2,R3) govern the evolution of these running couplings from lower to higher layers, just as renormalization-group equations do from microscopic to macroscopic scales. Such a recursive approach enables us to treat finite-width corrections to various observables, for networks with arbitrary activation functions.

The rest of the paper is structured as follows. In Section 2 we review and set up basic concepts. Our master recursive formulae (R1,R2,R3) are derived in Section 3, which control the flow of preactivation distributions. After an interlude with concrete examples in Section 4, we extend the Gaussian-process Bayesian inference to non-Gaussian priors in Section 5 and study inference of neural networks at finite widths. We conclude in Section 6 with dreams.

2. To infinity and beyond

In this paper we study real finite-width neural networks in the regime where the number of neurons in hidden layers is asymptotically large whereas input and output dimensions are kept constant.

2.1. Gaussian processes and neural networks at infinite widths

Let us focus on a class of neural networks termed multilayer perceptrons, with model parameters, $\boldsymbol{\theta} = \left\{ b_i^{(\ell)}, W_{ij}^{(\ell)} \right\}$, and an activation function, σ. For each input, $\boldsymbol{x} \in \mathbb{R}^{n_0}$, a neural network outputs a vector, $\boldsymbol{z}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{z}^{(L)} \in \mathbb{R}^{n_L}$, recursively defined as sequences of preactivations through

$$z_i^{(1)}(\boldsymbol{x}) = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1 \,, \tag{1}$$

$$z_i^{(\ell)}(\boldsymbol{x}) = b_i^{(\ell)} + \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \sigma\left[z_j^{(\ell-1)}(\boldsymbol{x}) \right] \quad \text{for} \quad i = 1, \ldots, n_\ell \,; \quad \ell = 2, \ldots, L \,. \tag{2}$$

Following Neal (1996), we assume priors for biases and weights given by independent and identically distributed Gaussian distributions with zero means, $\mathbb{E}\left[b_i^{(\ell)} \right] = \mathbb{E}\left[W_{ij}^{(\ell)} \right] = 0$, and variances

$$\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)} \right] = \delta_{i_1 i_2} C_b^{(\ell)} \,, \tag{3}$$

$$\mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)} \right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W^{(\ell)}}{n_{\ell-1}} \,. \tag{4}$$

Higher moments are then obtained by Wick's contractions (Wick, 1950; Zee, 2010). For instance,

$$\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)} b_{i_3}^{(\ell)} b_{i_4}^{(\ell)} \right] = \left[C_b^{(\ell)} \right]^2 \times \left(\delta_{i_1 i_2} \delta_{i_3 i_4} + \delta_{i_1 i_3} \delta_{i_2 i_4} + \delta_{i_1 i_4} \delta_{i_2 i_3} \right) \,. \tag{5}$$

For those unfamiliar with Wick's contractions and connected correlation functions (a.k.a. cumulants), a pedagogical review is provided in Appendix A as our formalism heavily relies on them.

In the infinite-width limit where $n_1, n_2, \ldots, n_{L-1} \to \infty$ (but finite n_0 and n_L), it has been argued – with varying degrees of rigor (Neal, 1996; Lee et al., 2018; Matthews et al., 2018) – that the prior distribution of outputs is governed by the Gaussian process with a kernel

$$K_{i_1 i_2; \alpha_1 \alpha_2} \equiv \mathbb{E}\left[z_{i_1}^{(L)}(\boldsymbol{x}_{\alpha_1}) z_{i_2}^{(L)}(\boldsymbol{x}_{\alpha_2}) \right] \tag{6}$$

and all the higher moments given by Wick's contractions. Here, the sample index α labels different inputs in a dataset. There exists a recursive formula that lets us evaluate this kernel for any pair of inputs (Lee et al., 2018) [c.f. Equation (R1)]. Importantly, once the values of the kernel are evaluated for all the pairs of $N_{\mathrm{D}} = N_{\mathrm{R}} + N_{\mathrm{E}}$ input data, $\{\boldsymbol{x}_\alpha\}_{\alpha=1,\ldots,N_{\mathrm{D}}}$, consisting of N_{R} training inputs with target outputs and N_{E} test inputs with unknown targets, we can perform exact Bayesian inference to yield mean outputs as predictions for N_{E} test data (Williams, 1997; Williams and Rasmussen, 2006) [c.f. Equation (GPM)]. This should be contrasted with stochastic gradient descent (SGD) optimization (Robbins and Monro, 1951), through which typically a single estimate for the optimal model parameters of the posterior, $\boldsymbol{\theta}_\star$, is obtained and used to predict outputs for test inputs; Bayesian inference instead marginalizes over all model parameters, performing an ensemble average over the posterior distribution (MacKay, 1995).

2.2. Beyond infinity

We shall now study real finite-width neural networks in the regime $n_1, \ldots, n_{L-1} \sim n \gg 1$.[1] At finite widths, there are corrections to Gaussian-process priors. In other words, a whole tower of nontrivial preactivation correlation functions beyond the kernel,

$$G^{(\ell)}_{i_1\ldots i_m;\alpha_1\ldots\alpha_m} \equiv \mathbb{E}\left[z^{(\ell)}_{i_1}(\boldsymbol{x}_{\alpha_1})\cdots z^{(\ell)}_{i_m}(\boldsymbol{x}_{\alpha_m})\right] , \tag{7}$$

collectively dictate the distribution of preactivations. Our aim is to trace the flow of these distributions progressively and cumulatively all the way up to the last layer whereat Bayesian inference is executed. More specifically, we shall inductively and self-consistently show that two-point preactivation correlation functions take the form[2]

$$G^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2} = \delta_{i_1 i_2}\left[\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2} + \frac{1}{n_{\ell-1}}\widetilde{S}^{(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n^2}\right)\right] \tag{KS}$$

and connected four-point preactivation correlation functions

$$\begin{aligned} &G^{(\ell)}_{i_1 i_2 i_3 i_4;\alpha_1\alpha_2\alpha_3\alpha_4}\Big|_{\text{connected}} \\ \equiv &G^{(\ell)}_{i_1 i_2 i_3 i_4;\alpha_1\alpha_2\alpha_3\alpha_4} - G^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2}G^{(\ell)}_{i_3 i_4;\alpha_3\alpha_4} - G^{(\ell)}_{i_1 i_3;\alpha_1\alpha_3}G^{(\ell)}_{i_2 i_4;\alpha_2\alpha_4} - G^{(\ell)}_{i_1 i_4;\alpha_1\alpha_4}G^{(\ell)}_{i_2 i_3;\alpha_2\alpha_3} \\ = &\frac{1}{n_{\ell-1}}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3,\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4}\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} + \delta_{i_1 i_4}\delta_{i_2 i_3}\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}\right] + O\left(\frac{1}{n^2}\right) , \end{aligned} \tag{V}$$

and higher cumulants are all suppressed by $O\left(\frac{1}{n^2}\right)$.[3] Here the Gaussian-process core kernel $\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}$ and the self-energy correction $\widetilde{S}^{(\ell)}_{\alpha_1\alpha_2}$ are symmetric under the exchange of sample indices $\alpha_1 \leftrightarrow \alpha_2$ and the four-point vertex $\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ is symmetric under $\alpha_1 \leftrightarrow \alpha_2$, $\alpha_3 \leftrightarrow \alpha_4$, and $(\alpha_1\alpha_2) \leftrightarrow$

1. Note that input and output dimensions, n_0 and n_L, are arbitrary. To be precise, defining $n_1, \ldots, n_{L-1} \equiv r_1 n, \ldots, r_{L-1} n$, we send $n \gg 1$ while keeping $\left\{C_b^{(\ell)}, C_W^{(\ell)}\right\}_{\ell=1,\ldots,L}$, $r_1, \ldots, r_{L-1}$, n_0, and n_L constants, and compute the leading $1/n$ corrections. In particular it is crucial to keep the number of outputs n_L constant in order to consistently perform Bayesian inference within our approach.
2. In the main text we place tildes on objects that depend only on sample indices α's in order to distinguish them from those that depend both on sample indices α's and neuron indices i's.
3. Given that the means of biases and weights are zero, $G^{(\ell)}_{i_1\ldots i_m;\alpha_1\ldots\alpha_m} = 0$ for all odd m.

$(\alpha_3\alpha_4)$. At the first layer the preactivation distribution is exactly Gaussian for any finite widths and hence Equations (KS) and (V) are trivially satisfied, with

$$\widetilde{K}^{(1)}_{\alpha_1\alpha_2} = C_b^{(1)} + C_W^{(1)} \cdot \left(\frac{\boldsymbol{x}_{\alpha_1} \cdot \boldsymbol{x}_{\alpha_2}}{n_0} \right) , \quad \widetilde{S}^{(1)}_{\alpha_1\alpha_2} = 0 , \quad \text{and} \quad \widetilde{V}^{(1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = 0 . \tag{R0}$$

Obtained in Section 3 are the recursive formulae that link these core kernel, self-energy, and four-point vertex at the ℓ-th layer to those at the $(\ell+1)$-th layer while in Section 5 these tensors at the last layer $\ell = L$ are used to yield the leading $1/n$ correction for Bayesian inference at finite widths.

2.3. Related work

Our Schwinger operator approach is orthogonal to the replica approach by Cohen et al. (2019) and, unlike the planar diagrammatic approach by Dyer and Gur-Ari (2019), applies to general activation functions, made possible by accumulating corrections layer by layer rather than dealing with them all at once. See also Antognini (2019). More substantially, in contrast to these previous approaches, we here study finite-width effects on Bayesian inference and find that the renormalization-group picture naturally emerges, with layers playing the role of scales.

3. Distributional flow

As auxiliary objects in recursive steps, let us introduce activation correlation functions

$$H^{(\ell)}_{i_1 \ldots i_m; \alpha_1 \ldots \alpha_m} \equiv \mathbb{E}\left\{ \sigma\left[z^{(\ell)}_{i_1}(\boldsymbol{x}_{\alpha_1}) \right] \cdots \sigma\left[z^{(\ell)}_{i_m}(\boldsymbol{x}_{\alpha_m}) \right] \right\} . \tag{8}$$

Our basic strategy is to establish relations

$$\left\{ \boldsymbol{G}^{(1)} \right\} \to \left\{ \boldsymbol{H}^{(1)} \right\} \to \left\{ \boldsymbol{G}^{(2)} \right\} \to \cdots \to \left\{ \boldsymbol{H}^{(L-1)} \right\} \to \left\{ \boldsymbol{G}^{(L)} \right\} , \tag{ZIGZAG}$$

zigzagging between sets of preactivation correlation functions and sets of activation correlation functions, keeping track of leading finite-width corrections. Below, relations $\boldsymbol{G}^{(\ell)} \to \boldsymbol{H}^{(\ell)}$ are obtained by integrating out preactivations while relations $\boldsymbol{H}^{(\ell)} \to \boldsymbol{G}^{(\ell+1)}$ are obtained by integrating out biases and weights. At first glance the algebra in this paper may look horrifying but repeated applications of Wick's contractions are all there is to it. The results are summarized in Section 3.2.

3.1. Zigzag relations for preactivation and activation correlation functions

Integrating over the Gaussian biases and weights at ℓ's connections yield the relations that link activation correlations $\boldsymbol{H}^{(\ell)}$ to preactivation correlations $\boldsymbol{G}^{(\ell+1)}$ at the next layer. Recalling Equations (KS) and (V), trivial Wick's contractions yield

$$\widetilde{K}^{(\ell+1)}_{\alpha_1\alpha_2} + \frac{1}{n_\ell} \widetilde{S}^{(\ell+1)}_{\alpha_1\alpha_2} = C_b^{(\ell+1)} + C_W^{(\ell+1)} \left[\frac{1}{n_\ell} \sum_{j=1}^{n_\ell} H^{(\ell)}_{jj;\alpha_1\alpha_2} \right] + O\left(\frac{1}{n^2} \right) \quad \text{and} \tag{9}$$

$$\widetilde{V}^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \frac{\left[C_W^{(\ell+1)} \right]^2}{n_\ell} \sum_{j,k=1}^{n_\ell} \left[H^{(\ell)}_{jjkk;\alpha_1\alpha_2\alpha_3\alpha_4} - H^{(\ell)}_{jj;\alpha_1\alpha_2} H^{(\ell)}_{kk;\alpha_3\alpha_4} \right] + O\left(\frac{1}{n} \right) . \tag{10}$$

The remaining task is to relate preactivation correlations $\boldsymbol{G}^{(\ell)}$ to activation correlations $\boldsymbol{H}^{(\ell)}$ within the same layer, which will complete the zigzag relation (ZIGZAG) for these correlation functions.[4]

With the mastery of Wick's contractions and connected correlation functions, it is simple to derive the following combinatorial hack (Appendix A.4): viewing prior preactivations

$$\mathbf{z} \equiv \left\{ \mathrm{z}_{i;\alpha} \equiv z_i^{(\ell)}\left(\boldsymbol{x}_\alpha\right) \right\}_{i=1,\ldots,n_\ell;\alpha=1,\ldots,N_\mathrm{D}}$$

at the ℓ-th layer as a random $(n_\ell N_\mathrm{D})$-dimensional vector and defining the Gaussian integral with the kernel $\left\langle \mathrm{z}_{i_1;\alpha_1}\mathrm{z}_{i_2;\alpha_2}\right\rangle_{K^{(\ell)}} = K^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2} \equiv \delta_{i_1 i_2}\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}$, the prior average

$$\mathbb{E}\left\{\mathcal{F}[\mathbf{z}]\right\} = \left\langle \mathcal{F}[\mathbf{z}]\right\rangle_{K^{(\ell)}} + \frac{1}{n_{\ell-1}}\left[\left\langle \mathcal{F}[\mathbf{z}]\mathcal{O}_S[\mathbf{z}] + \mathcal{F}[\mathbf{z}]\mathcal{O}_V[\mathbf{z}]\right\rangle_{K^{(\ell)}}\right] + O\left(\frac{1}{n^2}\right) \tag{HACK}$$

for any function $\mathcal{F}$. Here the operators $\mathcal{O}_S[\mathbf{z}]$ and $\mathcal{O}_V[\mathbf{z}]$ capture $1/n$ corrections due to self-energy and four-point vertex, respectively, and are defined as

$$\mathcal{O}_S[\mathbf{z}] \equiv \frac{1}{2}\sum_{\alpha_1,\alpha_2} \widetilde{S}^{\alpha_1\alpha_2}_{(\ell)}\left[\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right) - n_\ell \widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}\right] \quad \text{and} \tag{OS}$$

$$\begin{aligned}\mathcal{O}_V[\mathbf{z}] \equiv \frac{1}{8}\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4} \widetilde{V}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} \\ \times\left[\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right)\left(\sum_{j=1}^{n_\ell}\mathrm{z}_{j;\alpha_3}\mathrm{z}_{j;\alpha_4}\right) - 2n_\ell\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right)\widetilde{K}^{(\ell)}_{\alpha_3\alpha_4}\right. \\ \left. - 4\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_3}\right)\widetilde{K}^{(\ell)}_{\alpha_2\alpha_4} + n_\ell^2\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}\widetilde{K}^{(\ell)}_{\alpha_3\alpha_4} + 2n_\ell\widetilde{K}^{(\ell)}_{\alpha_1\alpha_3}\widetilde{K}^{(\ell)}_{\alpha_2\alpha_4}\right],\end{aligned} \tag{OV}$$

where the sample indices are raised by using the inverse core kernel as a metric, meaning

$$\widetilde{S}^{\alpha_1\alpha_2}_{(\ell)} \equiv \sum_{\alpha_1',\alpha_2'}\left(\widetilde{K}^{-1}_{(\ell)}\right)^{\alpha_1\alpha_1'}\left(\widetilde{K}^{-1}_{(\ell)}\right)^{\alpha_2\alpha_2'}\widetilde{S}^{(\ell)}_{\alpha_1'\alpha_2'} \quad \text{and} \tag{11}$$

$$\widetilde{V}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} \equiv \sum_{\alpha_1',\ldots,\alpha_4'}\left(\widetilde{K}^{-1}_{(\ell)}\right)^{\alpha_1\alpha_1'}\cdots\left(\widetilde{K}^{-1}_{(\ell)}\right)^{\alpha_4\alpha_4'}\widetilde{V}^{(\ell)}_{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}\,. \tag{12}$$

Using the above hack, we can evaluate the activation correlations by straightforward algebra with Wick's contractions. In particular, as the Gaussian integral is diagonal in the neuron index i, we just need to disentangle cases with repeated and unrepeated neuron indices. The solution for this exercise is in Appendix B: it is arguably the most cumbersome algebra in this paper.

4. The nontrivial parts of the inductive proof for Equations (KS) and (V) are to show (i) that the right-hand side of Equation (10) is finite as $n \to \infty$, (ii) that the leading contribution of Equation (9) is the Gaussian-process kernel, and (iii) that higher-point connected preactivation correlation functions are all suppressed by $O\left(\frac{1}{n^2}\right)$, all of which are verified in obtaining the recursive equations. See Appendix B for a full proof.

3.2. Master recursive flow equations

Denoting the Gaussian integral with the core kernel $\langle \tilde{z}_{\alpha_1} \tilde{z}_{\alpha_2} \rangle_{\widetilde{K}^{(\ell)}} = \widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}$ for a single-neuron random vector $\tilde{\mathbf{z}} \equiv \{\tilde{z}_\alpha\}_{\alpha=1,\ldots,n_{\mathrm{D}}}$, and plugging in results of Appendix B into Equations (9) and (10), we arrive at our master recursion relations

$$\widetilde{K}^{(\ell+1)}_{\alpha_1\alpha_2} = C_b^{(\ell+1)} + C_W^{(\ell+1)} \left\langle \sigma(\tilde{z}_{\alpha_1}) \sigma(\tilde{z}_{\alpha_2}) \right\rangle_{\widetilde{K}^{(\ell)}} , \tag{R1}$$

$$\begin{aligned} \widetilde{V}^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \left[C_W^{(\ell+1)}\right]^2 \Bigg[& \left\langle \sigma(\tilde{z}_{\alpha_1}) \sigma(\tilde{z}_{\alpha_2}) \sigma(\tilde{z}_{\alpha_3}) \sigma(\tilde{z}_{\alpha_4}) \right\rangle_{\widetilde{K}^{(\ell)}} \\ & - \left\langle \sigma(\tilde{z}_{\alpha_1}) \sigma(\tilde{z}_{\alpha_2}) \right\rangle_{\widetilde{K}^{(\ell)}} \left\langle \sigma(\tilde{z}_{\alpha_3}) \sigma(\tilde{z}_{\alpha_4}) \right\rangle_{\widetilde{K}^{(\ell)}} \\ & + \frac{1}{4} \left(\frac{n_\ell}{n_{\ell-1}}\right) \sum_{\alpha_1',\alpha_2',\alpha_3',\alpha_4'} \widetilde{V}^{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}_{(\ell)} \left\langle \sigma(\tilde{z}_{\alpha_1}) \sigma(\tilde{z}_{\alpha_2}) (\tilde{z}_{\alpha_1'} \tilde{z}_{\alpha_2'} - \widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'}) \right\rangle_{\widetilde{K}^{(\ell)}} \\ & \qquad \times \left\langle \sigma(\tilde{z}_{\alpha_3}) \sigma(\tilde{z}_{\alpha_4}) (\tilde{z}_{\alpha_3'} \tilde{z}_{\alpha_4'} - \widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'}) \right\rangle_{\widetilde{K}^{(\ell)}} \Bigg] , \quad \text{and} \end{aligned} \tag{R2}$$

$$\begin{aligned} \widetilde{S}^{(\ell+1)}_{\alpha_1\alpha_2} = \left(\frac{n_\ell}{n_{\ell-1}}\right) C_W^{(\ell+1)} \Bigg[& \frac{1}{2} \sum_{\alpha_1',\alpha_2'} \widetilde{S}^{\alpha_1'\alpha_2'}_{(\ell)} \left\langle \sigma(\tilde{z}_{\alpha_1}) \sigma(\tilde{z}_{\alpha_2}) (\tilde{z}_{\alpha_1'} \tilde{z}_{\alpha_2'} - \widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'}) \right\rangle_{\widetilde{K}^{(\ell)}} \\ & + \frac{1}{8} \sum_{\alpha_1',\alpha_2',\alpha_3',\alpha_4'} \widetilde{V}^{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}_{(\ell)} \Big\langle \sigma(\tilde{z}_{\alpha_1}) \sigma(\tilde{z}_{\alpha_2}) \\ & \quad \times \Big(\tilde{z}_{\alpha_1'} \tilde{z}_{\alpha_2'} \tilde{z}_{\alpha_3'} \tilde{z}_{\alpha_4'} - 2 \tilde{z}_{\alpha_1'} \tilde{z}_{\alpha_2'} \widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'} - 4 \tilde{z}_{\alpha_1'} \tilde{z}_{\alpha_3'} \widetilde{K}^{(\ell)}_{\alpha_2'\alpha_4'} \\ & \qquad + \widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'} \widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'} + 2 \widetilde{K}^{(\ell)}_{\alpha_1'\alpha_3'} \widetilde{K}^{(\ell)}_{\alpha_2'\alpha_4'} \Big) \Big\rangle_{\widetilde{K}^{(\ell)}} \Bigg] . \end{aligned} \tag{R3}$$

For $\ell = 1$, a special note about the ratio $\frac{n_\ell}{n_{\ell-1}}$ is in order: even though n_0 stays constant while $n_1 \gg 1$, the terms proportional to that ratio are identically zero due to the complete Gaussianity (R0).

The preactivation distribution in the first layer (R0) sets the initial condition for the flow from lower to higher layers dictated by these recursive equations. Evolving through these recursive equations, the running couplings – $\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}$, $\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$, and $\widetilde{S}^{(\ell)}_{\alpha_1\alpha_2}$ – then trace changes in the distributions of preactivations as the layer scale ℓ shifts, just as running couplings for physical systems track changes in effective Boltzmann distributions as the probing scale shifts. Once recursed up to the last layer $\ell = L$, the resulting distribution of outputs $\mathbf{z} = \mathbf{z}^{(L)}$ can be succinctly encoded by the probability distribution

$$p[\mathbf{z}] = \frac{e^{-\mathcal{H}[\mathbf{z}]}}{\int \mathrm{d}\mathbf{z}' e^{-\mathcal{H}[\mathbf{z}']}} \tag{D0}$$

with the potential $\mathcal{H}[\mathbf{z}] = \mathcal{H}_0[\mathbf{z}] + \epsilon\mathcal{H}_1[\mathbf{z}] + O(\epsilon^2)$ where $\epsilon \equiv \frac{1}{n_{L-1}} \ll 1$,

$$\mathcal{H}_0[\mathbf{z}] = \frac{1}{2}\sum_{\alpha_1,\alpha_2}\left(\widetilde{K}^{-1}_{(L)}\right)^{\alpha_1\alpha_2}\left(\sum_{i=1}^{n_L}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right), \quad \text{and} \tag{D1}$$

$$\mathcal{H}_1[\mathbf{z}] = -\frac{1}{2}\sum_{\alpha_1,\alpha_2}\widetilde{J}^{\alpha_1\alpha_2}\left(\sum_{i=1}^{n_L}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right) \tag{D2}$$

$$-\frac{1}{8}\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}\widetilde{V}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(L)}\left(\sum_{i=1}^{n_L}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right)\left(\sum_{j=1}^{n_L}\mathrm{z}_{j;\alpha_3}\mathrm{z}_{j;\alpha_4}\right) \quad \text{with}$$

$$\widetilde{J}^{\alpha_1\alpha_2} \equiv \widetilde{S}^{\alpha_1\alpha_2}_{(L)} - \sum_{\alpha_3,\alpha_4}\widetilde{K}^{(L)}_{\alpha_3\alpha_4}\left[\widetilde{V}^{(\alpha_1\alpha_3)(\alpha_2\alpha_4)}_{(L)} + \frac{n_L}{2}\widetilde{V}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(L)}\right]. \tag{13}$$

Again, this can be derived through Wick's contractions. It is important to note that n_L is constant and thus $\epsilon\mathcal{H}_1[\mathbf{z}]$ can consistently be treated perturbatively.[5]

4. Interlude: examples

The recursive relations obtained above can be evaluated numerically (Lee et al., 2018) [or sometimes analytically for rectified linear unit (ReLU) activation (Cho and Saul, 2009)], which is a perfectly adequate approach: at the leading order it involves four-dimensional Gaussian integrals at most. Here, continuing the theme of wearing out Wick's contractions, we develop an alternative analytic method that works for any polynomial activations (Liao and Poggio, 2017), providing another perfectly cromulent approach.

For a general polynomial activation of degree p, $\sigma(z) = \sum_{k=0}^{p} a_k z^k$, the nontrivial term in Equation (R1) can be expanded as

$$\langle\sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\rangle_{\widetilde{K}^{(\ell)}} = \sum_{k_1,k_2=0}^{p} a_{k_1}a_{k_2}\left\langle\left(\tilde{\mathrm{z}}_{\alpha_1}\right)^{k_1}\left(\tilde{\mathrm{z}}_{\alpha_2}\right)^{k_2}\right\rangle_{\widetilde{K}^{(\ell)}}. \tag{14}$$

Each term can then be evaluated by Wick's contractions and the same goes for all the terms in Equations (R2) and (R3).[6] Below and in Appendix C, we illustrate this procedure with simple examples.

4.1. Deep linear networks

When the activation function is linear, $\sigma(z) = z$, multilayer perceptrons are called deep linear networks (Saxe et al., 2013). Setting $C_b^{(\ell)} = 0$ and $C_W^{(\ell)} = 1$ for simplicity, our recursion relations reduce to $\widetilde{K}^{(\ell+1)}_{\alpha_1\alpha_2} = \widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}$,

$$\widetilde{V}^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \left[\widetilde{K}^{(\ell)}_{\alpha_1\alpha_3}\widetilde{K}^{(\ell)}_{\alpha_2\alpha_4} + \widetilde{K}^{(\ell)}_{\alpha_1\alpha_4}\widetilde{K}^{(\ell)}_{\alpha_2\alpha_3} + \left(\frac{n_\ell}{n_{\ell-1}}\right)\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}\right],$$

5. If n_L were of order $n \gg 1$, the potential $\mathcal{H}$ would become a large-n vector model, for which we would have to sum the infinite series of bubble diagrams (Moshe and Zinn-Justin, 2003).
6. The same approach could be adopted for an analytic function but it would in general be difficult to sum the resulting infinite series in a closed form. It could nonetheless be useful in, for example, proving convergence properties.

and $\widetilde{S}^{(\ell+1)}_{\alpha_1\alpha_2} = \left(\frac{n_\ell}{n_{\ell-1}}\right) \widetilde{S}^{(\ell)}_{\alpha_1\alpha_2}$. Solving them yields the layer-independent core kernel and zero self-energy

$$\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2} = \widetilde{K}^{(1)}_{\alpha_1\alpha_2} = \frac{\boldsymbol{x}_{\alpha_1} \cdot \boldsymbol{x}_{\alpha_2}}{n_0} \quad \text{and} \quad \widetilde{S}^{(\ell)}_{\alpha_1\alpha_2} = 0 \tag{15}$$

and the linearly layer-dependent four-point vertex

$$\frac{1}{n_{\ell-1}} \widetilde{V}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \left(\sum_{\ell'=1}^{\ell-1} \frac{1}{n_{\ell'}}\right) \left[\widetilde{K}^{(1)}_{\alpha_1\alpha_3} \widetilde{K}^{(1)}_{\alpha_2\alpha_4} + \widetilde{K}^{(1)}_{\alpha_1\alpha_4} \widetilde{K}^{(1)}_{\alpha_2\alpha_3}\right] . \tag{16}$$

It succinctly reproduces the result that can be obtained through planar diagrams in this special setup (Dyer and Gur-Ari, 2019). Quadratic activation (Li et al., 2018) is worked out in Appendix C.1.

4.2. ReLU with single input

The recursion relations simplify drastically for the case of a single input, $N_{\rm D} = 1$, as worked out in detail in Appendix C.2. For instance, for ReLU activation with $C_b^{(\ell)} = 0$ and $C_W^{(\ell)} = 2$, we obtain the layer-independent core kernel, zero self-energy, and the four-point vertex

$$\frac{1}{n_{\ell-1}} \widetilde{V}^{(\ell)}_{(\alpha\alpha)(\alpha\alpha)} = 5 \left(\sum_{\ell'=1}^{\ell-1} \frac{1}{n_{\ell'}}\right) \left(\widetilde{K}^{(1)}_{\alpha\alpha}\right)^2 . \tag{17}$$

Interestingly, as for deep linear networks, the factor $\sum_{\ell'}(1/n_{\ell'})$ appears again. This factor has also been found by Hanin and Rolnick (2018), which provides guidance for network architectural design through its minimization. We generalize this factor for monomial activations in Appendix C.2.1

4.3. Experimental verification: output distributions for a single input

Here we put our theory to the test. For concreteness, take a single black-white image of hand-written digits with 28-by-28 pixels (i.e. $n_0 = 784$) from the MNIST dataset (LeCun et al., 1998) without preprocessing, set depth $L = 3$, bias variance $C_b^{(\ell)} = 0$, weight variance $C_W^{(\ell)} = C_W$, and widths $(n_0, n_1, n_2, n_3) = (784, n, 2n, 1)$, and use activations $\sigma(z) = z$ (linear) with $C_W = 1$ and $\max(0, z)$ (ReLU) with $C_W = 2$. In Figure 1, for each width-parameter n of the hidden layers we record the prior distribution of outputs over 10^6 instances of Gaussian weights and compare it with the theoretical prediction – obtained by cranking the knob from the initial condition (R0) through the recursion relations (R1-R3) to the distribution (D0-D2). The prior distribution becomes increasingly non-Gaussian as networks narrow and the deviation from the Gaussian-process prior is correctly captured by our theory. Higher-order perturbative calculations are expected to systematically improve the quality – and extend the range – of the agreement. Additional experiments are performed in Appendix C.3, which further corroborates our theory.

5. Bayesian inference

Let us take off from the terminal point of Section 3: we have obtained the recursive equations (R0-R3) for the Gaussian-process kernel and the leading finite-width corrections and codified them in the weakly non-Gaussian prior distributions $p[\mathbf{z}]$ (D0-D2) of outputs

$$\mathbf{z} \equiv \left\{ \mathbf{z}_{i;\alpha} \equiv z_i^{(L)}(\boldsymbol{x}_\alpha) \right\}_{i=1,\ldots,n_L;\alpha=1,\ldots,N_{\rm D}} ,$$

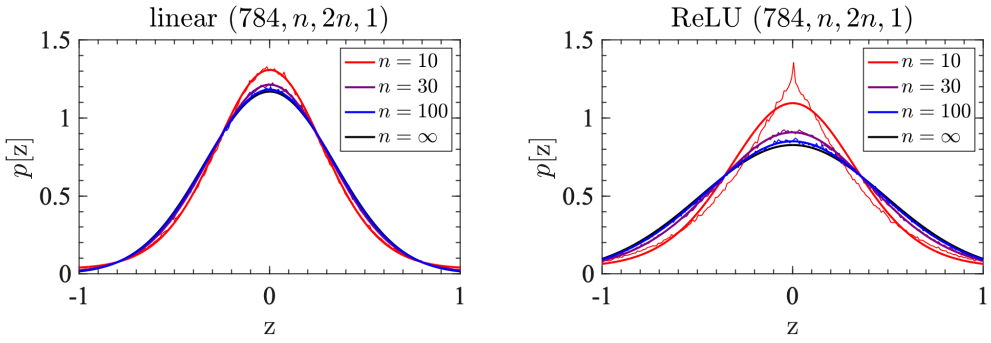

Figure 1: Comparison between theory and experiments for prior distributions of outputs for a single input. The agreement between our theoretical predictions (smooth thick lines) and experimental data (rugged thin lines) is superb, correctly capturing the initial deviations from Gaussian processes at $n = \infty$ (black), all the way down to $n \sim 10$ for linear activation and to $n \sim 30$ for ReLU activation.

dictated by the potential $\mathcal{H}[\mathbf{z}] = \mathcal{H}_0[\mathbf{z}] + \epsilon\mathcal{H}_1[\mathbf{z}] + O(\epsilon^2)$ with $\epsilon \equiv \frac{1}{n_{L-1}} \ll 1$. Examples in Section 4 illustrate that finite-width corrections stay perturbative typically when $\frac{\text{depth}}{\text{width}} \ll 1$. Let us now divide N_{D} inputs into N_{R} training and N_{E} test inputs as

$$\{\boldsymbol{x}_\alpha\}_{\alpha=1,\ldots,N_{\text{D}}} = \{(\boldsymbol{x}_{\text{R}})_{\bar{\beta}}\}_{\bar{\beta}=1,\ldots,N_{\text{R}}} \cup \{(\boldsymbol{x}_{\text{E}})_{\dot{\gamma}}\}_{\dot{\gamma}=1,\ldots,N_{\text{E}}}\,, \tag{18}$$

and the training inputs come with target outputs

$$\{(\boldsymbol{y}_{\text{R}})_{\bar{\beta}}\}_{\bar{\beta}=1,\ldots,N_{\text{R}}} = \{(y_{\text{R}})_{i;\bar{\beta}}\}_{\bar{\beta}=1,\ldots,N_{\text{R}};i=1,\ldots,n_L}\,. \tag{19}$$

We shall develop a procedure to infer outputs for test inputs a lá Bayes, perturbatively extending the textbook by Williams and Rasmussen (2006). For field theorists, our calculation is just a background-field calculation (Weinberg, 1996) in disguise.

Taking the liberty of notations, we let the number of input-data arguments dictate the summation over sample indices α inside the potential $\mathcal{H}$, and denote the joint probabilities

$$p[\mathbf{z}_{\text{R}}] = \frac{e^{-\mathcal{H}[\mathbf{z}_{\text{R}}]}}{\int \mathrm{d}\mathbf{z}'_{\text{R}} e^{-\mathcal{H}[\mathbf{z}'_{\text{R}}]}} \quad \text{and} \quad p[\mathbf{z}_{\text{R}}, \mathbf{z}_{\text{E}}] = \frac{e^{-\mathcal{H}[\mathbf{z}_{\text{R}}, \mathbf{z}_{\text{E}}]}}{\int \mathrm{d}\mathbf{z}'_{\text{R}} \mathrm{d}\mathbf{z}'_{\text{E}} e^{-\mathcal{H}[\mathbf{z}'_{\text{R}}, \mathbf{z}'_{\text{E}}]}}\,. \tag{20}$$

Given the training targets $\boldsymbol{y}_{\text{R}}$, the posterior distribution of test outputs are given by Bayes' rule:

$$p\left[\mathbf{z}_{\text{E}}|\boldsymbol{y}_{\text{R}}\right] = \frac{p[\boldsymbol{y}_{\text{R}}, \mathbf{z}_{\text{E}}]}{p[\boldsymbol{y}_{\text{R}}]} = \left(\frac{\int \mathrm{d}\mathbf{z}'_{\text{R}} e^{-\mathcal{H}[\mathbf{z}'_{\text{R}}]}}{\int \mathrm{d}\mathbf{z}'_{\text{R}} \mathrm{d}\mathbf{z}'_{\text{E}} e^{-\mathcal{H}[\mathbf{z}'_{\text{R}}, \mathbf{z}'_{\text{E}}]}}\right) e^{-(\mathcal{H}[\boldsymbol{y}_{\text{R}}, \mathbf{z}_{\text{E}}] - \mathcal{H}[\boldsymbol{y}_{\text{R}}])}\,. \tag{Bayes}$$

The leading Gaussian-process contributions can be segregated out through the textbook manipulation (Williams and Rasmussen, 2006) [c.f. Appendix D]: denoting the full Gaussian-process kernel in the last layer as

$$K_{i_1 i_2; \alpha_1 \alpha_2} = \delta_{i_1 i_2} \begin{pmatrix} \left(\widetilde{K}_{\text{RR}}\right)_{\bar{\beta}_1 \bar{\beta}_2} & \left(\widetilde{K}_{\text{RE}}\right)_{\bar{\beta}_1 \dot{\gamma}_2} \\ \left(\widetilde{K}_{\text{ER}}\right)_{\dot{\gamma}_1 \bar{\beta}_2} & \left(\widetilde{K}_{\text{EE}}\right)_{\dot{\gamma}_1 \dot{\gamma}_2} \end{pmatrix} \tag{21}$$

and the Gaussian-process posterior mean prediction as

$$\left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} \equiv \sum_{\bar{\beta}} \left[\widetilde{K}_{\mathrm{ER}} \widetilde{K}_{\mathrm{RR}}^{-1}\right]_{\dot{\gamma}}^{\bar{\beta}} \left(y_{\mathrm{R}}\right)_{i;\bar{\beta}} , \tag{GPM}$$

and defining a fluctuation $\left(\mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}} \equiv \left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} + \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}}$ and a matrix $\widetilde{K}_{\Delta} \equiv \widetilde{K}_{\mathrm{EE}} - \widetilde{K}_{\mathrm{ER}} \widetilde{K}_{\mathrm{RR}}^{-1} \widetilde{K}_{\mathrm{RE}}$,

$$\mathcal{H}_0\left[\boldsymbol{y}_{\mathrm{R}}, \mathbf{z}_{\mathrm{E}}\right] - \mathcal{H}_0\left[\boldsymbol{y}_{\mathrm{R}}\right] = \frac{1}{2} \sum_{i} \sum_{\dot{\gamma}_1, \dot{\gamma}_2} \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}_1} \left(\widetilde{K}_{\Delta}^{-1}\right)^{\dot{\gamma}_1 \dot{\gamma}_2} \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}_2} . \tag{GPΔ}$$

For any function $\mathcal{F}$, its expectation over the Bayesian posterior (Bayes) then turns into

$$\int \mathrm{d}\mathbf{z}_{\mathrm{E}} \mathcal{F}[\mathbf{z}_{\mathrm{E}}] p\left[\mathbf{z}_{\mathrm{E}} | \boldsymbol{y}_{\mathrm{R}}\right] = \widetilde{\mathcal{N}} \left\langle e^{-\epsilon \mathcal{H}_1\left[\boldsymbol{y}_{\mathrm{R}}, \mathbf{y}_{\mathrm{E}}^{\mathrm{GP}} + \delta \mathbf{z}_{\mathrm{E}}\right]} \mathcal{F}[\mathbf{y}_{\mathrm{E}}^{\mathrm{GP}} + \delta \mathbf{z}_{\mathrm{E}}] \right\rangle_{K_{\Delta}} \tag{22}$$

where the deviation kernel $\left\langle \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i_1;\dot{\gamma}_1} \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i_2;\dot{\gamma}_2} \right\rangle_{K_{\Delta}} \equiv \delta_{i_1 i_2} \left(\widetilde{K}_{\Delta}\right)_{\dot{\gamma}_1 \dot{\gamma}_2}$ and the normalization factor

$$\widetilde{\mathcal{N}} = \left[\left\langle e^{-\epsilon \mathcal{H}_1\left[\boldsymbol{y}_{\mathrm{R}}, \mathbf{y}_{\mathrm{E}}^{\mathrm{GP}} + \delta \mathbf{z}_{\mathrm{E}}\right]} \right\rangle_{K_{\Delta}} \right]^{-1} = 1 + O(\epsilon) . \tag{23}$$

In particular the mean posterior output is given by

$$\begin{aligned} \left(\overline{\mathrm{y}}_{\mathrm{E}}\right)_{i;\dot{\gamma}} \equiv \int \mathrm{d}\mathbf{z}_{\mathrm{E}} \left(\mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}} p\left[\mathbf{z}_{\mathrm{E}} | \boldsymbol{y}_{\mathrm{R}}\right] &= \left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} + \widetilde{\mathcal{N}} \left\langle \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}} e^{-\epsilon \mathcal{H}_1\left[\boldsymbol{y}_{\mathrm{R}}, \mathbf{y}_{\mathrm{E}}^{\mathrm{GP}} + \delta \mathbf{z}_{\mathrm{E}}\right]} \right\rangle_{K_{\Delta}} \\ &= \left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} - \epsilon \left\langle \left(\delta \mathrm{z}_{\mathrm{E}}\right)_{i;\dot{\gamma}} \mathcal{H}_1\left[\boldsymbol{y}_{\mathrm{R}}, \mathbf{y}_{\mathrm{E}}^{\mathrm{GP}} + \delta \mathbf{z}_{\mathrm{E}}\right] \right\rangle_{K_{\Delta}} + O(\epsilon^2) . \end{aligned} \tag{24}$$

Stringing together $\overline{\phi}_{i;\alpha} \equiv \left[\left(y_{\mathrm{R}}\right)_{i;\bar{\beta}}, \left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}}\right]$, recalling Equation (D2) for $\mathcal{H}_1$, and using Wick's contractions for one last time, the mean prediction becomes

$$\begin{aligned} &\left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} \\ &+ \epsilon \sum_{\alpha_1, \dot{\gamma}_1} \left(\widetilde{K}_{\Delta}\right)_{\dot{\gamma}\dot{\gamma}_1} \overline{\phi}_{i;\alpha_1} \left[\widetilde{S}^{\dot{\gamma}_1 \alpha_1} - \sum_{\alpha_2, \alpha_3} \widetilde{V}^{(\dot{\gamma}_1 \alpha_2)(\alpha_1 \alpha_3)} \widetilde{K}_{\alpha_2 \alpha_3} + \sum_{\dot{\gamma}_2, \dot{\gamma}_3} \widetilde{V}^{(\dot{\gamma}_1 \dot{\gamma}_2)(\alpha_1 \dot{\gamma}_3)} \left(\widetilde{K}_{\Delta}\right)_{\dot{\gamma}_2 \dot{\gamma}_3} \right. \\ &\left. + \frac{n_L}{2} \sum_{\dot{\gamma}_2, \dot{\gamma}_3} \widetilde{V}^{(\alpha_1 \dot{\gamma}_1)(\dot{\gamma}_2 \dot{\gamma}_3)} \left(\widetilde{K}_{\Delta}\right)_{\dot{\gamma}_2 \dot{\gamma}_3} + \sum_{\alpha_2, \alpha_3} \widetilde{V}^{(\dot{\gamma}_1 \alpha_1)(\alpha_2 \alpha_3)} \left(-\frac{n_L}{2} \widetilde{K}_{\alpha_2 \alpha_3} + \frac{1}{2} \sum_{j} \overline{\phi}_{j;\alpha_2} \overline{\phi}_{j;\alpha_3} \right) \right] . \end{aligned} \tag{NGPM}$$

With additional manipulations, this expression can be simplified into the actionable form that is amenable to use in practice [c.f. Equations (NGPM') and (NGPM") in Appendix D]. It turns out that for deep linear networks the leading finite-width correction vanishes, and the first correction is likely to show up at higher order in $1/n$ asymptotic expansion, which is not carried out in this paper. Here we instead use the $L = 2$ multilayer perceptron with the quadratic activation $\sigma(z) = z^2$, zero bias variance $C_b^{(\ell)} = 0$, and weight variance $C_W^{(\ell)} = 1/3$ for illustration, plugging Equations (S30,S31,S32) into Equations (NGPM') and (NGPM") and varying $\epsilon \equiv \frac{1}{n_{L-1}} = \frac{1}{n_1}$. Results in Figure 2 indicate the regularization effects of finite widths when the number of training samples, N_{R}, is small, resulting in peak performance at finite widths. This is in line with expectations that finite widths ameliorate overfitting and that non-Gaussian priors increase the expressivity of neural functions, but additional large-scale extensive experiments would be desirable in the future.

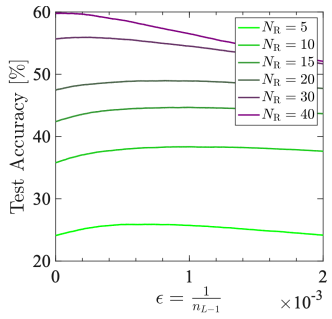

Figure 2: Test accuracy for $N_E = 10000$ MNIST test data as a function of the inverse width $\epsilon = 1/n_{L-1}$ of the hidden layer with quadratic activation. For each number N_R of subsampled training data, the result is averaged over 10 distinct choices of such subsamplings. For small numbers of training data, finite widths result in regularization effects, improving the test accuracy.

6. Dreams

In this paper, we have developed the perturbative formalism that captures the flow of preactivation distributions from lower to higher layers. The resemblance between our recursive equations and renormalization-group flow equations in high-energy and statistical physics is highly appealing. It would be exciting to investigate the structure of fixed points away from the Gaussian asymptopia (Schoenholz et al., 2016) and fully realize the dream articulated by Mehta and Schwab (2014) – the audacious hypothesis that neural networks wash away microscopic irrelevancies and extract relevant features – beyond their limited example of a mapping between two antiquated techniques.

In addition we have developed the perturbative Bayesian inference scheme universally applicable whenever prior distributions are weakly non-Gaussian, and have applied it to the specific cases of neural networks at finite widths. In light of possible finite-width regularization effects, it would be prudent to revisit the empirical comparison between SGD optimization and Bayesian inference at finite widths (Lee et al., 2018; Novak et al., 2019), especially for convolutional neural networks.

Finally, given surging interests in SGD dynamics within the large-width regime (Jacot et al., 2018; Chizat et al., 2018; Lee et al., 2019; Cohen et al., 2019; Dyer and Gur-Ari, 2019), it would be natural to adapt our formalism for investigating corrections to neural tangent kernels, and even aspire to capture a transition from lazy-learning to feature-learning regimes.

Acknowledgments

The author thanks Yasaman Bahri for the discussion that seeded the idea for this project, Boris L. Hanin for persistently preaching about Gaussian processes, and David J. Schwab for permission to call his example limited with our friendship intact. The author also thanks Ethan S. Dyer, Mario Geiger, Guy Gur-Ari, Eric T. Mintun, Stephen H. Shenker, and Lexing Ying for substantially useful discussions, and Daniel A. Roberts for the quality control of all the jokes and more.

References

Joseph M. Antognini. Finite size corrections for neural network Gaussian processes. *arXiv preprint arXiv:1908.10030*, 2019.

Robert Brown. XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4 (21):161–173, 1828.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pages 342–350, 2009.

Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for deep neural networks: a Gaussian field theory perspective. *arXiv preprint arXiv:1906.05301*, 2019.

Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from Feynman diagrams. *arXiv preprint arXiv:1909.11304*, 2019.

Albert Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560, 1905.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy learning in deep neural networks: an empirical study. *arXiv preprint arXiv:1906.08034*, 2019.

Nigel Goldenfeld. *Lectures on phase transitions and the renormalization group*. CRC Press, 2018.

Boris Hanin and David Rolnick. How to start training: the effect of initialization and architecture. In *Advances in Neural Information Processing Systems*, pages 571–581, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Leo P. Kadanoff. Scaling laws for Ising models near T_c. *Physics Physique Fizika*, 2(6):263–272, 1966.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

Qianli Liao and Tomaso Poggio. Theory II: landscape of the empirical risk in deep learning. *arXiv preprint arXiv:1703.09833*, 2017.

David J. C. MacKay. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning. *arXiv preprint arXiv:1410.3831*, 2014.

Moshe Moshe and Jean Zinn-Justin. Quantum field theory in the large N limit: a review. *Physics Reports*, 385(3-6):69–228, 2003.

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

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.

Giorgio Parisi and Francesco Zamponi. Mean-field theory of hard sphere glasses and jamming. *Reviews of Modern Physics*, 82(1):789–845, 2010.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

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

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2016. URL https://openreview.net/forum?id=H1W1UN9gg.

Steven Weinberg. *The quantum theory of fields: Volume I: Foundations*. Cambridge University Press, 1995.

Steven Weinberg. *The quantum theory of fields: Volume II: Modern Applications*. Cambridge University Press, 1996.

Gian-Carlo Wick. The evaluation of the collision matrix. *Physical Review*, 80(2):268–272, 1950.

Christopher K. I. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, pages 295–301, 1997.

Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006.

Kenneth G. Wilson. Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Physical Review B*, 4(9):3174–3183, 1971.

Anthony Zee. *Quantum field theory in a nutshell*. Princeton University Press, 2010.

Appendix A. Wick's tricks

Here is all you need to know in order to follow the calculations in the paper. In the main text, Wick's contractions are used both for trivially integrating out biases and weights as straightforward applications of Appendix A.1 and for nontrivially integrating out preactivations, with concepts of cumulants reviewed in Appendix A.2 and A.3, culminating in the hack derived in Appendix A.4. The random variables are generically indexed by $\mu = 1, \ldots, N$ throughout this Appendix: when applying formulae for biases, $\mu = i$; for weights $\mu = (i, j)$; for full preactivations $\mu = (i, \alpha)$; for single-neuron preactivations $\mu = \alpha$.

A.1. Wick's contractions

For Gaussian-distributed variables $\mathbf{z} = \{\mathrm{z}_\mu\}_{\mu=1,\ldots,N}$ with a kernel $K_{\mu\mu'}$, moments

$$\left\langle \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \cdots \mathrm{z}_{\mu_m} \right\rangle_K \equiv \frac{\int \mathrm{d}\mathbf{z} e^{-\mathcal{H}_0[\mathbf{z}]} \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \cdots \mathrm{z}_{\mu_m}}{\int \mathrm{d}\mathbf{z} e^{-\mathcal{H}_0[\mathbf{z}]}} \quad \text{with} \quad \mathcal{H}_0[\mathbf{z}] \equiv \frac{1}{2} \sum_{\mu,\mu'=1}^{N} \mathrm{z}_\mu \left(K^{-1}\right)^{\mu\mu'} \mathrm{z}_{\mu'} \,. \tag{S1}$$

For any odd m such moments identically vanish. For even m, Isserlis-Wick's theorem states that

$$\left\langle \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \cdots \mathrm{z}_{\mu_m} \right\rangle_K = \sum_{\text{all pairing}} K_{\mu_{k_1}\mu_{k_2}} \cdots K_{\mu_{k_{m-1}}\mu_{k_m}} \tag{S2}$$

where the sum is over all the possible pairings of m variables, $(k_1, k_2), \ldots, (k_{m-1}, k_m)$. In general, there are $(m-1)!! = (m-1)\cdot(m-3)\cdots 1$ such pairings. For a proof, see for example Zee (2010). In order to understand and use the theorem, it is instructive to look at a few examples:

$$\left\langle \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \right\rangle_K = K_{\mu_1\mu_2} \,; \tag{S3}$$

$$\left\langle \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \mathrm{z}_{\mu_3} \mathrm{z}_{\mu_4} \right\rangle_K = K_{\mu_1\mu_2} K_{\mu_3\mu_4} + K_{\mu_1\mu_3} K_{\mu_2\mu_4} + K_{\mu_1\mu_4} K_{\mu_2\mu_3} \,; \tag{S4}$$

and

$$\begin{aligned} & \left\langle \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \mathrm{z}_{\mu_3} \mathrm{z}_{\mu_4} \mathrm{z}_{\mu_5} \mathrm{z}_{\mu_6} \right\rangle_K \\ = \; & K_{\mu_1\mu_2} K_{\mu_3\mu_4} K_{\mu_5\mu_6} + K_{\mu_1\mu_3} K_{\mu_2\mu_4} K_{\mu_5\mu_6} + K_{\mu_1\mu_4} K_{\mu_2\mu_3} K_{\mu_5\mu_6} \\ + \; & K_{\mu_1\mu_2} K_{\mu_3\mu_5} K_{\mu_4\mu_6} + K_{\mu_1\mu_3} K_{\mu_2\mu_5} K_{\mu_4\mu_6} + K_{\mu_1\mu_5} K_{\mu_2\mu_3} K_{\mu_4\mu_6} \\ + \; & K_{\mu_1\mu_2} K_{\mu_5\mu_4} K_{\mu_3\mu_6} + K_{\mu_1\mu_5} K_{\mu_2\mu_4} K_{\mu_3\mu_6} + K_{\mu_1\mu_4} K_{\mu_2\mu_5} K_{\mu_3\mu_6} \\ + \; & K_{\mu_1\mu_5} K_{\mu_3\mu_4} K_{\mu_2\mu_6} + K_{\mu_1\mu_3} K_{\mu_5\mu_4} K_{\mu_2\mu_6} + K_{\mu_1\mu_4} K_{\mu_5\mu_3} K_{\mu_2\mu_6} \\ + \; & K_{\mu_5\mu_2} K_{\mu_3\mu_4} K_{\mu_1\mu_6} + K_{\mu_5\mu_3} K_{\mu_2\mu_4} K_{\mu_1\mu_6} + K_{\mu_5\mu_4} K_{\mu_2\mu_3} K_{\mu_1\mu_6} \,. \end{aligned} \tag{S5}$$

A.2. Connected correlations

Given general (not necessarily Gaussian) random variables, connected correlation functions are defined inductively through

$$\begin{aligned} & \mathbb{E}\left[\mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \cdots \mathrm{z}_{\mu_m}\right] \\ \equiv \; & \mathbb{E}\left[\mathrm{z}_{\mu_1} \mathrm{z}_{\mu_2} \cdots \mathrm{z}_{\mu_m}\right]\Big|_{\text{connected}} \\ + \; & \sum_{\text{all subdivisions}} \mathbb{E}\left[\mathrm{z}_{\mu_{k_1^{[1]}}} \cdots \mathrm{z}_{\mu_{k_{\nu_1}^{[1]}}}\right]\Big|_{\text{connected}} \cdots \mathbb{E}\left[\mathrm{z}_{\mu_{k_1^{[s]}}} \cdots \mathrm{z}_{\mu_{k_{\nu_s}^{[s]}}}\right]\Big|_{\text{connected}} \end{aligned} \tag{S6}$$

where the sum is over all the possible subdivisions of m variables into $s > 1$ clusters of sizes $(\nu_1, \ldots, \nu_s)$ as $(k_1^{[1]}, \ldots, k_{\nu_1}^{[1]}), \ldots, (k_1^{[s]}, \ldots, k_{\nu_s}^{[s]})$. In order to understand the definition, it is again instructive to look at a few examples. Assuming that all the odd moments vanish,

$$\begin{aligned} \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right] &= \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right]\big|_{\text{connected}} \quad \text{and} && \text{(S7)} \\ \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right] &= \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} && \text{(S8)} \\ &\quad + \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right]\big|_{\text{connected}} \mathbb{E}\left[\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} \\ &\quad + \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_3}\right]\big|_{\text{connected}} \mathbb{E}\left[\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} \\ &\quad + \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} \mathbb{E}\left[\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\right]\big|_{\text{connected}} . \end{aligned}$$

Rearranging them in particular yields

$$\begin{aligned} & \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} && \text{(S9)} \\ = \; & \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right] \\ & - \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right]\mathbb{E}\left[\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right] - \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_3}\right]\mathbb{E}\left[\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_4}\right] - \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_4}\right]\mathbb{E}\left[\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\right] . \end{aligned}$$

If these examples do not suffice, here is yet another example to chew on:

$$\begin{aligned} \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right] &= \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right]\big|_{\text{connected}} && \text{(S10)} \\ &\quad + \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right]\big|_{\text{connected}} \mathbb{E}\left[\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} \mathbb{E}\left[\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right]\big|_{\text{connected}} \\ &\quad + [14 \text{ other } (2,2,2) \text{ subdivisions}] \\ &\quad + \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} \mathbb{E}\left[\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right]\big|_{\text{connected}} \\ &\quad + [14 \text{ other } (4,2) \text{ subdivisions}] \end{aligned}$$

and hence

$$\begin{aligned} & \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right]\big|_{\text{connected}} && \text{(S11)} \\ = \; & \mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right] \\ & - \left\{\mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\mathbb{E}\left[\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right] + [14 \text{ other } (4,2) \text{ subdivisions}]\right\} \\ & + 2\left\{\mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right]\mathbb{E}\left[\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\mathbb{E}\left[\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right] + [14 \text{ other } (2,2,2) \text{ subdivisions}]\right\} . \end{aligned}$$

We emphasize that these are just renderings of the definition (S6). The power of this definition will be illustrated in the next two subsections.

A.3. Hierarchical clustering

We often encounter situations with the hierarchy

$$\mathbb{E}\left[\mathrm{z}_{\mu_1}\cdots\mathrm{z}_{\mu_m}\right]\big|_{\text{connected}} = O(\epsilon^{\frac{m-2}{2}}) \tag{S12}$$

where $\epsilon \ll 1$ is a small perturbative parameter and here again odd moments are assumed to vanish. Often comes with the hierarchical structure is the asymptotic limit $\epsilon \to 0$ where

$$\mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right] = K_{\mu_1\mu_2} + \epsilon S_{\mu_1\mu_2} + O(\epsilon^2) \tag{S13}$$

with the Gaussian kernel $K_{\mu_1\mu_2}$ at zero ϵ and the leading self-energy correction $S_{\mu_1\mu_2}$. Let us also denote the leading four-point vertex

$$\mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}} = \epsilon V_{\mu_1\mu_2\mu_3\mu_4} + O(\epsilon^2)\,. \tag{S14}$$

For instance this hierarchy holds for weakly-coupled field theories – from which we are importing names such as self-energy, vertex, and metric – and, in this paper, such hierarchical structure is inductively shown to hold for prior preactivations $\mathbf{z}^{(\ell)}$ with $\epsilon = \frac{1}{n_{\ell-1}}$ in the regime $n_1, \ldots, n_{L-1} \sim n \gg 1$. Note that, by definition, $K_{\mu_1\mu_2}$ and $S_{\mu_1\mu_2}$ are symmetric under $\mu_1 \leftrightarrow \mu_2$ and $V_{\mu_1\mu_2\mu_3\mu_4}$ is symmetric under permutations of $(\mu_1, \mu_2, \mu_3, \mu_4)$.[7]

A.4. Combinatorial hack

So far we have reviewed the standard technology of Wick's contractions, connected correlation functions, and all that. Our objective now is to develop a method to evaluate $\mathbb{E}\left[\mathrm{z}_{\mu_1}\cdots\mathrm{z}_{\mu_m}\right]$ for random variables obeying the hierarchical-clustering property (S12),[8] which is the inductive hypothesis made in the main text; by extension, the resulting method (HACK') lets us perturbatively evaluate $\mathbb{E}\left\{\mathcal{F}[\mathbf{z}]\right\}$ for any function $\mathcal{F}$ that can be obtained as a limit of a sequence of analytic functions.

With the review of connected correlation functions passed us, first note that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{z}_{\mu_1}\cdots\mathrm{z}_{\mu_m}\right] && \text{(CLUSTER)}\\
=&\mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\right]\cdots\mathbb{E}\left[\mathrm{z}_{\mu_{m-1}}\mathrm{z}_{\mu_m}\right] + \left\{\left[(m-1)!!-1\right]\text{ other pairings}\right\}\\
&+\mathbb{E}\left[\mathrm{z}_{\mu_1}\mathrm{z}_{\mu_2}\mathrm{z}_{\mu_3}\mathrm{z}_{\mu_4}\right]\big|_{\text{connected}}\mathbb{E}\left[\mathrm{z}_{\mu_5}\mathrm{z}_{\mu_6}\right]\cdots\mathbb{E}\left[\mathrm{z}_{\mu_{m-1}}\mathrm{z}_{\mu_m}\right]\\
&+\left\{\left[\binom{m}{4}\times(m-5)!!-1\right]\text{ other }(4,2,2,\ldots,2)\text{ clusterings}\right\}+O(\epsilon^2)\\
=&K_{\mu_1\mu_2}\cdots K_{\mu_{m-1}\mu_m} + \left\{\left[(m-1)!!-1\right]\text{ other pairings}\right\}\\
&+\epsilon S_{\mu_1\mu_2}K_{\mu_3\mu_4}\cdots K_{\mu_{m-1}\mu_m}\\
&+\left\{\left[\binom{m}{2}\times(m-3)!!-1\right]\text{ other such clusterings}\right\}\\
&+\epsilon V_{\mu_1\mu_2\mu_3\mu_4}K_{\mu_5\mu_6}\cdots K_{\mu_{m-1}\mu_m}\\
&+\left\{\left[\binom{m}{4}\times(m-5)!!-1\right]\text{ other }(4,2,2,\ldots,2)\text{ clusterings}\right\}+O(\epsilon^2)\\
=&\left\langle\mathrm{z}_{\mu_1}\cdots\mathrm{z}_{\mu_m}\right\rangle_K && \text{(CLUSTER')}\\
&+\epsilon S_{\mu_1\mu_2}\left\langle\mathrm{z}_{\mu_3}\cdots\mathrm{z}_{\mu_m}\right\rangle_K+\left\{\left[\binom{m}{2}-1\right]\text{ other self-energy contractions}\right\}\\
&+\epsilon V_{\mu_1\mu_2\mu_3\mu_4}\left\langle\mathrm{z}_{\mu_5}\cdots\mathrm{z}_{\mu_m}\right\rangle_K+\left\{\left[\binom{m}{4}-1\right]\text{ other vertex contractions}\right\}+O(\epsilon^2)\,.
\end{aligned}$$

where in the last equality Wick's theorem was used backward.

7. In the main text the connected four-point preactivation correlation functions are symmetric under the permutations of four (sample, neuron) indices, $\{(i_1,\alpha_1),(i_2,\alpha_2),(i_3,\alpha_3),(i_4,\alpha_4)\}$.
8. More precisely, we shall inductively use only the weaker proposition that $\mathbb{E}\left[\mathrm{z}_{\mu_1}\cdots\mathrm{z}_{\mu_m}\right]\big|_{\text{connected}} = O(\epsilon^2)$ for $m \geq 6$ along with Equations (S13) and (S14).

Below, let us use the inverse kernel $\left(K^{-1}\right)^{\mu_1\mu_2}$ as a metric to raise indices:

$$S^{\mu_1\mu_2} \equiv \sum_{\mu_1',\mu_2'} \left(K^{-1}\right)^{\mu_1\mu_1'} \left(K^{-1}\right)^{\mu_2\mu_2'} S_{\mu_1'\mu_2'} \quad \text{and} \tag{S15}$$

$$V^{\mu_1\mu_2\mu_3\mu_4} \equiv \sum_{\mu_1',\ldots,\mu_4'} \left(K^{-1}\right)^{\mu_1\mu_1'} \cdots \left(K^{-1}\right)^{\mu_4\mu_4'} V_{\mu_1'\mu_2'\mu_3'\mu_4'} \,. \tag{S16}$$

Then, in order to simplify the second set of terms in Equation (CLUSTER') involving self-energy, note that

$$\begin{aligned}
&\left\langle \mathrm{z}_{\mu_1} \cdots \mathrm{z}_{\mu_m} \left(\sum_{\mu_1',\mu_2'} S^{\mu_1'\mu_2'} \mathrm{z}_{\mu_1'} \mathrm{z}_{\mu_2'} \right) \right\rangle_K \\
= &\sum_{\mu_1',\mu_2'} S^{\mu_1'\mu_2'} \Bigg[\left\langle \mathrm{z}_{\mu_1'} \mathrm{z}_{\mu_2'} \right\rangle_K \left\langle \mathrm{z}_{\mu_1} \cdots \mathrm{z}_{\mu_m} \right\rangle_K \\
&\qquad + 2 \left\langle \mathrm{z}_{\mu_1} \mathrm{z}_{\mu_1'} \right\rangle_K \left\langle \mathrm{z}_{\mu_2} \mathrm{z}_{\mu_2'} \right\rangle_K \left\langle \mathrm{z}_{\mu_3} \cdots \mathrm{z}_{\mu_m} \right\rangle_K + \left\{ \left[\binom{m}{2} - 1 \right] \text{ other } (\mu_1, \mu_2) \right\} \Bigg] \\
= &\left(\sum_{\mu_1',\mu_2'} S^{\mu_1'\mu_2'} K_{\mu_1'\mu_2'} \right) \left\langle \mathrm{z}_{\mu_1} \cdots \mathrm{z}_{\mu_m} \right\rangle_K \\
&+ 2 S_{\mu_1\mu_2} \left\langle \mathrm{z}_{\mu_3} \cdots \mathrm{z}_{\mu_m} \right\rangle_K + \left\{ \left[\binom{m}{2} - 1 \right] \text{ other } (\mu_1, \mu_2) \right\}
\end{aligned}$$

where the symmetry $\mu_1 \leftrightarrow \mu_2$ of $S_{\mu_1\mu_2}$ was used. Hence, defining

$$\mathcal{O}_S[\mathbf{z}] \equiv \frac{1}{2} \sum_{\mu_1',\mu_2'} S^{\mu_1'\mu_2'} \left(\mathrm{z}_{\mu_1'} \mathrm{z}_{\mu_2'} - K_{\mu_1'\mu_2'} \right) \,, \tag{OS'}$$

we obtain

$$\epsilon S_{\mu_1\mu_2} \left\langle \mathrm{z}_{\mu_3} \cdots \mathrm{z}_{\mu_m} \right\rangle_K + \left\{ \left[\binom{m}{2} - 1 \right] \text{ other } (\mu_1, \mu_2) \right\} \tag{S17}$$

$$= \epsilon \left\langle \mathrm{z}_{\mu_1} \cdots \mathrm{z}_{\mu_m} \mathcal{O}_S[\mathbf{z}] \right\rangle_K \,. \tag{S18}$$

The similar algebraic exercise renders the other term in Equation (CLUSTER') to be

$$\epsilon V_{\mu_1\mu_2\mu_3\mu_4} \left\langle \mathrm{z}_{\mu_5} \cdots \mathrm{z}_{\mu_m} \right\rangle_K + \left\{ \left[\binom{m}{4} - 1 \right] \text{ other } (\mu_1, \mu_2, \mu_3, \mu_4) \right\} \tag{S19}$$

$$= \epsilon \left\langle \mathrm{z}_{\mu_1} \cdots \mathrm{z}_{\mu_m} \mathcal{O}_V[\mathbf{z}] \right\rangle_K \tag{S20}$$

with

$$\mathcal{O}_V[\mathbf{z}] \equiv \frac{1}{24} \sum_{\mu_1',\ldots,\mu_4'} V^{\mu_1'\mu_2'\mu_3'\mu_4'} \left(\mathrm{z}_{\mu_1'} \mathrm{z}_{\mu_2'} \mathrm{z}_{\mu_3'} \mathrm{z}_{\mu_4'} - 6 \mathrm{z}_{\mu_1'} \mathrm{z}_{\mu_2'} K_{\mu_3'\mu_4'} + 3 K_{\mu_1'\mu_2'} K_{\mu_3'\mu_4'} \right) \,. \tag{OV'}$$

In summary, for any function $\mathcal{F}[\mathbf{z}]$ of random variables z_μ

$$\mathbb{E}\left\{\mathcal{F}[\mathbf{z}]\right\} = \left\langle \mathcal{F}[\mathbf{z}]\right\rangle_K + \epsilon \left\langle \mathcal{F}[\mathbf{z}]\mathcal{O}_S[\mathbf{z}]\right\rangle_K + \epsilon \left\langle \mathcal{F}[\mathbf{z}]\mathcal{O}_V[\mathbf{z}]\right\rangle_K + O(\epsilon^2)\,. \tag{HACK'}$$

In order to get the expressions used in the main text at the ℓ-th layer, we need only to replace $\mu \to (i,\alpha)$, identify $\epsilon = \frac{1}{n_{\ell-1}}$, and use the inductive hypotheses (KS)

$$\mathbb{E}\left[\mathrm{z}_{i_1;\alpha_1}\mathrm{z}_{i_2;\alpha_2}\right] = \delta_{i_1 i_2}\left[\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2} + \frac{1}{n_{\ell-1}}\widetilde{S}^{(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n^2}\right)\right]$$

and (V)

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{z}_{i_1;\alpha_1}\mathrm{z}_{i_2;\alpha_2}\mathrm{z}_{i_3;\alpha_3}\mathrm{z}_{i_4;\alpha_4}\right]\Big|_{\text{connected}} \\
&= \frac{1}{n_{\ell-1}}\Big[\delta_{i_1 i_2}\delta_{i_3 i_4}\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4}\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} \\
&\qquad + \delta_{i_1 i_4}\delta_{i_2 i_3}\widetilde{V}^{(\ell)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}\Big] + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

The operators in Equations (OS') and (OV') then become

$$\begin{aligned}
\mathcal{O}_S[\mathbf{z}] &= \frac{1}{2}\sum_{\alpha_1,\alpha_2}\widetilde{S}^{\alpha_1\alpha_2}_{(\ell)}\left[\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right) - n_\ell\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}\right] \quad \text{and} \\
\mathcal{O}_V[\mathbf{z}] &= \frac{1}{8}\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}\widetilde{V}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} \\
&\quad\times\Bigg[\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right)\left(\sum_{j=1}^{n_\ell}\mathrm{z}_{j;\alpha_3}\mathrm{z}_{j;\alpha_4}\right) - 2n_\ell\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_2}\right)\widetilde{K}^{(\ell)}_{\alpha_3\alpha_4} \\
&\quad -4\left(\sum_{i=1}^{n_\ell}\mathrm{z}_{i;\alpha_1}\mathrm{z}_{i;\alpha_3}\right)\widetilde{K}^{(\ell)}_{\alpha_2\alpha_4} + n_\ell^2\widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}\widetilde{K}^{(\ell)}_{\alpha_3\alpha_4} + 2n_\ell\widetilde{K}^{(\ell)}_{\alpha_1\alpha_3}\widetilde{K}^{(\ell)}_{\alpha_2\alpha_4}\Bigg],
\end{aligned}$$

i.e., the operators in Equations (OS) and (OV) in the main text.

Appendix B. Full condensed proof

In this Appendix, we provide a full inductive proof for one of the main claims in the paper, streamlined in the main text. Namely, we assume at the ℓ-th layer that Equations (KS) and (V) hold and that all the higher-point connected preactivation correlation functions are of order $O\left(\frac{1}{n^2}\right)$ – which are trivially true at $\ell = 1$ – and prove the same for the $(\ell+1)$-th layer. We assume the full mastery of Appendix A or, conversely, this section can be used to test the mastery of Wick's tricks.

First, trivial Wick's contractions yield

$$\begin{aligned} & G^{(\ell+1)}_{i_1 \ldots i_{2m};\alpha_1 \ldots \alpha_{2m}} \\ = \;& \delta_{i_1 i_2} \cdots \delta_{i_{2m-1} i_{2m}} \sum_{k=0}^{m} \left[C_b^{(\ell+1)}\right]^{m-k} \left[C_W^{(\ell+1)}\right]^{k} \\ & \times \left\{ \frac{1}{n_\ell^k} \sum_{j_1,\ldots,j_k=1}^{n_\ell} H^{(\ell)}_{j_1 j_1 \ldots j_k j_k;\alpha_1 \alpha_2 \ldots \alpha_{2k-1}\alpha_{2k}} + \left[\binom{m}{k} - 1 \text{ others} \right] \right\} \\ & + \left[(2m-1)!! - 1 \text{ other pairings}\right] . \end{aligned} \tag{S21}$$

Studiously disentangling cases with different numbers of repetitions in neuron indices $(j_1, \ldots, j_k)$, we notice that at order $O\left(\frac{1}{n}\right)$, terms without repetition or with only one repetition contribute, finding

$$\begin{aligned} & \frac{1}{n_\ell^k} \sum_{j_1,\ldots,j_k=1}^{n_\ell} H^{(\ell)}_{j_1 j_1 \ldots j_k j_k;\alpha_1 \alpha_2 \ldots \alpha_{2k-1}\alpha_{2k}} \\ = \;& \left[\left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\right\rangle_{\widetilde{K}^{(\ell)}} \cdots \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_{2k-1}})\sigma(\tilde{\mathrm{z}}_{\alpha_{2k}})\right\rangle_{\widetilde{K}^{(\ell)}}\right] \\ & + \frac{1}{n_\ell} \left\{ \left[\left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\sigma(\tilde{\mathrm{z}}_{\alpha_3})\sigma(\tilde{\mathrm{z}}_{\alpha_4})\right\rangle_{\widetilde{K}^{(\ell)}} - \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\right\rangle_{\widetilde{K}^{(\ell)}} \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_3})\sigma(\tilde{\mathrm{z}}_{\alpha_4})\right\rangle_{\widetilde{K}^{(\ell)}}\right] \right. \\ & \qquad \times \left[\left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_5})\sigma(\tilde{\mathrm{z}}_{\alpha_6})\right\rangle_{\widetilde{K}^{(\ell)}} \cdots \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_{2k-1}})\sigma(\tilde{\mathrm{z}}_{\alpha_{2k}})\right\rangle_{\widetilde{K}^{(\ell)}}\right] \\ & \qquad \left. + \left[\binom{k}{2} - 1 \text{ others} \right] \right\} \\ & + \frac{1}{n_{\ell-1}} \left\langle \left[\sigma(\mathrm{z}_{1;\alpha_1})\sigma(\mathrm{z}_{1;\alpha_2}) \cdots \sigma(\mathrm{z}_{k;\alpha_{2k-1}})\sigma(\mathrm{z}_{k;\alpha_{2k}})\right] \left\{\mathcal{O}_S[\mathbf{z}] + \mathcal{O}_V[\mathbf{z}]\right\} \right\rangle_{K^{(\ell)}} \\ & + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{S22}$$

where we used the inductive hierarchical assumption at the ℓ-th layer, i.e., its consequence (HACK) and denoted a single-neuron random vector $\tilde{\mathbf{z}} = \{\tilde{\mathrm{z}}_\alpha\}_{\alpha=1,\ldots,N_\mathrm{D}}$ and the Gaussian integral with the core kernel $\left\langle \tilde{\mathrm{z}}_{\alpha_1}\tilde{\mathrm{z}}_{\alpha_2}\right\rangle_{\widetilde{K}^{(\ell)}} = \widetilde{K}^{(\ell)}_{\alpha_1\alpha_2}$. Plugging in expressions (OS,OV) for operators $\mathcal{O}_{S,V}[\mathbf{z}]$,

$$\begin{aligned} & \left\langle \left[\sigma(\mathrm{z}_{1,\alpha_1})\sigma(\mathrm{z}_{1,\alpha_2}) \cdots \sigma(\mathrm{z}_{k,\alpha_{2k-1}})\sigma(\mathrm{z}_{k,\alpha_{2k}})\right] \mathcal{O}_S[\mathbf{z}] \right\rangle_{K^{(\ell)}} \\ = \;& \frac{1}{2} \sum_{\alpha_1',\alpha_2'} \widetilde{S}^{\alpha_1'\alpha_2'}_{(\ell)} \left\{ \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2}) \left(\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'} - \widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'}\right) \right\rangle_{\widetilde{K}^{(\ell)}} \right. \\ & \qquad \left. \times \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_3})\sigma(\tilde{\mathrm{z}}_{\alpha_4})\right\rangle_{\widetilde{K}^{(\ell)}} \cdots \left\langle \sigma(\tilde{\mathrm{z}}_{\alpha_{2k-1}})\sigma(\tilde{\mathrm{z}}_{\alpha_{2k}})\right\rangle_{\widetilde{K}^{(\ell)}} + [(k-1) \text{ others}] \right\} \end{aligned} \tag{S23}$$

and

$$
\begin{aligned}
&\left\langle \left[\sigma(\mathrm{z}_{1,\alpha_1})\sigma(\mathrm{z}_{1,\alpha_2})\cdots\sigma(\mathrm{z}_{k,\alpha_{2k-1}})\sigma(\mathrm{z}_{k,\alpha_{2k}})\right]\mathcal{O}_V[\mathbf{z}]\right\rangle_{K^{(\ell)}} \qquad \text{(S24)}\\
=\;& \frac{1}{8}\sum_{\alpha_1',\alpha_2',\alpha_3',\alpha_4'} \widetilde{V}_{(\ell)}^{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}\\
&\times\Bigg\{\Big\langle\sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\Big(\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'}\tilde{\mathrm{z}}_{\alpha_3'}\tilde{\mathrm{z}}_{\alpha_4'}-2\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'}\widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'}-4\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_3'}\widetilde{K}^{(\ell)}_{\alpha_2'\alpha_4'}\\
&\qquad\qquad +\widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'}\widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'}+2\widetilde{K}^{(\ell)}_{\alpha_1'\alpha_3'}\widetilde{K}^{(\ell)}_{\alpha_2'\alpha_4'}\Big)\Big\rangle_{\widetilde{K}^{(\ell)}}\\
&\quad\times\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_3})\sigma(\tilde{\mathrm{z}}_{\alpha_4})\right\rangle_{\widetilde{K}^{(\ell)}}\cdots\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_{2k-1}})\sigma(\tilde{\mathrm{z}}_{\alpha_{2k}})\right\rangle_{\widetilde{K}^{(\ell)}}+[(k-1)\text{ others}]\Bigg\}\\
&+\frac{1}{4}\sum_{\alpha_1',\alpha_2',\alpha_3',\alpha_4'} \widetilde{V}_{(\ell)}^{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}\\
&\times\Bigg\{\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\left(\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'}-\widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'}\right)\right\rangle_{\widetilde{K}^{(\ell)}}\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_3})\sigma(\tilde{\mathrm{z}}_{\alpha_4})\left(\tilde{\mathrm{z}}_{\alpha_3'}\tilde{\mathrm{z}}_{\alpha_4'}-\widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'}\right)\right\rangle_{\widetilde{K}^{(\ell)}}\\
&\quad\times\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_5})\sigma(\tilde{\mathrm{z}}_{\alpha_6})\right\rangle_{\widetilde{K}^{(\ell)}}\cdots\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_{2k-1}})\sigma(\tilde{\mathrm{z}}_{\alpha_{2k}})\right\rangle_{\widetilde{K}^{(\ell)}}+\left[\binom{k}{2}-1\text{ others}\right]\Bigg\}
\end{aligned}
$$

As special cases, we obtain expressions advertised in the main text to be contained in this Appendix:

$$
\begin{aligned}
&\widetilde{A}^{(\ell)}_{\alpha_1\alpha_2}\equiv\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}H^{(\ell)}_{jj;\alpha_1\alpha_2} \qquad \text{(S25)}\\
=\;&\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\right\rangle_{\widetilde{K}^{(\ell)}}\\
&+\frac{1}{n_{\ell-1}}\Bigg[\frac{1}{2}\sum_{\alpha_1',\alpha_2'}S^{\alpha_1'\alpha_2'}_{(\ell)}\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})(\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'}-\widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'})\right\rangle_{\widetilde{K}^{(\ell)}}\\
&\qquad+\frac{1}{8}\sum_{\alpha_1',\alpha_2',\alpha_3',\alpha_4'}\widetilde{V}_{(\ell)}^{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}\left\langle\sigma(\tilde{\mathrm{z}}_{\alpha_1})\sigma(\tilde{\mathrm{z}}_{\alpha_2})\right.\\
&\qquad\times\Big(\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'}\tilde{\mathrm{z}}_{\alpha_3'}\tilde{\mathrm{z}}_{\alpha_4'}-2\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_2'}\widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'}-4\tilde{\mathrm{z}}_{\alpha_1'}\tilde{\mathrm{z}}_{\alpha_3'}\widetilde{K}^{(\ell)}_{\alpha_2'\alpha_4'}\\
&\qquad+\widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'}\widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'}+2\widetilde{K}^{(\ell)}_{\alpha_1'\alpha_3'}\widetilde{K}^{(\ell)}_{\alpha_2'\alpha_4'}\Big)\Big\rangle_{\widetilde{K}^{(\ell)}}\Bigg]+O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

and

$$\widetilde{B}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \equiv \frac{1}{n_\ell^2}\sum_{j_1,j_2=1}^{n_\ell}\left[H^{(\ell)}_{j_1j_1j_2j_2;\alpha_1\alpha_2\alpha_3\alpha_4} - H^{(\ell)}_{j_1j_1;\alpha_1\alpha_2}H^{(\ell)}_{j_2j_2;\alpha_3\alpha_4}\right] \tag{S26}$$

$$= \frac{1}{n_\ell}\Bigg\{\langle\sigma(\tilde{z}_{\alpha_1})\sigma(\tilde{z}_{\alpha_2})\sigma(\tilde{z}_{\alpha_3})\sigma(\tilde{z}_{\alpha_4})\rangle_{\widetilde{K}^{(\ell)}} - \langle\sigma(\tilde{z}_{\alpha_1})\sigma(\tilde{z}_{\alpha_2})\rangle_{\widetilde{K}^{(\ell)}}\langle\sigma(\tilde{z}_{\alpha_3})\sigma(\tilde{z}_{\alpha_4})\rangle_{\widetilde{K}^{(\ell)}}$$
$$+\frac{1}{4}\left(\frac{n_\ell}{n_{\ell-1}}\right)\sum_{\alpha_1',\alpha_2',\alpha_3',\alpha_4'}\widetilde{V}_{(\ell)}^{(\alpha_1'\alpha_2')(\alpha_3'\alpha_4')}\left\langle\sigma(\tilde{z}_{\alpha_1})\sigma(\tilde{z}_{\alpha_2})(\tilde{z}_{\alpha_1'}\tilde{z}_{\alpha_2'} - \widetilde{K}^{(\ell)}_{\alpha_1'\alpha_2'})\right\rangle_{\widetilde{K}^{(\ell)}}$$
$$\times\left\langle\sigma(\tilde{z}_{\alpha_3})\sigma(\tilde{z}_{\alpha_4})(\tilde{z}_{\alpha_3'}\tilde{z}_{\alpha_4'} - \widetilde{K}^{(\ell)}_{\alpha_3'\alpha_4'})\right\rangle_{\widetilde{K}^{(\ell)}}\Bigg\} + O\left(\frac{1}{n^2}\right).$$

Assembling everything,

$$\begin{aligned}
& G^{(\ell+1)}_{i_1\ldots i_{2m};\alpha_1\ldots\alpha_{2m}} \\
= \; & \delta_{i_1i_2}\cdots\delta_{i_{2m-1}i_{2m}}\prod_{k=1}^{m}\left[C_b^{(\ell+1)} + C_W^{(\ell+1)}\widetilde{A}^{(\ell)}_{\alpha_{2k-1}\alpha_{2k}}\right] \\
& + [(2m-1)!! - 1 \text{ other pairings}] \\
& + \delta_{i_1i_2}\cdots\delta_{i_{2m-1}i_{2m}}\widetilde{B}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}\prod_{k=3}^{m}\left[C_b^{(\ell+1)} + C_W^{(\ell+1)}\widetilde{A}^{(\ell)}_{\alpha_{2k-1}\alpha_{2k}}\right] \\
& + \left\{\left[3\times\binom{2m}{4}\times(2m-5)!! - 1\right] \text{ other } (4,2,2,\ldots,2) \text{ clusterings}\right\} \\
& + O\left(\frac{1}{n^2}\right).
\end{aligned} \tag{S27}$$

In particular,

$$G^{(\ell+1)}_{i_1i_2;\alpha_1\alpha_2} = \delta_{i_1i_2}\left[C_b^{(\ell+1)} + C_W^{(\ell+1)}\widetilde{A}^{(\ell)}_{\alpha_1\alpha_2}\right] + O\left(\frac{1}{n^2}\right), \tag{S28}$$

$$\begin{aligned}
G^{(\ell+1)}_{i_1i_2i_3i_4;\alpha_1\alpha_2\alpha_3\alpha_4}\Big|_{\text{connected}} &= \delta_{i_1i_2}\delta_{i_3i_4}\widetilde{B}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1i_3}\delta_{i_2i_4}\widetilde{B}^{(\ell)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} \\
&\quad + \delta_{i_1i_4}\delta_{i_2i_3}\widetilde{B}^{(\ell)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)} + O\left(\frac{1}{n^2}\right), \quad \text{and}
\end{aligned}$$

$$G^{(\ell+1)}_{i_1i_2\ldots i_{2m-1}i_{2m};\alpha_1\alpha_2\ldots\alpha_{2m-1}\alpha_{2m}}\Big|_{\text{connected}} = O\left(\frac{1}{n^2}\right), \quad \text{for} \quad 2m \geq 6\,. \tag{S29}$$

completing our inductive proof. Note that $\widetilde{B}^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = O\left(\frac{1}{n}\right)$.

Nowhere in our derivation had we assumed anything about the form of activation functions. The only potential exceptions to our formalism are exponentially growing activation functions – which we never see in practice – that would make the Gaussian integrals unintegrable.

Appendix C. Bestiary of concrete examples

C.1. Quadratic activation

Let us take multilayer perceptrons with quadratic activation, $\sigma(z) = z^2$, and study the distributions of preactivations in the second layer as another illustration of our technology. From the master recursion relations (R1-R3) with the initial condition (R0), Wick's contractions yield

$$\widetilde{K}^{(2)}_{\alpha_1\alpha_2} = C_b^{(2)} + C_W^{(2)}\left[\widetilde{K}^{(1)}_{\alpha_1\alpha_1}\widetilde{K}^{(1)}_{\alpha_2\alpha_2} + 2\widetilde{K}^{(1)}_{\alpha_1\alpha_2}\widetilde{K}^{(1)}_{\alpha_1\alpha_2}\right], \tag{S30}$$

$$\begin{aligned}\frac{\widetilde{V}^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}}{\left[C_W^{(2)}\right]^2} =& 2\Big[\widetilde{K}^{(1)}_{\alpha_1\alpha_1}\widetilde{K}^{(1)}_{\alpha_3\alpha_3}\left(\widetilde{K}^{(1)}_{\alpha_2\alpha_4}\right)^2 + \widetilde{K}^{(1)}_{\alpha_1\alpha_1}\widetilde{K}^{(1)}_{\alpha_4\alpha_4}\left(\widetilde{K}^{(1)}_{\alpha_2\alpha_3}\right)^2 \\ &+ \widetilde{K}^{(1)}_{\alpha_2\alpha_2}\widetilde{K}^{(1)}_{\alpha_3\alpha_3}\left(\widetilde{K}^{(1)}_{\alpha_1\alpha_4}\right)^2 + \widetilde{K}^{(1)}_{\alpha_2\alpha_2}\widetilde{K}^{(1)}_{\alpha_4\alpha_4}\left(\widetilde{K}^{(1)}_{\alpha_1\alpha_3}\right)^2\Big] \\ &+ 4\left[\left(\widetilde{K}^{(1)}_{\alpha_1\alpha_3}\right)^2\left(\widetilde{K}^{(1)}_{\alpha_2\alpha_4}\right)^2 + \left(\widetilde{K}^{(1)}_{\alpha_1\alpha_4}\right)^2\left(\widetilde{K}^{(1)}_{\alpha_2\alpha_3}\right)^2\right] \\ &+ 8\Big[\widetilde{K}^{(1)}_{\alpha_1\alpha_1}\widetilde{K}^{(1)}_{\alpha_2\alpha_3}\widetilde{K}^{(1)}_{\alpha_3\alpha_4}\widetilde{K}^{(1)}_{\alpha_4\alpha_2} + \widetilde{K}^{(1)}_{\alpha_2\alpha_2}\widetilde{K}^{(1)}_{\alpha_3\alpha_4}\widetilde{K}^{(1)}_{\alpha_4\alpha_1}\widetilde{K}^{(1)}_{\alpha_1\alpha_3} \\ &\quad + \widetilde{K}^{(1)}_{\alpha_3\alpha_3}\widetilde{K}^{(1)}_{\alpha_4\alpha_1}\widetilde{K}^{(1)}_{\alpha_1\alpha_2}\widetilde{K}^{(1)}_{\alpha_2\alpha_4} + \widetilde{K}^{(1)}_{\alpha_4\alpha_4}\widetilde{K}^{(1)}_{\alpha_1\alpha_2}\widetilde{K}^{(1)}_{\alpha_2\alpha_3}\widetilde{K}^{(1)}_{\alpha_3\alpha_1}\Big] \\ &+ 16\left[\widetilde{K}^{(1)}_{\alpha_1\alpha_2}\widetilde{K}^{(1)}_{\alpha_1\alpha_3}\widetilde{K}^{(1)}_{\alpha_2\alpha_4}\widetilde{K}^{(1)}_{\alpha_3\alpha_4} + \widetilde{K}^{(1)}_{\alpha_1\alpha_2}\widetilde{K}^{(1)}_{\alpha_1\alpha_4}\widetilde{K}^{(1)}_{\alpha_2\alpha_3}\widetilde{K}^{(1)}_{\alpha_3\alpha_4}\right] \\ &+ 16\widetilde{K}^{(1)}_{\alpha_1\alpha_3}\widetilde{K}^{(1)}_{\alpha_1\alpha_4}\widetilde{K}^{(1)}_{\alpha_2\alpha_3}\widetilde{K}^{(1)}_{\alpha_2\alpha_4}, \quad \text{and}\end{aligned} \tag{S31}$$

$$\widetilde{S}^{(2)}_{\alpha_1\alpha_2} = 0\,. \tag{S32}$$

where $\widetilde{K}^{(1)}_{\alpha_1\alpha_2} = C_b^{(1)} + C_W^{(1)} \cdot \left(\frac{\boldsymbol{x}_{\alpha_1}\cdot\boldsymbol{x}_{\alpha_2}}{n_0}\right)$. These expressions are used in the main text for the experimental study of finite-width corrections on Bayesian inference.

C.2. Details for single-input cases

The recursive relations simplify drastically for the case of a single input, $N_{\rm D} = 1$. Setting $C_b^{(\ell)} = 0$ for simplicity and dropping α index, our recursive equations reduce to

$$\widetilde{K}^{(\ell+1)} = C_W^{(\ell+1)} \left\langle [\sigma(\tilde{\mathrm{z}})]^2 \right\rangle_{\widetilde{K}^{(\ell)}} , \tag{S33}$$

$$\frac{\widetilde{V}^{(\ell+1)}}{\left(\widetilde{K}^{(\ell+1)}\right)^2} = \left(\frac{\left\langle [\sigma(\tilde{\mathrm{z}})]^4 \right\rangle_{\widetilde{K}^{(\ell)}}}{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \right\rangle^2_{\widetilde{K}^{(\ell)}}} - 1 \right) + \frac{1}{4} \left(\frac{n_\ell}{n_{\ell-1}} \right) \left(\frac{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \, \tilde{\mathrm{z}}^2 \right\rangle_{\widetilde{K}^{(\ell)}}}{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \right\rangle_{\widetilde{K}^{(\ell)}} \widetilde{K}^{(\ell)}} - 1 \right)^2 \cdot \frac{\widetilde{V}^{(\ell)}}{\left(\widetilde{K}^{(\ell)}\right)^2} , \text{ and} \tag{S34}$$

$$\begin{aligned} \frac{\widetilde{S}^{(\ell+1)}}{\widetilde{K}^{(\ell+1)}} =& \frac{1}{2} \left(\frac{n_\ell}{n_{\ell-1}} \right) \left(\frac{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \, \tilde{\mathrm{z}}^2 \right\rangle_{\widetilde{K}^{(\ell)}}}{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \right\rangle_{\widetilde{K}^{(\ell)}} \widetilde{K}^{(\ell)}} - 1 \right) \cdot \frac{\widetilde{S}^{(\ell)}}{\widetilde{K}^{(\ell)}} \\ &+ \frac{1}{8} \left(\frac{n_\ell}{n_{\ell-1}} \right) \left(\frac{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \, \tilde{\mathrm{z}}^4 \right\rangle_{\widetilde{K}^{(\ell)}}}{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \right\rangle_{\widetilde{K}^{(\ell)}} \left(\widetilde{K}^{(\ell)}\right)^2} - 6 \frac{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \, \tilde{\mathrm{z}}^2 \right\rangle_{\widetilde{K}^{(\ell)}}}{\left\langle [\sigma(\tilde{\mathrm{z}})]^2 \right\rangle_{\widetilde{K}^{(\ell)}} \widetilde{K}^{(\ell)}} + 3 \right) \cdot \frac{\widetilde{V}^{(\ell)}}{\left(\widetilde{K}^{(\ell)}\right)^2} . \end{aligned} \tag{S35}$$

C.2.1. MONOMIALS WITH SINGLE INPUT

For monomial activations, $\sigma(z) = z^p$, such as in deep linear networks (Saxe et al., 2013) and quadratic activations (Li et al., 2018),

$$\widetilde{K}^{(\ell+1)} = \left[(2p-1)!! C_W^{(\ell+1)} \right] \left(\widetilde{K}^{(\ell)} \right)^p , \tag{S36}$$

$$\frac{\widetilde{V}^{(\ell+1)}}{\left(\widetilde{K}^{(\ell+1)}\right)^2} = \left\{ \frac{(4p-1)!!}{[(2p-1)!!]^2} - 1 \right\} + p^2 \left(\frac{n_\ell}{n_{\ell-1}} \right) \frac{\widetilde{V}^{(\ell)}}{\left(\widetilde{K}^{(\ell)}\right)^2} , \quad \text{and} \tag{S37}$$

$$\frac{\widetilde{S}^{(\ell+1)}}{\widetilde{K}^{(\ell+1)}} = \left(\frac{n_\ell}{n_{\ell-1}} \right) \left[p \frac{\widetilde{S}^{(\ell)}}{\widetilde{K}^{(\ell)}} + \frac{p(p-1)}{2} \frac{\widetilde{V}^{(\ell)}}{\left(\widetilde{K}^{(\ell)}\right)^2} \right] . \tag{S38}$$

In particular the four-point vertex solution is given by

$$\frac{1}{n_{\ell-1} p^{2(\ell-1)}} \frac{\widetilde{V}^{(\ell)}}{\left(\widetilde{K}^{(\ell)}\right)^2} = \left\{ \frac{(4p-1)!!}{[(2p-1)!!]^2} - 1 \right\} \left(\sum_{\ell'=1}^{\ell-1} \frac{1}{n_{\ell'} p^{2\ell'}} \right) . \tag{S39}$$

The factor $\left(\sum_{\ell'} \frac{1}{n_{\ell'} p^{2\ell'}} \right)$ generalizes the factor $\left(\sum_{\ell'} \frac{1}{n_{\ell'}} \right)$ for linear and ReLU activations. Following Hanin and Rolnick (2018), this factor guides us to narrow hidden layers as we pass through nonlinear activations for $p > 1$.

C.2.2. RELU WITH SINGLE INPUT

ReLU activation, $\sigma(z) = \max(0, z)$, can also be worked out for a single input through Wick's contractions, noting that the Gaussian integral is halved, yielding

$$\widetilde{K}^{(\ell+1)} = \left[\frac{C_W^{(\ell+1)}}{2}\right] \widetilde{K}^{(\ell)}\,, \tag{S40}$$

$$\frac{\widetilde{V}^{(\ell+1)}}{\left(\widetilde{K}^{(\ell+1)}\right)^2} = 5 + \left(\frac{n_\ell}{n_{\ell-1}}\right) \frac{\widetilde{V}^{(\ell)}}{\left(\widetilde{K}^{(\ell)}\right)^2}\,, \quad \text{and} \tag{S41}$$

$$\frac{\widetilde{S}^{(\ell+1)}}{\widetilde{K}^{(\ell+1)}} = \left(\frac{n_\ell}{n_{\ell-1}}\right) \frac{\widetilde{S}^{(\ell)}}{\widetilde{K}^{(\ell)}}\,. \tag{S42}$$

Setting $C_W^{(\ell)} = 2$ for simplicity, these equations can be solved, leading to

$$\widetilde{K}^{(\ell)} = \widetilde{K}^{(1)} = \frac{||\boldsymbol{x}||_2^2}{n_0}\,, \tag{S43}$$

$$\frac{1}{n_{\ell-1}} \widetilde{V}^{(\ell)} = 5 \left(\sum_{\ell'=1}^{\ell-1} \frac{1}{n_{\ell'}}\right) \left(\widetilde{K}^{(1)}\right)^2\,, \quad \text{and} \tag{S44}$$

$$\widetilde{S}^{(\ell)} = 0\,. \tag{S45}$$

C.3. More experiments on output distributions

Here is an extended version of experiments in Section 4.3. As in the main text, take a single black-white image of hand-written digits from the MNIST dataset as an $n_0 = 784$-dimensional input, without preprocessing. Set bias variance $C_b^{(\ell)} = 0$, weight variance $C_W^{(\ell)} = C_W$, and use activations $\sigma(z) = z$ (linear) with $C_W = 1$, $\sigma(z) = z^2$ (quadratic) with $C_W = \frac{1}{3}$, and $\sigma(z) = \max(0, z)$ (ReLU) with $C_W = 2$. For all three cases, we consider both depth $L = 2$ with widths $(n_0, n_1, n_2) = (784, n, 1)$ and depth $L = 3$ with widths $(n_0, n_1, n_2, n_3) = (784, n, 2n, 1)$. As in Figure 1, in Figure S1, for each width-parameter n of the hidden layers we record the prior distribution of outputs over 10^6 instances of Gaussian weights and compare it with the theoretical prediction. Results again corroborate our theory.

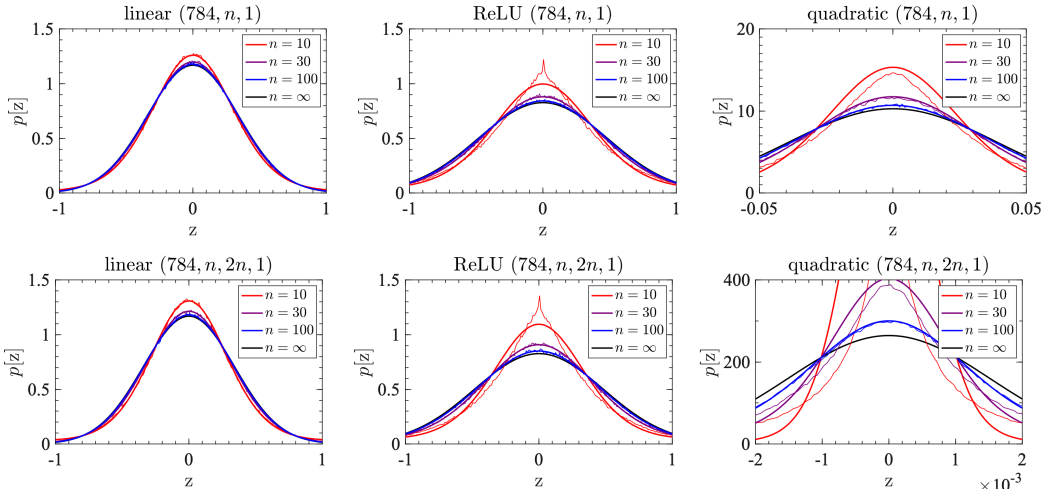

Figure S1: Comparison between theory and experiments for prior distributions of outputs for a single input. Our theoretical predictions (smooth thick lines) and experimental data (rugged thin lines) agree, correctly capturing the initial deviations from the Gaussian processes (black, $n = \infty$), at least down to $n = n_\star$ with $n_\star \sim 10$ for linear cases, $n_\star \sim 30$ for ReLU cases and depth $L = 2$ quadratic case, and $n_\star \sim 100$ for depth $L = 3$ quadratic case. This also illustrates that nonlinear activations quickly amplify non-Gaussianity.

Appendix D. Finite-width corrections on Bayesian inference

In order to massage Equation (NGPM) into an actionable form, first playing with the metric inversions and defining $\overline{\phi}_i^{\ \alpha} \equiv \sum_{\alpha'} (\widetilde{K}^{-1})^{\alpha\alpha'} \overline{\phi}_{i;\alpha'}$, the mean prediction becomes

$$\begin{aligned}
&\left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} + \epsilon \sum_{\alpha_1,\dot{\gamma}_1,\alpha_0} \left(\widetilde{K}_\Delta\right)_{\dot{\gamma}\dot{\gamma}_1} \overline{\phi}_i^{\ \alpha_1} \left(\widetilde{K}^{-1}\right)^{\dot{\gamma}_1\alpha_0} \\
&\quad \times \left\{ \widetilde{S}^{(L)}_{\alpha_0\alpha_1} - \sum_{\alpha_2,\alpha_3} \left[\widetilde{V}^{(L)}_{(\alpha_0\alpha_2)(\alpha_1\alpha_3)} + \frac{n_L}{2} \widetilde{V}^{(L)}_{(\alpha_0\alpha_1)(\alpha_2\alpha_3)} \right] \right. \\
&\qquad \times \left[\left(\widetilde{K}^{-1}\right)^{\alpha_2\alpha_3} - \left(\widetilde{K}^{-1}\right)^{\alpha_2\dot{\gamma}_2} \left(\widetilde{K}_\Delta\right)_{\dot{\gamma}_2\dot{\gamma}_3} \left(\widetilde{K}^{-1}\right)^{\dot{\gamma}_3\alpha_3} \right] \\
&\qquad \left. + \frac{1}{2} \sum_{\alpha_2,\alpha_3} \widetilde{V}^{(L)}_{(\alpha_0\alpha_1)(\alpha_2\alpha_3)} \left(\sum_j \overline{\phi}_j^{\ \alpha_2} \overline{\phi}_j^{\ \alpha_3} \right) \right\}.
\end{aligned} \tag{S46}$$

This expression simplifies drastically through the identity

$$\begin{pmatrix} \widetilde{K}_{\mathrm{RR}} & \widetilde{K}_{\mathrm{RE}} \\ \widetilde{K}_{\mathrm{ER}} & \widetilde{K}_{\mathrm{EE}} \end{pmatrix}^{-1} = \begin{pmatrix} \widetilde{K}_{\mathrm{RR}}^{-1} + \widetilde{K}_{\mathrm{RR}}^{-1} \widetilde{K}_{\mathrm{RE}} \widetilde{K}_\Delta^{-1} \widetilde{K}_{\mathrm{ER}} \widetilde{K}_{\mathrm{RR}}^{-1} & -\widetilde{K}_{\mathrm{RR}}^{-1} \widetilde{K}_{\mathrm{RE}} \widetilde{K}_\Delta^{-1} \\ -\widetilde{K}_\Delta^{-1} \widetilde{K}_{\mathrm{ER}} \widetilde{K}_{\mathrm{RR}}^{-1} & \widetilde{K}_\Delta^{-1} \end{pmatrix}, \tag{S47}$$

which can be checked explicitly, recalling $\widetilde{K}_\Delta \equiv \widetilde{K}_{\mathrm{EE}} - \widetilde{K}_{\mathrm{ER}} \widetilde{K}_{\mathrm{RR}}^{-1} \widetilde{K}_{\mathrm{RE}}$. Incidentally, this identity can also be used to prove Equation (GPΔ). Now equipped with this identity, recalling $\overline{\phi}_{i;\alpha} \equiv [(y_{\mathrm{R}})_{i;\bar{\beta}}, \left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}}]$, we notice that $\overline{\phi}_i^{\ \bar{\beta}} = \sum_{\bar{\beta}'} \left(\widetilde{K}_{\mathrm{RR}}^{-1}\right)^{\bar{\beta}\bar{\beta}'} (y_{\mathrm{R}})_{i;\bar{\beta}'}$ and $\overline{\phi}_i^{\ \dot{\gamma}_1} = 0$. Similarly

$$\left[\left(\widetilde{K}^{-1}\right)^{\bar{\beta}_2\bar{\beta}_3} - \sum_{\dot{\gamma}_2,\dot{\gamma}_3} \left(\widetilde{K}^{-1}\right)^{\bar{\beta}_2\dot{\gamma}_2} \left(\widetilde{K}_\Delta\right)_{\dot{\gamma}_2\dot{\gamma}_3} \left(\widetilde{K}^{-1}\right)^{\dot{\gamma}_3\bar{\beta}_3} \right] = \left(\widetilde{K}_{\mathrm{RR}}^{-1}\right)^{\bar{\beta}_2\bar{\beta}_3}$$

and other components [i.e. with one or both of training components $(\bar{\beta}_2, \bar{\beta}_3)$ replaced by test components $\dot{\gamma}$] vanish. Equation (S46) thus simplifies to

$$\begin{aligned}
&\left(\mathrm{y}_{\mathrm{E}}^{\mathrm{GP}}\right)_{i;\dot{\gamma}} + \epsilon \sum_{\bar{\beta}_1,\dot{\gamma}_1,\alpha_0} \left(\widetilde{K}_\Delta\right)_{\dot{\gamma}\dot{\gamma}_1} \overline{\phi}_i^{\ \bar{\beta}_1} \left(\widetilde{K}^{-1}\right)^{\dot{\gamma}_1\alpha_0} \left[\widetilde{S}^{(L)}_{\alpha_0\bar{\beta}_1} + \frac{1}{2} \sum_{\bar{\beta}_2,\bar{\beta}_3} \widetilde{V}^{(L)}_{(\alpha_0\bar{\beta}_1)(\bar{\beta}_2\bar{\beta}_3)} \left(\sum_j \overline{\phi}_j^{\ \bar{\beta}_2} \overline{\phi}_j^{\ \bar{\beta}_3} \right) \right. \\
&\qquad \left. - \sum_{\bar{\beta}_2,\bar{\beta}_3} \left(\widetilde{V}^{(L)}_{(\alpha_0\bar{\beta}_2)(\bar{\beta}_1\bar{\beta}_3)} + \frac{n_L}{2} \widetilde{V}^{(L)}_{(\alpha_0\bar{\beta}_1)(\bar{\beta}_2\bar{\beta}_3)} \right) \left(\widetilde{K}_{\mathrm{RR}}^{-1}\right)^{\bar{\beta}_2\bar{\beta}_3} \right]
\end{aligned} \tag{S48}$$

Finally, denoting the matrix inside the parenthesis to be

$$\begin{aligned}
A_{\alpha_0\bar{\beta}_1} \equiv & S^{(L)}_{\alpha_0\bar{\beta}_1} + \frac{1}{2} \sum_{\bar{\beta}_2,\bar{\beta}_3,\bar{\beta}_2',\bar{\beta}_3'} \widetilde{V}^{(L)}_{(\alpha_0\bar{\beta}_1)(\bar{\beta}_2\bar{\beta}_3)} \left(\sum_j \overline{\phi}_j^{\ \bar{\beta}_2} \overline{\phi}_j^{\ \bar{\beta}_3} \right) \\
& - \sum_{\bar{\beta}_2,\bar{\beta}_3} \left(\widetilde{V}^{(L)}_{(\alpha_0\bar{\beta}_2)(\bar{\beta}_1\bar{\beta}_3)} + \frac{n_L}{2} \widetilde{V}^{(L)}_{(\alpha_0\bar{\beta}_1)(\bar{\beta}_2\bar{\beta}_3)} \right) \left(\widetilde{K}_{\mathrm{RR}}^{-1}\right)^{\bar{\beta}_2\bar{\beta}_3},
\end{aligned} \tag{NGPM'}$$

and noticing $\sum_{\dot\gamma_1}\left(\widetilde{K}_\Delta\right)_{\dot\gamma\dot\gamma_1}\left(\widetilde{K}^{-1}\right)^{\dot\gamma_1\bar\beta_0} = -\left(\widetilde{K}_{\text{ER}}\widetilde{K}_{\text{RR}}^{-1}\right)_{\dot\gamma}{}^{\bar\beta_0}$ and $\sum_{\dot\gamma_1}\left(\widetilde{K}_\Delta\right)_{\dot\gamma\dot\gamma_1}\left(\widetilde{K}^{-1}\right)^{\dot\gamma_1\dot\gamma_0} = \delta_{\dot\gamma}^{\dot\gamma_0}$,

$$\left(\mathrm{y}_{\text{E}}^{\text{GP}}\right)_{i;\dot\gamma} + \epsilon\sum_{\bar\beta_1}\overline{\phi}_i^{\bar\beta_1}\left[A_{\dot\gamma\bar\beta_1} - \sum_{\bar\beta_0}\left(\widetilde{K}_{\text{ER}}\widetilde{K}_{\text{RR}}^{-1}\right)_{\dot\gamma}{}^{\bar\beta_0}A_{\bar\beta_0\bar\beta_1}\right] \tag{NGPM''}$$

is the mean prediction. Equations (NGPM') and (NGPM'') with $\overline{\phi}_i^{\bar\beta} = \sum_{\bar\beta'}\left(\widetilde{K}_{\text{RR}}^{-1}\right)^{\bar\beta\bar\beta'}\left(y_{\text{R}}\right)_{i;\bar\beta'}$ are actionable, i.e., easy to program.