Journal of Machine Learning Research 25 (2024) 1-58 Submitted 5/23; Revised 6/24; Published 8/24

Random Fully Connected Neural Networks as Perturbatively Solvable Hierarchies

Boris Hanin BHANIN@PRINCETON.EDU
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Editor: Joan Bruna

Abstract

We study the distribution of fully connected neural networks with Gaussian random weights/biases and L hidden layers, each of width proportional to a large parameter n. For polynomially bounded non-linearities we give sharp estimates in powers of $1/n$ for the joint cumulants of the network output and its derivatives. We further show that network cumulants form a perturbatively solvable hierarchy in powers of $1/n$. That is, the k-th order cumulants in each layer are determined to leading order in $1/n$ by cumulants of order at most k computed at the previous layer. By explicitly deriving and then solving several such recursions, we find that the depth-to-width ratio L/n plays the role of an effective network depth, controlling both the distance to Gaussianity and the size of inter-neuron correlations.

Keywords: Deep Learning, Neural Networks, Finite Width Corrections, Cumulants, Quantitative CLT

1. Introduction

We live in an era of big data and cheap computation. This has led to remarkable progress in domains ranging from self-driving cars (Krizhevsky et al. (2012)) to automatic drug discovery (Jumper et al. (2021)) and machine translation (Brown et al. (2020)). Underpinning many of these exciting practical developments is a class of computational models called neural networks. While they were originally developed in the 1940's and 1950's (see e.g. Hebb (1949); Rosenblatt (1958)), the complexity of state-of-the-art neural nets is unprecedented.

The undeniable empirical utility of modern neural networks has led over the past decade or so to significant interest in principled theoretical approaches to understanding deep learning (e.g. Bartlett et al. (2021); Belkin (2021); Jacot et al. (2018); Kawaguchi (2016); Roberts et al. (2022)). One of the most well-developed lines of such work focuses on analyzing networks in the so-called NTK regime. As we explain more fully in §2.1, the NTK regime captures the structure of neural networks asymptotically when the network depth and training set size are held fixed, while the hidden layer widths tend to infinity.

In the NTK regime neural networks are surprisingly simple. At the start of training, when the network parameters are chosen at random, the network output converges to a Gaussian process. The limiting covariance function, moreover, satisfies an explicit recursion

with respect to network depth (see Theorem 2 and Lee et al. (2018); Matthews et al. (2018); Neal (1996); Novak et al. (2018); Yang (2019a,b); Yang and Schoenholz (2018)). Further, when network parameters are optimized by using gradient descent to minimize a mean squared error objective, the full optimization trajectory coincides with that of the network's linearization around the start of training (see Bartlett et al. (2021); Chizat and Bach (2018); Du et al. (2018); Jacot et al. (2018); Lee et al. (2019); Liu et al. (2022); Roberts et al. (2022)). Neural networks in the NTK regime are thus equivalent to linear models.

While the NTK regime sheds light on the empirical success of optimization in modern deep learning, it is therefore too rigid to capture the ability of real work networks to learn data-dependent features (see e.g. Hanin and Nica (2020a); Huang and Yau (2020); Roberts et al. (2022); Yang and Hu (2021)). Since it is precisely such feature learning that is believed to be crucial for understanding why neural networks generalize well in practice, it is imperative to understand neural networks beyond the NTK regime. Prior work has approached this in several ways:

- **Mean field instead of NTK initialization.** The simplification of neural networks in the NTK regime alluded to above depends on initializing networks in the specific way typically done in practice. A range of articles such as Bordelon and Pehlevan (2023); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2020); Woodworth et al. (2020); Yang et al. (2022) point out that alternative mean-field initialization schemes can lead to feature learning and non-linear training dynamics, even at infinite width.

- **Growing dataset size.** Even with the NTK initialization, if the size of the training dataset and network width tend to infinity together, neural networks need not become linear models (e.g. Cui et al. (2023); Hanin and Zlokapa (2022); Li and Sompolinsky (2021); Seroussi and Ringel (2021)). Characterizing the simultaneous limit of wide networks training on growing dataset sizes remains an important open question.

- **Large learning rates.** A key requirement for the asymptotic linearization of neural network training in the NTK regime is that learning rates (i.e. step sizes used in gradient descent) should tend to zero as the network width tends to infinity. Articles such as Lewkowycz et al. (2020) and Zhu et al. (2022), however, show that larger learning rates lead to non-NTK behavior.

- **Finite width corrections to the NTK regime.** Neural networks at large but finite width are neither Gaussian processes at initialization nor linear models during training (see Hanin (2018); Hanin and Paouris (2021); Hanin and Nica (2020b,a); Hanin and Zlokapa (2022); Huang and Yau (2020); Roberts et al. (2022); Yaida (2020)). A key message of these articles is that depth amplifies finite width effects, including feature learning. This setting is the focus of the present article.

A key difficulty in the finite width approaches of the last bullet point is that, at initialization, the distribution of network outputs is both non-Gaussian and involves complicated correlations between neurons. The purpose of the present article is to develop, in simplest important setting of fully connected networks (see §1.1 for the definition), a flexible set of probabilistic tools to characterize these non-Gaussian effects. We summarize our main

contributions in §1.2. Before doing so, we introduce in the next section some notation and a precise statement of the problem statement.

1.1 Notation and Problem Statement

By definition, a fully connected network is a map that associates to each network input $x_\alpha \in \mathbb{R}^{n_0}$ an output $z_\alpha^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ through a sequence of intermediate representations $z_\alpha^{(\ell)} \in \mathbb{R}^{n_\ell}$

$$z_\alpha^{(\ell+1)} := \begin{cases} b^{(\ell+1)} + W^{(\ell+1)}\sigma(z_\alpha^{(\ell)}), & \ell \geq 1 \\ b^{(1)} + W^{(1)}x_\alpha, & \ell = 0 \end{cases}, \tag{1}$$

where

$$W^{(\ell+1)} \in \mathbb{R}^{n_{\ell+1}\times n_\ell},\, b^{(\ell+1)} \in \mathbb{R}^{n_{\ell+1}}.$$

In the recursion (1), the univariate function $\sigma : \mathbb{R} \to \mathbb{R}$ applied to a vector $z_\alpha^{(\ell)} \in \mathbb{R}^{n_\ell}$ is short-hand for applying it separately to each component and the widths $n_0, \ldots, n_{L+1}$ are known a priori. The entries of the matrices $W^{(\ell)}$ and the components of the vectors $b^{(\ell)}$ are called the *weights and biases* in layer ℓ, respectively. We shall typically refer to

$$z_\alpha^{(\ell)} = \left(z_{1;\alpha}^{(\ell)}, \ldots, z_{n_\ell;\alpha}^{(\ell)}\right) \in \mathbb{R}^{n_\ell}$$

as the *vector of pre-activations at layer* ℓ corresponding to the input x_α. The most popular choices of σ in practice include $\mathrm{ReLU}(t) := \max\{0, t\}$ as well the hyperbolic tangent and their variations. We will analyze these cases in detail later (see §B.2), but for our general results make only the following mild assumption

Assumption 1 *There exists $r \geq 1$ so that the r-th derivative of σ exists almost everywhere and grows at most polynomially:*

$$\exists k \geq 1 \ s.t. \ \left|\left|(1+|x|)^{-k}\frac{d^r}{dx^r}\sigma(x)\right|\right|_{L^\infty(\mathbb{R},dx)} < \infty.$$

The primary objects of study in this article are *random fully connected neural networks*, obtained by choosing network weights and biases of a fully connected network to be independent centered Gaussians:

$$W_{ij}^{(\ell)} \sim \mathcal{N}(0, C_W/n_{\ell-1}), \qquad b_i^{(\ell)} \sim \mathcal{N}(0, C_b) \qquad \text{independent.} \tag{2}$$

Here $C_b \geq 0, C_W > 0$ are fixed constants. The $1/n_{\ell-1}$ scaling in the weight variance ensures that the moments of the outputs $z_\alpha^{(L+1)}$ remain uniformly bounded as $n_1, \ldots, n_L, \to \infty$ (see e.g. Theorem 2). The distribution (2) is used in practice to initialize neural networks at the start of training (e.g. by gradient descent on an empirical loss)[1].

The main problem we take up in the present article is to characterize the joint distribution of any finite number of components in a random neural network $x_\alpha \mapsto z_\alpha^{(L+1)}$ evaluated

1. While (2) is indeed the default initialization scheme mainly used in practice, there exists an important alternative often referred to as a mean-field initialization in which the final layer weights $W^{(L+1)}$ have a much smaller variance Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Yang et al. (2022). In this context, our analysis applies directly to pre-activations $z_{i;\alpha}^{(\ell)}$ in hidden layers with $\ell \leq L$.

at potentially different inputs in the regime where the input dimension n_0 is arbitrary, the inputs x_α satisfy

$$\frac{1}{n_0}\,||x_\alpha||_2^2 < \infty,$$

the output dimension n_{L+1} is fixed, and the hidden layer widths n_ℓ are taken large but finite:

$$\exists c, C > 0 \text{ s.t.} \qquad cn \leq n_1, \ldots, n_L \leq Cn, \qquad\qquad n \gg 1. \tag{3}$$

Our approach will be to describe the random field $x_\alpha \mapsto z_\alpha^{(L+1)}$ perturbatively (i.e. as a power series) in $1/n$ and recursively in L.

1.2 Main Contributions

The main results/contributions of the present article are:

- **Sharp Estimates for Cumulants at Finite Width.** At any fixed finite depth L and large width n, we prove that the k-th cumulant of a random neural network vanishes if k is odd and tends to zero like $n^{-k/2+1}$ at large n when k is even. See Theorem 3 and Corollary 6. This estimate is sharp in terms of its dependence on n and can be viewed as a quantitative Central Limit Theorem for the finite-dimensional distributions in wide neural networks (see Theorem 2).

- **Hierarchy of Layer-wise Cumulant Recursions.** At any fixed finite depth L and large width n, we prove for any $\ell = 1, \ldots, L$ that in a random neural network the k-th cumulants of pre-activations $z_\alpha^{(\ell+1)}$ in layer $\ell + 1$ are determined to leading order in $1/n$ by the cumulants of order at most k in layer ℓ. In this way, the distribution of a random neural network forms a *perturbatively solvable hierarchy* which extends the infinite width covariance recursion from Theorem 2. See Theorem 4 and Corollary 5. This is similar in spirit, though with a rather different focus, to the breakthrough work Huang and Yau (2020), which provides a hierarchy in powers of $1/n$ for the dynamics of the NTK during network training.

- **Emergence of Effective Network Depth.** By solving explicitly cumulant recursions for the $k = 4^{th}, 6^{th}, 8^{th}$ cumulants we observe a remarkable phenomenon. Namely, while the k-th cumulant goes to zero like $n^{-k/2+1}$, we also find that it grows like $L^{k/2-1}$ at large depth L. Taken together, this shows that random neural networks that are both deep and wide are not close to Gaussian processes, with depth amplifying finite width effects. Specifically, it is the *effective network depth*, given by the depth-to-width ratio L/n, rather than the apparent depth L that is a more informative measure of neural network depth and complexity. This suggests a non-trivial double scaling limit for random neural networks in which

  $$n, L \to \infty \qquad \text{and} \qquad L/n \to \xi \in [0, \infty).$$

  See Conjecture 11. For non-linear networks this scaling limit has only started to be considered Hanin (2018); Hanin and Nica (2020b); Huang and Yau (2020); Li et al. (2022); Roberts et al. (2022). Even in the very special case of product of L iid random $n \times n$ matrices (sometimes called deep linear networks) the simultaneous large n, L

regime has revealed a range of interesting and not fully understood properties (see e.g. references in Akemann and Burda (2012); Akemann et al. (2019); Gorin and Sun (2018); Hanin and Nica (2020b); Hanin and Paouris (2021); Hanin and Zlokapa (2022); Liu et al. (2016)). This stands in contrast to the $\xi = 0$ regime often considered in previous work on neural networks (c.f. e.g. Du et al. (2018, 2019); Jacot et al. (2018); Liu et al. (2022)).

Several key ideas used to derive the results outlined above were first derived at a physics level of rigor in Roberts et al. (2022). This monograph, based on joint work of the author with Roberts and Yaida, develops a set of techniques that allow one to characterize properties of wide network at first order in $1/n$. A key strength of these techniques is that they can be applied both at initialization (e.g. to understand objects such as the 4-th cumulant of the network output and the variance of the NTK) and also to networks *after training* (e.g. to obtain layer-wise recursions such as those in Chapter ∞ of Roberts et al. (2022) for network predictions after training). However, not only are these techniques developed only at a physics level of rigor but they are not well-adapted to studying all order effects in powers of $1/n$. One of the contributions of the present article is therefore to both make mathematically rigorous the ideas in Roberts et al. (2022) and extend them to the point where one can tractably study higher cumulants, though for now only at initialization. We refer the interested reader to §3 for a brief technical summary of how we are able to systematically capture the structure of all effects at all orders in powers of $1/n$.

1.3 Outline for Rest of the Article

The remainder of this article is structured as follows. We begin in §2.1 by providing some background about prior work on why at the start of training neural networks in the NTK regime are Gaussian processes. We then present a range of results about the distribution of wide but not infinitely wide networks at the start of training. As a first result we provide in §2.2 sharp estimates, in terms of powers of $1/n$ for the cumulants of the output of a random neural network and its derivatives (see Theorem 3). Then, in §2.3 we give expansions, as a series in $1/n$, for expectations of observables evaluated on the finite-dimensional distribution of the output of a random neural network (see Theorem 4). This yields two Corollaries. The first, Corollary 5 shows formally that cumulants form a perturbatively solvable hierarchy with respect to network depth. Corollary 6 then makes these recursions explicit in some special cases. As a final result, we present in §2.4 the solutions to the cumulant recursions from Corollary 6, in which the effective network depth enters explicitly.

After presenting our results, we give some background on cumulants and Gaussian integration by parts in §4. We then explain the main idea behind the proof our most technical results, Theorems 3 and 4, in §3 before providing detailed proofs of our results starting in §5. Finally, in the Appendix, we recapitulate the discussion of criticality and universality in deep neural networks from Roberts et al. (2022), recall a trick for obtaining the exact distribution of the output a ReLU network evaluated a single input, and give a detailed analysis at large depth of the infinite width behavior of the Gaussian processes obtained from networks with tanh-like non-linearities.

2. Results

2.1 Background on Neural Networks at Infinite Width

To put the results of this article into context, note that in practice neural networks have many parameters. Thus, to study the structure of neural networks at the start of training, it is sensible to first understand various limits in which the number of parameters tends to infinity. The most well-studied regime of this type (though not the only option cf eg Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2020, 2021); Yang and Hu (2021)) is the NTK regime alluded to in the Introduction. By definition, this regime is accessed by fixing the depth L, the input and output dimensions n_0, n_{L+1}, the non-linearity σ, the initialization scheme (2) and considering the limit when $n_1, \ldots, n_L \to \infty$. In view of the relation (3) the NTK regime is obtained by taking $n \to \infty$ at fixed L. At the start of training, neural networks converge to a Gaussian process (see Theorem 2 below and Lee et al. (2018); Matthews et al. (2018); Neal (1996); Novak et al. (2018); Yang (2019a,b); Yang and Schoenholz (2018)). More precisely, we have the following result:

Theorem 2 (Gaussian Process Limit of Wide Networks) *Fix a non-negative integer $r \geq 0$, and suppose $\sigma : \mathbb{R} \to \mathbb{R}$ is r-times differentiable and that its r-th derivative is polynomially bounded:*

$$\exists k \geq 1 \ \textit{s.t.} \ \sup_{x \in \mathbb{R}} \left| (1 + |x|)^{-k} \frac{d^r}{dx^r} \sigma(x) \right| < \infty.$$

Then the finite-dimensional distributions of the stochastic process $x_\alpha \mapsto z_\alpha^{(L+1)}$ and its partial derivatives (with respect to the input) of order up to r converge to those of a centered Gaussian process with n_{L+1} iid components. The limiting covariance function of each component

$$K_{\alpha\beta}^{(L+1)} := \lim_{n_1, \ldots, n_L \to \infty} \mathrm{Cov}\left(z_{i;\alpha}^{(L+1)}, z_{i;\beta}^{(L+1)} \right), \qquad x_\alpha, x_\beta \in \mathbb{R}^{n_0},$$

satisfies the recursion

$$K_{\alpha\beta}^{(\ell+1)} = \begin{cases} C_b + C_W \left\langle \sigma(z_\alpha)\sigma(z_\beta) \right\rangle_{K^{(\ell)}}, & \ell \geq 1 \\ C_b + \frac{C_W}{n_0} \sum_{j=1}^{n_0} x_{j;\alpha} x_{j;\beta}, & \ell = 0 \end{cases}. \tag{4}$$

In the statement of Theorem 2 and henceforth we reserve the symbol $\langle f(z_\alpha, z_\beta) \rangle_\kappa$ to denote the expectation of $f(z_\alpha, z_\beta)$ with respect to the Gaussian distribution

$$(z_\alpha, z_\beta) \sim \mathcal{N}\left(0, \begin{pmatrix} \kappa_{\alpha\alpha} & \kappa_{\alpha\beta} \\ \kappa_{\alpha\beta} & \kappa_{\beta\beta} \end{pmatrix} \right),$$

where $\kappa_{\alpha\beta} = \kappa(x_\alpha, x_\beta)$ is a given covariance function. The conclusion in Theorem 2 is not new, having been obtained many times and under a variety of different assumptions, including for more general architectures. See Hanin (2021); Lee et al. (2018); Matthews et al. (2018); Poole et al. (2016); Yang (2019b). We refer the interested reader to Hanin (2021) for a discussion of prior work and note only that convergence of the derivatives of the field $z_\alpha^{(L+1)}$ to its Gaussian limit does not seem to have been previously considered. We

give a short proof that includes convergence of derivatives along the lines of the arguments in Hanin et al. (2022); Lee et al. (2018) in Appendix §A.

In the NTK regime, not only are neural networks at initialization given by Gaussian processes but, for the purposes of optimization of a squared loss with a small learning rate, the network can be replaced by *its linearization at the start of training* (see Bartlett et al. (2021); Chizat and Bach (2018); Du et al. (2018); Jacot et al. (2018); Lee et al. (2019); Liu et al. (2022)). Taken together, these two points show that at least at infinite width and finite depth, it is the structure of the network at initialization that determines not only the start of training but really the entire training trajectory. However, as we've already mentioned, by virtue of the second point, the infinite width limit is too rigid to capture the ability of real work networks to learn data-dependent features (see e.g. Hanin and Nica (2020a); Huang and Yau (2020); Roberts et al. (2022); Yang and Hu (2021)). With the NTK initialization (2), only finite width networks can capture these effects! It is the study of such networks that we take up in this article.

2.2 Results on the Size of Cumulants at Finite Width

Since in the infinite width limit, the field $z_\alpha^{(L+1)}$ is Gaussian (see Theorem 2), it is natural to study finite width corrections to this regime by considering the behavior of the cumulants of $z_\alpha^{(L+1)}$ and its derivatives at large but finite values of the network width n. Recall that a Gaussian process is determined by the condition that the mixed cumulants of order three and higher vanish. Our first result, Theorem 3, gives sharp estimates on the rate of vanishing in powers of $1/n$ for the cumulants of the finite-dimensional distributions of $z_\alpha^{(L+1)}$ at large width, providing a quantitative version of Theorem 2.

In order to state Theorem 3, we need some notation. First, given random variables $X_1, \dots, X_k$ with finite moments defined on the same probability space, let us denote their mixed cumulant by

$$\kappa\left(X_1, \dots, X_k\right) := i^k \frac{\partial^k}{\partial t_1 \cdots \partial t_k}\bigg|_{t=0} \log \mathbb{E}\left[\exp\left[-i(t_1 X_1 + \cdots + t_k X_k)\right]\right]. \tag{5}$$

Thus, for example, $\kappa(X_1) = \mathbb{E}\left[X_1\right]$ and $\kappa(X_1, X_2) = \mathrm{Cov}(X_1, X_2)$. We refer the reader to §4.1 for background on cumulants. Next, let us we fix a finite collection

$$\{x_\alpha,\, \alpha \in \mathcal{A}\} \subseteq \mathbb{R}^{n_0},$$

of $|\mathcal{A}|$ distinct network inputs. Moreover, let us also fix a collection of p directional derivatives:

$$D = (d_1, \dots, d_p), \qquad d_j := \nabla_{v_j} = \sum_{i=1}^{n_0} v_{ij} \partial_{x_i} \tag{6}$$

and for any multi-index $J = (j_1, \dots, j_p) \in \mathbb{N}^p$ denote by

$$D_\alpha^J := d_1^{j_1} \cdots d_m^{j_p}\Big|_{x=x_\alpha}$$

the corresponding differential operator of order $|J| := j_1 + \cdots + j_p$.

Theorem 3 (size of cumulants in random networks) *Fix $r, L \geq 1$ and suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies Assumption (1) with this value of r. Suppose further that one of the following two conditions holds:*

- *σ is smooth*
- *the limiting covariance matrix*

$$\left(\lim_{n\to\infty} \mathrm{Cov}\left(D^{J_1}_{\alpha_1} z^{(\ell)}_{1;\alpha_1}, D^{J_2}_{\alpha_2} z^{(\ell)}_{1;\alpha_2}\right)\right)_{\substack{|J_1|,|J_2|\leq r \\ \alpha_1,\alpha_2\in\mathcal{A}}} \tag{7}$$

of derivatives of order at most r in the directional derivatives $d_1, \ldots, d_p$ of the scalar field $z^{(\ell)}_{1;\alpha}$ is strictly positive definite in the infinite width limit for all $\ell \leq L$.

Then, for each $k \geq 1$ and $1 \leq \ell \leq L+1$, as $n \to \infty$

$$\kappa\left(D^{J_1}_{\alpha_1} z^{(\ell)}_{i_1;\alpha_1}, \ldots, D^{J_k}_{\alpha_k} z^{(\ell)}_{i_k;\alpha_k}\right) = \begin{cases} 0, & k \text{ odd} \\ O(n^{-\frac{k}{2}+1}), & k \text{ even} \end{cases}, \tag{8}$$

where the implicit constant in the error term depends on k, the inputs $x_{\alpha_1}, \ldots, x_{\alpha_k}$, the multi-indices $J_1, \ldots, J_k$, the weight and bias variances C_b, C_W, the non-linearity σ, and the layer index ℓ.

We prove Theorem 3 in §5. At a physics level of rigor, Theorem 3 with $k = 4$ and no derivatives was already derived in the breakthrough work of Yaida (2020). In fact, Yaida's original article went much further: it obtained a recursive formula with respect to ℓ for the fourth cumulant $\kappa(z^{(\ell)}_{i_1;\alpha_1}, \ldots, z^{(\ell)}_{i_4;\alpha_4})$ at layer ℓ in terms of the second and fourth cumulants at layer $\ell - 1$. This is analogous to the recursion (4) for the infinite width covariance $K^{(\ell)}_{\alpha_1\alpha_2}$. This theme was then picked up and significantly expanded upon in the physics monograph Roberts et al. (2022), which computes, among other things, at order $1/n$ the leading corrections to the field $z^{(\ell)}_{\alpha}$ and its derivatives with respect to both x_α and model parameters. We will reproduce some of these recursions and obtain new ones of a similar flavor below (see Corollary 6).

Compared to prior work the main novelty of Theorem 3 is two-fold. First, it gives sharp estimates in powers of $1/n$ for cumulants of all orders (the sharpness can already be seen when $\ell = 2$). Second, it treats in a uniform way the cumulants for not only the values but also all derivatives of $z^{(\ell)}_{\alpha}$.

2.3 Results on Layer-wise Recursions for Cumulants

The estimate (8) in Theorem 3 only gives the order of magnitude in powers of $1/n$ for the cumulants $\kappa(D^{J_1}_{\alpha_1} z^{(\ell)}_{i_1;\alpha_1}, \ldots, D^{J_k}_{\alpha_k} z^{(\ell)}_{i_k;\alpha_k})$ but does not directly provide information about their structural dependence on the remaining model parameters C_b, C_W, σ, ℓ. Our next set of results supplies such information.

To state them, let us denote by $\mathcal{F}^{(\ell)}$ the sigma algebra generated by the weights and biases in layers up to and including ℓ. Since we've assumed weights and biases to be Gaussian and independent for different neurons, note that conditional on $\mathcal{F}^{(\ell)}$ the vectors

$$z^{(\ell+1)}_{i;\mathcal{A}} := \left(z^{(\ell+1)}_{i;\alpha},\, \alpha \in \mathcal{A}\right)$$

are independent centered Gaussians with the conditional covariance

$$\Sigma^{(\ell)}_{\alpha\alpha'} := \mathrm{Cov}\left(z^{(\ell+1)}_{i;\alpha_1}, z^{(\ell+1)}_{i;\alpha_2} \mid \mathcal{F}^{(\ell)}\right) = C_b + \frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell} \sigma\left(z^{(\ell)}_{j;\alpha_1}\right)\sigma\left(z^{(\ell)}_{j;\alpha_2}\right). \tag{9}$$

Let us denote by

$$\kappa^{(\ell)}_{\alpha_1\alpha_2} := \mathrm{Cov}\left(z^{(\ell+1)}_{i;\alpha_1}, z^{(\ell+1)}_{i;\alpha_2}\right) = \mathbb{E}\left[\Sigma^{(\ell)}_{\alpha_1\alpha_2}\right]$$

the finite width covariance and by

$$\Delta^{(\ell)}_{\alpha_1\alpha_2} := \Sigma^{(\ell)}_{\alpha_1\alpha_2} - \kappa^{(\ell)}_{\alpha_1\alpha_2}$$

the difference between the conditional covariance matrix $\Sigma^{(\ell)}$ and its mean $\kappa^{(\ell)}$. Further, let us agree to denote by $\langle\cdot\rangle_{\kappa^{(\ell)}}$ the expectation with respect to a collection of centered jointly Gaussian random vectors

$$D^{\le r}_{\mathcal{A}} z_i = \left(D^J_\alpha z_{i;\alpha},\, \alpha\in\mathcal{A},\, |J|\le r\right)$$

which match the covariance of the neural network derivatives

$$D^{\le r}_{\mathcal{A}} z^{(\ell)}_i := \left(D^J_\alpha z^{(\ell)}_{i;\alpha},\, \alpha\in A,\, |J|\le r\right)$$

in the sense that

$$\mathrm{Cov}\left(D^{J_1}_{\alpha_1} z_{i_1;\alpha_1},\, D^{J_2}_{\alpha_2} z_{i_2;\alpha_2}\right) := \mathrm{Cov}\left(D^{J_1}_{\alpha_1} z^{(\ell)}_{i_1;\alpha_1},\, D^{J_2}_{\alpha_2} z^{(\ell)}_{i_2;\alpha_2}\right) = \delta_{i_1 i_2} D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \kappa^{(\ell)}_{\alpha_1\alpha_2}$$

in each component separately but are defined to have zero covariance for different neurons.

Theorem 4 ($1/n$ expansion of expectations at finite width) *Fix an integer $r \ge 0$ and suppose that f is continuous and polynomially bounded. Then for any $q_* \ge 0$ we have*

$$\begin{aligned}
&\mathbb{E}\left[f\left(D^{\le r} z^{(\ell+1)}_{1,\mathcal{A}}, \ldots, D^{\le r} z^{(\ell+1)}_{m,\mathcal{A}}\right)\right] \\
&\quad = \sum_{q=0}^{2q_*} \frac{(-1)^q}{2^q q!} \left\langle \mathbb{E}\left[\left(\sum_{\substack{|J|,|J'|\le r \\ \alpha,\alpha'\in\mathcal{A}}} \Delta^{JJ',(\ell)}_{\alpha\alpha'} \sum_{j=1}^m \partial_{D^J_\alpha z_{j;\alpha}} \partial_{D^{J'}_{\alpha'} z_{j;\alpha'}}\right)^q\right] f\left(D^{\le r}_{\mathcal{A}} z_1, \ldots, D^{\le r}_{\mathcal{A}} z_m\right)\right\rangle_{\kappa^{(\ell+1)}} \\
&\quad + O(n^{-q_*-1}),
\end{aligned} \tag{10}$$

where the sum is over multi-indices $J, J' \in \mathbb{N}^p$ of order at most r, we've set

$$\Delta^{JJ',(\ell)}_{\alpha\alpha'} := D^J_\alpha D^{J'}_{\alpha'} \Sigma^{(\ell)}_{\alpha\alpha'} - \mathbb{E}\left[D^J_\alpha D^{J'}_{\alpha'} \Sigma^{(\ell)}_{\alpha\alpha'}\right], \tag{11}$$

and the derivatives $\partial_{D^J_\alpha z_{j;\alpha}}$ are interpreted in the weak sense if f is not differentiable.

For example, taking $r = 0, m = 1$, and $\mathcal{A}$ be the singleton $\{\alpha\}$ gives

$$\mathbb{E}\left[f\left(z^{(\ell+1)}_{1;\alpha}\right)\right] = \sum_{q=0}^{2q_*} \mathbb{E}\left[\left(\Delta^{(\ell)}_{\alpha\alpha}\right)^q\right] \left\langle \left(\partial_{z_{1;\alpha}}\right)^{2q} f\left(z_{1;\alpha}\right)\right\rangle_{\kappa^{(\ell+1)}} + O(n^{-q_*-1}). \tag{12}$$

As we explain in Lemma 18, Theorem 3 immediately yields

$$\mathbb{E}\left[\prod_{i=1}^{q}\Delta_{\alpha_i\alpha_i'}^{J_iJ_i',(\ell)}\right]=O\left(n^{-\lceil\frac{q}{2}\rceil}\right),$$

showing that the expansions in (10) and (12) are indeed series in decreasing powers of $1/n$. Moreover, each term in these power series is simply given by computing a Gaussian integral with covariance $\kappa^{(\ell+1)}$ in which different neuron are independent.

We prove Theorem 4 in §7 and give the main idea of the proof in §3. By substituting various polynomials in $z_\alpha^{(\ell+1)}$ for f into the perturbative expansion (10), it is now possible in principle to deduce recursions for the cumulants at layer $\ell+1$ in terms of objects of the same type at layer ℓ. In particular, we have the following

Corollary 5 (hierarchy of cumulants in $1/n$) *With the assumptions of Theorem 3, the mixed cumulant*

$$\kappa\left(D_{\alpha_1}^{J_1}z_{i_1;\alpha_1}^{(\ell+1)},\ldots,D_{\alpha_{2k}}^{J_{2k}}z_{i_{2k};\alpha_{2k}}^{(\ell+1)}\right)$$

equals

$$\sum_{\substack{j\le k\\ J_i',\,i=1,\ldots,2j\\ |J_1'|+\cdots+|J_{2j}'|\\ \le|J_1|+\cdots+|J_{2k}|}} C(J_i',K_{\alpha_i\alpha_{i'}}^{(\ell)}i,i'=1,\ldots,2j)\;\kappa\left(D_{\alpha_1}^{J_1'}z_{1;\alpha_1}^{(\ell)},\ldots,D_{\alpha_{2j}}^{J_{2j}'}z_{2j;\alpha_{2j}}^{(\ell)}\right)+O\left(n^{-k}\right), \tag{13}$$

where the sum is over multi-indices J_i', the constants $C(J_i',K_{\alpha_i\alpha_j}^{(\ell)}i,j=1,\ldots,2k)$ depend only on the multi-indices J_i' and the infinite width covariance $K^{(\ell)}$, while the implicit constant in the error term depends on k, the inputs $x_{\alpha_1},\ldots,x_{\alpha_k}$, the multi-indices $J_1,\ldots,J_k$, the weight and bias variances C_b,C_W, the non-linearity σ, and the layer index ℓ.

We do not know how to efficiently compute the coefficients C in the recursion (13) for general cumulants. Instead, we compute then by hand for small values of k and a single input $x_\alpha\in\mathbb{R}^{n_0}$:

$$\kappa_{2k;\alpha}^{(\ell)}:=\frac{1}{(2k-1)!!}\kappa\Big(\underbrace{z_{i;\alpha}^{(\ell+1)},\ldots,z_{i;\alpha}^{(\ell+1)}}_{2k\text{ times}}\Big) \tag{14}$$

when $k=2,3,4$. See Corollary 6. In order to facilitate a compact form for the recursions described in first bullet point, let us write

$$T_{i,j;\alpha}^{(\ell)}:=C_W^j\left\langle\partial_z^i\left\{\left(\sigma^2(z)-\left\langle\sigma^2(z)\right\rangle_{K_{\alpha\alpha}^{(\ell)}}\right)^j\right\}\right\rangle_{K_{\alpha\alpha}^{(\ell)}}, \tag{15}$$

with the derivatives interpreted in the weak sense if σ is not sufficiently smooth and where we remind the reader of our standing notation

$$\langle f(z)\rangle_K=\int_{-\infty}^{\infty}f(zK^{1/2})e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}},\qquad K\ge 0.$$

Corollary 6 *Fix $r \geq 1$ and suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies Assumption (1) with this value of r. Consider a depth L random neural network with input dimension n_0, hidden layer widths $n_1, \ldots, n_L$, output dimension n_{L+1} and non-linearity σ. Fix $x_\alpha \in \mathbb{R}^{n_0}$ and define*

$$\chi_{||;\alpha}^{(\ell)} := \frac{1}{2} T_{2,1;\alpha}^{(\ell)} = \frac{C_W}{2} \left\langle \partial_z^2 \sigma(z)^2 \right\rangle_{K_{\alpha\alpha}^{(\ell)}},$$

where the second derivative is interpreted in the weak sense if σ is not twice differentiable. For each $\ell = 1, \ldots, L$, in the notation of (14), the fourth cumulant satisfies

$$\kappa_{4;\alpha}^{(\ell+1)} = \frac{T_{0,2;\alpha}^{(\ell)}}{n_\ell} + \left(\chi_{||;\alpha}^{(\ell)}\right)^2 \kappa_{4,\alpha}^{(\ell)} + O(n^{-2}). \tag{16}$$

Further, the 6-th cumulant satisfies

$$\kappa_{6;\alpha}^{(\ell+1)} = \frac{T_{0,3;\alpha}}{n_\ell^2} + \frac{3T_{2,2;\alpha}^{(\ell)}}{2n_\ell} \chi_{||;\alpha}^{(\ell)} \kappa_{4;\alpha}^{(\ell)} + \frac{3T_{4,1;\alpha}^{(\ell)}}{4} \left(\chi_{||;\alpha}^{(\ell)} \kappa_{4;\alpha}^{(\ell)}\right)^2 + \left(\chi_{||;\alpha}^{(\ell)}\right)^3 \kappa_{6;\alpha}^{(\ell)} + O(n^{-3}). \tag{17}$$

Finally, the 8-th cumulant satisfies:

$$\begin{aligned}
\kappa_{8;\alpha}^{(\ell+1)} = & \frac{1}{n_\ell^3} \left(T_{0,4;\alpha}^{(\ell)} - 3 \left(T_{0,2;\alpha}^{(\ell)} \right)^2 \right) \\
& + \frac{1}{n_\ell^2} \left[2T_{2,3;\alpha}^{(\ell)} \chi_{||;\alpha}^{(\ell)} - 12 T_{0,2;\alpha}^{(\ell)} \left(\chi_{||;\alpha}^{(\ell)}\right)^2 + \frac{3}{2} \left(T_{2,2;\alpha}^{(\ell)} \right)^2 - \frac{3}{2} T_{4,1;\alpha}^{(\ell)} T_{0,2;\alpha}^{(\ell)} \right] \kappa_{4;\alpha}^{(\ell)} \\
& - \frac{1}{n_\ell} \left[2T_{2,2;\alpha}^{(\ell)} T_{4,1;\alpha}^{(\ell)} \chi_{||;\alpha}^{(\ell)} - \frac{1}{2} T_{4,2;\alpha}^{(\ell)} \left(\chi_{||;\alpha}^{(\ell)}\right)^2 + \left(\chi_{||;\alpha}^{(\ell)}\right)^4 \right] \left(\kappa_{4;\alpha}^{(\ell)}\right)^2 \\
& + \frac{1}{n_\ell} \left[5T_{0,2;\alpha}^{(\ell)} T_{4,1;\alpha}^{(\ell)} \chi_{||;\alpha}^{(\ell)} + 12 T_{2,2;\alpha}^{(\ell)} \left(\chi_{||;\alpha}^{(\ell)}\right)^2 \right] \kappa_{6;\alpha}^{(\ell)} \\
& + \frac{3}{32} \left(T_{4,1;\alpha}^{(\ell)} \right)^2 \left(\chi_{||;\alpha}^{(\ell)}\right)^2 \left(\kappa_{4;\alpha}^{(\ell)}\right)^3 - \frac{1}{2} \left(\chi_{||;\alpha}^{(\ell)}\right)^3 T_{4,1;\alpha}^{(\ell)} \kappa_{4;\alpha}^{(\ell)} \kappa_{6;\alpha}^{(\ell)} \\
& + \left(\chi_{||;\alpha}^{(\ell)}\right)^4 \kappa_{8;\alpha}^{(\ell)} + O(n^{-4}).
\end{aligned} \tag{18}$$

The initial condition for the recursions (16)-(18) is that $\kappa_{2k;\alpha}^{(1)} = 0$ for all $k \geq 2$.

Remark 7 *Note that for $k = 2, 3, 4$, we therefore see that to leading order in $1/n$ the recursions for $\kappa_{2k;\alpha}^{(\ell+1)}$ depends only on $\kappa_{2j;\alpha}^{(\ell)}$ for $j \leq k$. This allows us to interpret (16) - (18) as a forming the start of hierarchy in powers of $1/n$ describing the cumulants of the output of a random neural network.*

Let us briefly compare Corollary 6 to results in prior work:

- In the special case when σ is $1-$homogeneous (i.e. is linear, ReLU, leaky ReLU, etc, see (74)), the full distribution of a neuron pre-activation $z_{i;\alpha}^{(\ell)}$ can be worked out in closed form. Namely, as we explain in §B.2.1 and Appendix D, is has the same distribution as a Gaussian with an independent random variance given by a product of independent weighted chi-squared random variables. This was first pointed out in Hanin (2018); Hanin and Nica (2020b) and described in the language of special functions (namely Meijer G functions) in Zavatone-Veth and Pehlevan (2021). For such non-linearities obtaining the recursions (16)-(18) is not new.

- The breakthrough work Yaida (2020) was the first to obtain, at a physics level of rigor, the recursion (16) and probe its solutions at large ℓ.
- The ideas in Yaida (2020) then seeded the development in the monograph Roberts et al. (2022) a much richer analysis, producing at a physics level of rigor many recursions similar in flavor to (16)-(18) that describe the the behavior of objects such as network derivatives at the start of training, the NTK at the start of training, and even the change in the NTK and the resulting output of a *trained* network. Many of these results go far beyond what we are currently capable of doing mathematically.
- The analysis in Roberts et al. (2022) never required studying cumulants $\kappa^{(\ell)}_{2k;\alpha}$ for $k \geq 3$, and while the techniques developed there can certainly be used to obtain the recursions (17) and (18) we take a rather different approach in this article that produces such recursions more directly.

The functional $\chi^{(\ell)}_{||;\alpha}$ plays a fundamental role in the recursive description of random neural networks supplied by Corollary 6, whose proof is in §7. In §B we explain a principled procedure, called *tuning to criticality*, that reveals its origin (as well as that of a similar object we denote $\chi^{(\ell)}_{\perp;\alpha}$) and explains how to choose C_b, C_W so that these functionals are approximately equal to 1 at large ℓ. As we will see, such a choice will ensure that the recursions in Corollary 6 and their infinite width counterpart (4) have well-behaved solutions at large ℓ. We will then return in §B.2.1 and §B.2.2 to solving the recursions from Corollary 6 in random networks tuned to criticality (see (76) and Corollary 8).

2.4 Results on Effective Depth and Low Order Cumulants

As a final result, we record the following consequence of Corollary 6, which shows that it is the effective network depth that controls the size of low order cumulants.

Corollary 8 *Suppose σ is either* ReLU *or* tanh *and consider a depth L random neural network with input dimension n_0, output dimension n_{L+1}, hidden layer widths satisfying*

$$n_1, \ldots, n_L = n \gg 1,$$

and non-linearity σ with $C_W = 1$ if $\sigma = \tanh$ and $C_W = 2$ if $\sigma = \mathrm{ReLU}$ as well as $C_b = 0$ in both cases. Write $\xi = L/n$ and define the normalized cumulants

$$\widehat{\kappa}^{(L)}_{2k;\alpha} := \frac{\kappa^{(L)}_{2k;\alpha}}{(K^{(L)}_{\alpha\alpha})^k}$$

of pre-activations in layer L corresponding to a fixed network input x_α. We have for $k = 2, 3, 4$ that

$$\widehat{\kappa}^{(L)}_{2k;\alpha} = C_{2k}\xi^{k-1}\left(1 + O(L^{-1})\right) + O(n^{-k}), \tag{19}$$

where C_{2k} are some positive universal constants depending on σ. The implicit constants in error terms $O(L^{-1})$ depend on σ, C_b, C_W and constants in $O(n^{-j})$, $j = 2, 3, 4$ may depend in addition on L.

Remark 9 *Although we have formulated Corollary 8 only for ReLU and* tanh *non-linearities we actually prove it for a more general classes of ReLU-like and tanh-like non-linearities defined in §B.*

Remark 10 *Combining estimate* (19) *for* $k = 2$ *with the definition* (79) *of the* $K_* = 0$ *universality class and* (14) *yields that in the setting of Corollary 6 we have to first order in* $1/n$ *and to leading order in* $1/L$ *that*

$$\widehat{\mathrm{Var}}\left[\left(z_{1;\alpha}^{(\ell+1)}\right)^2\right] = const \times (1+\xi), \quad \mathrm{Corr}\left(\left(z_{1;\alpha}^{(\ell+1)}\right)^2, \left(z_{2;\alpha}^{(\ell+1)}\right)^2\right) = const \times \xi, \quad \xi := \frac{L}{n},$$

where $\widehat{\mathrm{Var}}[X] = \mathrm{Var}[X]/\mathbb{E}\left[X^2\right]$. *Thus, both the fluctuations of a single neuron pre-activation and the correlation between different neurons is controlled to first order in* $1/n, 1/L$ *by the effective network depth* ξ.

We prove Corollary 6 in §7.1 and note that formula (19) was derived in Yaida (2020) for $k = 2$ at a physics level of rigor. While we do not know how to generalize the results in (19) to obtain the corresponding formulas for general k, we make the following conjecture:

Conjecture 11 (Double Scaling Limit for Random Neural Networks) *Consider a random depth* L *neural network with input dimension* n_0, *hidden layer widths*

$$n_1, \ldots, n_L = n \gg 1,$$

output dimension n_{L+1} *and non-linearity* σ. *Suppose further that this network is tuned to criticality in the sense that* (73) *is satisfied. Fix a non-zero network input* $x_\alpha \in \mathbb{R}^{n_0}$ *and write* $\xi = L/n$. *For each* $k \geq 1$ *there exists* $C_{2k} > 0$ *depending on the universality class of* σ *so that*

$$\frac{\kappa_{2k;\alpha}^{(L)}}{\left(K_{2;\alpha}^{(L)}\right)^k} = C_{2k}\xi^{k-1} + O\left(\xi^k\right).$$

Moreover, for each $\xi \in [0, \infty)$ *there exists a probability distribution* $\mathbb{P}_{\xi,\sigma}$ *on* $\mathbb{R}$, *depending only on* ξ *and* σ, *such that in the double scaling limit*

$$n, L \to \infty, \qquad \frac{L}{n} \to \xi,$$

the random variable $z_{i;\alpha}^{(\ell)}$ *converges in distribution to a random variable with law* $\mathbb{P}_{\xi,\sigma}$.

3. Overview of Proofs

In this section, we present the essential idea for how we analyze a random fully connected neural network $x_\alpha \mapsto z_\alpha^{(L+1)}$ at finite width. Our approach is based on the following structural properties:

- The sequence of fields $z_\alpha^{(\ell)}$ is a Markov Chain with respect to ℓ.

- Conditional on the sigma algebra $\mathcal{F}^{(\ell)}$ defined by $z_\alpha^{(\ell)}$ is a Gaussian field with independent components $z_{i;\alpha}^{(\ell+1)}$. See Lemma 14.

- The variance of each component $z_{i;\alpha}^{(\ell+1)}$ depends on $z_\alpha^{(\ell)}$ only through random variables of the form

$$\mathcal{O}_f^{(\ell)} := \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} f(z_{i;\alpha}^{(\ell)}), \qquad f:\mathbb{R}\to\mathbb{R} \text{ polynomially bounded}$$

  which we refer to as collective observables. See (9).

- Centered moments of collective observables depend on n as if the random variables $f(z_{i;\alpha}^{(\ell)})$ were independent:

$$\mathbb{E}\left[\left(\mathcal{O}_f^{(\ell)} - \mathbb{E}\left[\mathcal{O}_f^{(\ell)}\right]\right)^q\right] = O_q\left(n^{-\lceil \frac{q}{2}\rceil}\right), \qquad q\geq 0. \tag{20}$$

  Establishing this is the most difficult technical aspect of the present article. See Theorem 16 and Lemma 18.

Let us briefly explain, mostly dispensing with rigor, how these four ideas come together to obtain a recursive description of the distribution of the field $z_\alpha^{(\ell+1)}$ in terms of that of $z_\alpha^{(\ell)}$. To keep the notation to a minimum, we fix a network input x_α and focus on describing the joint distribution of $z_{i;\alpha}^{(\ell+1)}$, $i=1,\dots,m$. Extensions to multiple inputs and derivatives proceed along very similar lines. Denoting by $\xi=(\xi_1,\dots,\xi_m)$ dual variables, consider the characteristic function

$$p^{(\ell+1)}(\xi) := \mathbb{E}\left[\exp\left[-i\sum_{j=1}^m \xi_j z_{j;\alpha}^{(\ell+1)}\right]\right].$$

Conditioning on $z_\alpha^{(\ell)}$ and using (9) allows us to write

$$p^{(\ell+1)}(\xi) := \mathbb{E}\left[\exp\left[-\frac{1}{2}\,||\xi||^2\,\Sigma_{\alpha\alpha}^{(\ell)}\right]\right],$$

where we remind the reader that

$$\Sigma_{\alpha\alpha}^{(\ell)} = \mathrm{Var}\left[z_{i;\alpha}^{(\ell+1)} \;\middle|\; \mathcal{F}^{(\ell)}\right] = C_b + \frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\sigma(z_{j;\alpha}^{(\ell)})^2$$

is a collective observable *at the previous layer*. Writing

$$\kappa_{\alpha\alpha}^{(\ell)} := \mathbb{E}\left[\Sigma_{\alpha\alpha}^{(\ell)}\right], \qquad \Delta_{\alpha\alpha}^{(\ell)} := \Sigma_{\alpha\alpha}^{(\ell)} - \mathbb{E}\left[\Sigma_{\alpha\alpha}^{(\ell)}\right],$$

we find

$$p^{(\ell+1)}(\xi) := \mathbb{E}\left[\exp\left[-\frac{1}{2}\,||\xi||^2\,\Delta_{\alpha\alpha}^{(\ell)}\right]\right]\exp\left[-\frac{1}{2}\,||\xi||^2\,\kappa_{\alpha\alpha}^{(\ell)}\right].$$

The second term is precisely the characteristic function of a centered m-dimensional Gaussian with iid components of variance $\kappa^{(\ell)}_{\alpha\alpha}$. Moreover, at least heuristically, the first term can be written

$$\mathbb{E}\left[\exp\left[-\frac{1}{2}\,||\xi||^2\,\Delta^{(\ell)}_{\alpha\alpha}\right]\right] = \sum_{q\geq 0}\mathbb{E}\left[\left(\Delta^{(\ell)}_{\alpha\alpha}\right)^q\right]\frac{(-1)^q}{2^q q!}\,||\xi||^{2q}\,.$$

The concentration estimates (20) ensure that this series converges. Moreover, since the Fourier transform turns polynomials into derivatives we have

$$-\,||\xi||^2 = \text{ Laplacian in the variables } z^{(\ell+1)}_{i;\alpha}.$$

Hence, we obtain for any reasonable test function f that

$$\mathbb{E}\left[f(z^{(\ell+1)}_{i;\alpha},\, i=1,\dots,m)\right] = \sum_{q=0}^{\infty}\frac{1}{2^q q!}\mathbb{E}\left[\left(\Delta^{(\ell)}_{\alpha\alpha}\right)^q\right]\left\langle\left(\sum_{i=1}^m \partial^2_{z_{i;\alpha}}\right)^q f(z_{i;\alpha},\, i=1,\dots,m)\right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}},$$

where $(z_{i;\alpha},\, i=1,\dots,m)$ is a vector of iid centered Gaussians with variance $\kappa^{(\ell)}_{\alpha\alpha}$. The concentration estimates (20) ensure that this expression is a power series in $1/n$. In particular,

$$\begin{aligned}\mathbb{E}\left[f(z^{(\ell+1)}_{i;\alpha},\, i=1,\dots,m)\right] &= \left\langle f(z_{i;\alpha},\, i=1,\dots,m)\right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}} \\ &\quad + \frac{\mathbb{E}\left[(\Delta^{(\ell)}_{\alpha\alpha})^2\right]}{8}\left\langle\left(\sum_{i=1}^m \partial^2_{z_{i;\alpha}}\right)^2 f(z_{i;\alpha},\, i=1,\dots,m)\right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}} + O(n^{-2}).\end{aligned} \tag{21}$$

This is the essence of Theorem 4. To derive usable recursions for cumulants of $z^{(\ell+1)}_{i;\alpha}$, note for instance that, in the notation of Corollary 6,

$$\kappa^{(\ell)}_{4;\alpha} := \frac{1}{3}\kappa\left(z^{(\ell+1)}_{i;\alpha}, z^{(\ell+1)}_{i;\alpha}, z^{(\ell+1)}_{i;\alpha}, z^{(\ell+1)}_{i;\alpha}\right) = \mathbb{E}\left[(\Delta^{(\ell)}_{\alpha\alpha})^2\right].$$

Writing

$$X_j := \sigma(z^{(\ell+1)}_{j;\alpha})^2 - \mathbb{E}\left[\sigma(z^{(\ell+1)}_{j;\alpha})^2\right]$$

we thus have

$$\kappa^{(\ell+1)}_{\alpha\alpha} = \mathbb{E}\left[\left(\Delta^{(\ell)}_{\alpha\alpha}\right)^2\right] = \frac{C_W^2}{n_\ell}\mathbb{E}\left[X_1^2\right] + C_W^2\left(1-n_\ell^{-1}\right)\mathbb{E}\left[X_1X_2\right].$$

Applying the expansion (21) to both these terms and a bit of algebra already yields

$$\begin{aligned}\kappa^{(\ell+1)}_{\alpha\alpha} &= \mathbb{E}\left[\left(\Delta^{(\ell)}_{\alpha\alpha}\right)^2\right] \\ &= \frac{C_W^2}{n_\ell}\left(\left\langle\sigma^4\right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}} - \left\langle\sigma^2\right\rangle^2_{\kappa^{(\ell)}_{\alpha\alpha}}\right) \\ &\quad + C_W^2\left(1-n_\ell^{-1}\right)\left(\left(\left\langle\sigma^2\right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}} - \mathbb{E}\left[\sigma(z^{(\ell)}_{1;\alpha})^2\right]\right)^2 + \frac{1}{4}\left\langle\partial^2\sigma^2\right\rangle^2_{\kappa^{(\ell)}_{\alpha\alpha}}\kappa^{(\ell)}_{4;\alpha}\right) + O(n^{-2}).\end{aligned}$$

A short argument supplied in §7 shows that

$$\left\langle \sigma^2 \right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}} = \mathbb{E}\left[\sigma(z^{(\ell)}_{1;\alpha})^2\right] + O(n^{-1})$$

and that we may replace $\kappa^{(\ell)}_{\alpha\alpha}$ by its infinite width limit $K^{(\ell)}_{\alpha\alpha}$ in all remaining expectations at the cost of an $O(n^{-1})$ error. This already yields the recursion (16) of Corollary 6.

4. Background

4.1 Properties of Cumulants

Recall that, given random variables $X_1, \ldots, X_k$ on the same probability space, we denote their mixed cumulant by

$$\kappa\left(X_1, \ldots, X_k\right) := i^k \frac{\partial^k}{\partial t_1 \cdots \partial t_k}\bigg|_{t=0} \log \mathbb{E}\left[\exp\left[-i(t_1 X_1 + \cdots + t_k X_k)\right]\right].$$

In the following result, we recall the key properties of these mixed cumulants that we will need.

Proposition 12 (See Theorem 2.3.1 in Brillinger (2001)) *Mixed cumulants satisfy the following properties.*

1. *Suppose $X = (X_1, \ldots, X_k)$ is a random vector with finite moments of all orders. Then, for any sub-sigma algebra $\mathcal{F}$ of the probability space on which X is defined*

$$\kappa\left(X_1, \ldots, X_k\right) = \sum_{\pi=(\pi_1,\ldots\pi_b)} \kappa\left(\kappa\left(X_{\pi_1} \mid \mathcal{F}\right), \ldots, \kappa\left(X_{\pi_b} \mid \mathcal{F}\right)\right), \tag{22}$$

   where the sum is over all partitions π of $[k]$ and for each $a = 1, \ldots, b$

$$X_{\pi_a} := \left(X_i,\, i \in \pi_a\right).$$

   This is known as the law of total cumulance. See Brillinger (1969).

2. *Suppose $X = (X_1, \ldots, X_k)$ is a random vector with finite moments of all orders. When X can be partitioned into two independent subsets, the mixed cumulant $\kappa(X)$ vanishes. More precisely, suppose $I \subseteq [k]$ and that $I, I^c \neq \emptyset$. Then*

$$X_I := \left(X_i,\, i \in I\right) \perp X_{I^c} = \left(X_i,\, i \notin I\right) \qquad \Longrightarrow \qquad \kappa\left(X_1, \ldots, X_k\right) = 0. \tag{23}$$

3. *Mixed cumulants are multi-linear. More precisely if*

$$\left\{X_{i,j},\, 1 \leq j \leq k,\, i \leq T_j\right\}$$

   are random variables with finite moments defined on the same probability space, then

$$\kappa\left(\sum_{i_1=1}^{T_1} a_{i_1,1} X_{i_1,1}, \ldots, \sum_{i_k=1}^{T_k} a_{i_k,k} X_{i_k,k}\right) = \sum_{i_1=1}^{T_1} \cdots \sum_{i_k=1}^{T_k} a_{i_1,1} \cdots a_{i_k,k} \kappa\left(X_{i_1,1}, \ldots, X_{i_k,k}\right) \tag{24}$$

   for any $a_{i,j} \in \mathbb{R}$.

4. *Suppose $X = (X_1, \ldots, X_j) \sim \mathcal{N}(0, \Sigma)$ is a centered Gaussian with covariance Σ. Then for any $i_1, \ldots, i_k \in [j]$*

$$\kappa\left(X_{i_1}, \ldots, X_{i_k}\right) = \begin{cases} \Sigma_{i_1 i_2}, & k = 2 \\ 0, & \text{otherwise} \end{cases}. \tag{25}$$

5. *Moments are polynomials in cumulants. Specifically, suppose $X = (X_1, \ldots, X_k)$ is a random vector with finite moments of all orders. Then,*

$$\mathbb{E}\left[X_1 \cdots X_k\right] = \sum_{\pi = (\pi_1, \ldots, \pi_b)} \prod_{a=1}^{b} \kappa\left(X_{\pi_a}\right), \tag{26}$$

where the sum is over all partitions of $[k]$ and for each $a \in [b]$ we've written

$$X_{\pi_a} := \left(X_i,\, i \in \pi_a\right).$$

6. *Cumulants are polynomials in moments. Specifically,*

$$\kappa\left(X_1, \ldots, X_k\right) = \sum_{\pi = (\pi_1, \ldots, \pi_b)} (-1)^{b-1}(b-1)! \prod_{a=1}^{b} \mathbb{E}\left[\prod_{i \in \pi_a} X_i\right], \tag{27}$$

where the sum is over all partitions of $[k]$ and for each $a \in [b]$ we've written

$$X_{\pi_a} := \left(X_i,\, i \in \pi_a\right).$$

4.2 Gaussian Integration Lemma

Lemma 13 *Fix $r \geq 1$, a $r \times r$ matrix Σ and measurable function $g : \mathbb{R}^r \to \mathbb{R}$ that is polynomially bounded:*

$$\exists r \geq 1 \text{ s.t. } \sup_{x \in \mathbb{R}^r} \left|(1 + ||x||)^{-r}\, g(x)\right| < \infty.$$

If X is a standard Gaussian random vector in $\mathbb{R}^r$, then the function

$$\Sigma \mapsto \mathbb{E}\left[g\left(\Sigma^{1/2} X\right)\right] \tag{28}$$

is smooth on the open set of strictly positive definite $k \times k$ matrices. Further, if g is a smooth function and each of its derivatives is polynomially bounded, then the map (28) is extends to a smooth function on the closed set of positive semi-definite matrices and, in particular,

$$\frac{\partial}{\partial \Sigma_{ij}} \mathbb{E}\left[g\left(\Sigma^{1/2} X\right)\right] = \mathbb{E}\left[(\partial_i \partial_j g)(\Sigma^{1/2} X)\right]. \tag{29}$$

Proof On the open set of strictly positive definite matrices, the Gaussian density

$$\Sigma \mapsto \exp\left[-\frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2} \log\det(2\pi\Sigma)\right]$$

is a smooth function of Σ with derivatives that are polynomials in x and the entries of Σ, Σ^{-1}. The assumption that f is polynomially bounded shows that we may differentiate under the integral sign and see that that

$$\mathbb{E}\left[g(\Sigma^{1/2}X)\right] = \int_{\mathbb{R}^r} g(x) \exp\left[-\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\log\det(2\pi\Sigma)\right] dx$$

is indeed a smooth function of Σ. Suppose instead that g is a smooth function and that it's derivatives are all polynomially bounded. Suppose first that g is in fact a Schwartz function. Then, writing $\widehat{g}$ for its Fourier transform we have

$$\mathbb{E}\left[g(\Sigma^{1/2}X)\right] = \int_{\mathbb{R}^r} \widehat{g}(\xi) \exp\left[-\frac{1}{2}\xi^T\Sigma\xi\right] d\xi.$$

Since $\widehat{g}$ is also Schwartz, we may differentiate under the integral sign to obtain

$$\frac{\partial}{\partial\Sigma_{ij}}\mathbb{E}\left[g(\Sigma^{1/2}X)\right] = -\int_{\mathbb{R}^r} \xi_i\xi_j\widehat{g}(\xi) \exp\left[-\frac{1}{2}\xi^T\Sigma\xi\right] d\xi = \mathbb{E}\left[\partial_{x_i}\partial_{x_j}\Big|_{x=\Sigma^{1/2}X} g(x)\right]. \quad (30)$$

Finally, if g is not Schwartz but is smooth with all derivatives being polynomially bounded, we consider the convolution

$$g_\epsilon(x) := (g * \psi_\epsilon)(x), \qquad \psi_\epsilon(y) = \exp\left[-\frac{||y||^2}{2\epsilon} - \frac{1}{2}\log(2\pi\epsilon)\right].$$

Then, g_ϵ is Schwartz for all $\epsilon > 0$. Moreover, note that $g_\epsilon(\Sigma^{1/2}x)$ is also polynomially bounded for any PSD matrix Σ. Specifically, for any fixed PSD matrix Σ_0 we have for any $k \geq 1$

$$\begin{aligned}
&\sup_{\epsilon\in[0,1]} \sup_{\substack{||\Sigma-\Sigma_0||\leq 1 \\ \Sigma \text{ PSD}}} \sup_{x\in\mathbb{R}^r} \left|(1+||x||)^{-k} g_\epsilon(\Sigma^{1/2}x)\right| \\
&\quad = \sup_{\epsilon\in[0,1]} \sup_{\substack{||\Sigma-\Sigma_0||\leq 1 \\ \Sigma \text{ PSD}}} \sup_{x\in\mathbb{R}^r} \left|(1+||x||)^{-k} \int_{\mathbb{R}^r} g(\Sigma^{1/2}(x-y))\psi_\epsilon(y)dy\right| \\
&\quad \leq \sup_{\epsilon\in[0,1]} \sup_{\substack{||\Sigma-\Sigma_0||\leq 1 \\ \Sigma \text{ PSD}}} \sup_{x\in\mathbb{R}^r} \left\{(1+||x||)^{-k} \int_{\mathbb{R}^r} \left(1 + \left|\left|\Sigma^{1/2}(x-y)\right|\right|^k\right)\psi_\epsilon(y)dy\right\} \\
&\quad < \infty, \qquad (31)
\end{aligned}$$

Note that there exists $K > 0$ depending only k, r, Σ_0 so that

$$\sup_{||\Sigma-\Sigma_0||\leq 1} \left|\left|\Sigma^{1/2}(x-y)\right|\right|^k \leq K\left(1 + \left|\left|\Sigma_0^{1/2}\right|\right|\right)^k (||x||^k + ||y||^k).$$

Hence, since

$$\sup_{\epsilon\in[0,1]} \int_{\mathbb{R}^r} ||y||^k \psi_\epsilon(y)dy < \infty$$

we find that

$$\sup_{\epsilon\in[0,1]}\sup_{\substack{||\Sigma-\Sigma_0||\leq 1\\ \Sigma \text{ PSD}}}\sup_{x\in\mathbb{R}^r}\left|(1+||x||)^{-k}g_\epsilon(\Sigma^{1/2}x)\right|<\infty. \tag{32}$$

The estimate above allows us to use dominate convergence to see that for any PSD Σ

$$\mathbb{E}\left[g(\Sigma^{1/2}X)\right]=\lim_{\epsilon\to 0}\mathbb{E}\left[g_\epsilon(\Sigma^{1/2}X)\right]. \tag{33}$$

To complete the proof we note that g_ϵ and $\partial_i\partial_j g_\epsilon$ are both Schwartz for any positive ϵ. Moreover, $\partial_i\partial_j\partial_k\partial_m g_\epsilon$ satisfies (32). Hence, we conclude by applying (30) that for any PSD matrix Σ_0 there exists $C>0$ so that

$$\begin{aligned}
&\sup_{\substack{||\Sigma-\Sigma_0||\leq 1\\ \Sigma \text{ PSD}}}\sup_{\epsilon\in[0,1]}\frac{\left|\mathbb{E}\left[g_\epsilon(\Sigma^{1/2}X)\right]-\mathbb{E}\left[g_\epsilon(\Sigma_0^{1/2}X)\right]-\sum_{i,j=1}^r\mathbb{E}\left[(\partial_i\partial_j g_\epsilon)(\Sigma_0^{1/2}X)\right](\Sigma-\Sigma_0)_{ij}\right|}{||\Sigma-\Sigma_0||^2}\\
&\quad\leq\sup_{\epsilon\in[0,1]}\sup_{||\Sigma-\Sigma_0||\leq 1}\sum_{i,j,k,m=1,\ldots,r}\left|\mathbb{E}\left[(\partial_i\partial_j\partial_k\partial_m)g_\epsilon(\Sigma^{1/2}X)\right]\right|\\
&\quad\leq C.
\end{aligned}$$

Thus, if $\Sigma-\Sigma_0/\left|\left|\Sigma-\Sigma_0\right|\right|\to\Sigma_1$, we find by applying (33) to $\partial_i\partial_j g$ that

$$\begin{aligned}
\lim_{\Sigma\to\Sigma_0}\frac{\mathbb{E}\left[g(\Sigma^{1/2}X)\right]-\mathbb{E}\left[g(\Sigma_0^{1/2}X)\right]}{||\Sigma-\Sigma_0||}&=\lim_{\Sigma\to\Sigma_0}\lim_{\epsilon\to 0}\frac{\mathbb{E}\left[g_\epsilon(\Sigma^{1/2}X)\right]-\mathbb{E}\left[g_\epsilon(\Sigma_0^{1/2}X)\right]}{||\Sigma-\Sigma_0||}\\
&=\lim_{\Sigma\to\Sigma_0}\lim_{\epsilon\to 0}\left\{\sum_{i,j=1}^r\mathbb{E}\left[(\partial_i\partial_j g_\epsilon(\Sigma_0^{1/2}X))\right]\frac{(\Sigma-\Sigma_0)_{ij}}{||\Sigma-\Sigma_0||}\right\}\\
&=\sum_{i,j=1}^r\mathbb{E}\left[\partial_i\partial_j g(\Sigma_0^{1/2}X)\right](\Sigma_1)_{ij}.
\end{aligned}$$

This shows that (29) holds for any g that is smooth, completing the proof of Lemma 13. ■

5. Proof of Theorem 3

Let us first recall the notation. We fix $r\geq 1$ and assume that $\sigma:\mathbb{R}\to\mathbb{R}$ satisfies assumption 1 with this value of r. We also fix a finite collection $x_{\mathcal{A}}=\{x_\alpha,\alpha\in\mathcal{A}\}\subseteq\mathbb{R}^{n_0}$ of distinct network inputs and p directional derivatives $d_1,\ldots,d_p$ as in (6). We denote by

$$N(p,r)=\#\left\{J=(j_1,\ldots,j_p)\in\mathbb{N}^p\mid j_1+\cdots+j_p\leq r\right\},$$

which computes the number of possible derivatives of order at most r in the p directional derivatives d_j. We also denote by $\mathcal{F}^{(\ell)}$ the sigma algebra generated by the weights and biases in layers up to and including ℓ. The starting point for proving Theorem 3 is the following simple but fundamental observation.

Lemma 14 *For each $\ell \geq 0$, conditional on $\mathcal{F}^{(\ell)}$,*

$$\left\{ \left(D_\alpha^J z_{i;\alpha}^{(\ell+1)},\, \alpha \in \mathcal{A},\, |J| \leq r \right) \right\}_{i=1}^{n_{\ell+1}}$$

is a collection of $n_{\ell+1}$ iid centered Gaussians of dimension $N(p,r)$.

Proof The defining recursion (1) of a fully connected network yields for each α, J

$$D_\alpha^J z_{i;\alpha}^{(\ell+1)} = D_\alpha^J \left\{ b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(z_{j;\alpha}^{(\ell)} \right) \right\} = \delta_{|J|=0} b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} D_\alpha^J \sigma\left(z_{j;\alpha}^{(\ell)} \right). \tag{34}$$

Note that $D_\alpha^J \sigma(z_{j;\alpha}^{(\ell)})$ are measurable with respect to $\mathcal{F}^{(\ell)}$. The conclusion now follows since the weights $W_{ij}^{(\ell+1)}$, $j = 1 \ldots, n_\ell$ and bias $b_i^{(\ell+1)}$ are centered Gaussians and are independent for different i. ∎

Thus, the structure of $z_\alpha^{(\ell+1)}$ and its derivatives is always that of a Gaussian field after conditioning on $\mathcal{F}^{(\ell)}$. To ease the notation in what comes given $f : \mathbb{R}^{|\mathcal{A}| \times N(n_0,r)} \to \mathbb{R}$, let us abbreviate

$$f\left(z_{j;\mathcal{A}}^{(\ell)} \right) := f\left(D_\alpha^J z_{j;\alpha}^{(\ell)},\, \alpha \in \mathcal{A},\, |J| \leq r \right), \qquad j \in [n_\ell].$$

Next, we remind the reader that given $f : \mathbb{R}^{|\mathcal{A}| \times N(n_0,r)} \to \mathbb{R}$, which is measurable and polynomially bounded, the corresponding collective observable $\mathcal{O}_f^{(\ell)}$ at layer ℓ is

$$\mathcal{O}_f^{(\ell)} = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} f\left(z_{j;\mathcal{A}}^{(\ell)} \right)$$

and that the statement (67) in Proposition 21 ensures

$$\sup_{n \geq 1} \mathbb{E}\left[\left| \mathcal{O}_f^{(\ell)} \right| \right] < \infty. \tag{35}$$

Recall also our notation for the conditional covariance

$$\Sigma_{\alpha_1 \alpha_2}^{(\ell)} := \mathrm{Cov}\left(z_{i;\alpha_1}^{(\ell+1)}, z_{i;\alpha_2}^{(\ell+1)} \mid \mathcal{F}^{(\ell)} \right) = C_b + \frac{C_W}{n_\ell} \sum_{j=1}^{n_\ell} \sigma\left(z_{j;\alpha_1}^{(\ell)} \right) \sigma\left(z_{j;\alpha_2}^{(\ell)} \right)$$

and note that both it and its derivatives

$$\begin{aligned} D_{\alpha_1}^{J_1} D_{\alpha_2}^{J_2} \Sigma_{\alpha_1 \alpha_2}^{(\ell)} &= \mathrm{Cov}\left(D_{\alpha_1}^{J_1} z_{i;\alpha_1}^{(\ell+1)}, D_{\alpha_2}^{J_2} z_{i;\alpha_2}^{(\ell+1)} \mid \mathcal{F}^{(\ell)} \right) \\ &= D_{\alpha_1}^{J_1} D_{\alpha_2}^{J_2} \left(C_b + \frac{C_W}{n_\ell} \sum_{j=1}^{n_\ell} \sigma\left(z_{j;\alpha_1}^{(\ell)} \right) \sigma\left(z_{j;\alpha_2}^{(\ell)} \right) \right) \end{aligned}$$

are collective observables at layer ℓ. Our first application of Lemma 14 is the following reduction of the study of cumulants of $D_\alpha^J z_{i;\alpha}^{(\ell+1)}$ to the cumulants of certain collective observables.

Proposition 15 *Fix $k, \ell \geq 1$ and p-dimensional multi-indices $J_1, \ldots, J_k$ with $|J_i| \leq r$. If k is odd, then*

$$\kappa\left(D^{J_1}_{\alpha_1} z^{(\ell+1)}_{i_1;\alpha_1}, \ldots, D^{J_k}_{\alpha_k} z^{(\ell+1)}_{i_k;\alpha_k}\right) = 0$$

In contrast, if k is even

$$\kappa\left(D^{J_1}_{\alpha_1} z^{(\ell+1)}_{i_1;\alpha_1}, \ldots, D^{J_k}_{\alpha_k} z^{(\ell+1)}_{i_k;\alpha_k}\right) \;\; = \;\; \textit{finite sums of } \kappa\left(\mathcal{O}^{(\ell)}_{f_1}, \ldots, \mathcal{O}^{(\ell)}_{f_{k/2}}\right),$$

where $\mathcal{O}^{(\ell)}_{f_j}$ are collective observables of the form

$$D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}, \qquad |J_1|, |J_2| \leq r. \tag{36}$$

Proof Using (22) and recalling that $\mathcal{F}^{(\ell)}$ is the sigma algebra generated by weights and biases in layers up to and including ℓ, we have that $\kappa\left(D^{J_1}_{\alpha_1} z^{(\ell+1)}_{i_1;\alpha_1}, \ldots, D^{J_k}_{\alpha_k} z^{(\ell+1)}_{i_k;\alpha_k}\right)$ equals

$$\sum_{\pi=(\pi_1,\ldots,\pi_B)} \kappa\left(\kappa\left(\left(D^J z^{(\ell+1)}\right)_{\pi_1} \big| \mathcal{F}^{(\ell)}\right), \ldots, \kappa\left(\left(D^J z^{(\ell+1)}\right)_{\pi_B} \big| \mathcal{F}^{(\ell)}\right)\right), \tag{37}$$

where the sum is over partitions π of $[k]$ and for $b = 1, \ldots, B$ we've abbreviated

$$\left(D^J z^{(\ell+1)}\right)_{\pi_b} := \left(D^{J_t}_{\alpha_t} z^{(\ell+1)}_{i_t;\alpha_t}, \quad t \in \pi_b\right).$$

By Lemma 14, $\{(D^J_\alpha z^{(\ell+1)}_{i;\alpha}, \alpha \in \mathcal{A}, |J| \leq d), i = 1, \ldots, n_{\ell+1}\}$ are iid centered Gaussians conditional on $\mathcal{F}^{(\ell)}$. Hence, by the properties (23) and (24) and (25) from Proposition 12, in the sum over partitions above, a term is non-zero only if

$$\forall b \in [B], \qquad |\pi_b| = 2 \quad \text{and} \quad i_{\pi_b(1)} = i_{\pi_b(2)}$$

This proves that $\kappa\left(D^{J_1} z^{(\ell+1)}_{i_1;\alpha_1}, \ldots, D^{J_k} z^{(\ell+1)}_{i_k;\alpha_k}\right)$ vanishes if k is odd. To treat the case when k is even observe that by (34)

$$\kappa\left(D^{J_1} z^{(\ell+1)}_{i_1;\alpha_1}, D^{J_2} z^{(\ell+1)}_{i_2;\alpha_2} \big| \mathcal{F}^{(\ell)}\right) = \delta_{i_1 i_2} D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}.$$

Substituting this into (37) completes the proof. ■

When $k = 2$, Proposition 15 and our assumption (1) shows that for each $\ell \geq 1$, any $i_1, i_2 \in [n_{\ell+1}]$, $\alpha \in \mathcal{A}$, and multi-indices J_1, J_2 of order at most d, there exists a polynomially bounded function $f : \mathbb{R}^{|\mathcal{A}| \times N(n_0,d)} \to \mathbb{R}$ for which

$$\kappa\left(D^{J_1}_{\alpha_1} z^{(\ell+1)}_{i_1;\alpha_1}, D^{J_2}_{\alpha_2} z^{(\ell+1)}_{i_2;\alpha_2}\right) = \mathbb{E}\left[\mathcal{O}^{(\ell)}_f\right]$$

In light of (35) this proves Theorem 3 when $k = 2$. Further, since the cumulant of 2 or more random variables is shift-invariant, we may assume for $k \geq 3$ that the collective observables $D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}$ in Proposition 15 are replaced by their zero mean versions:

$$\Delta^{J_1,J_2,(\ell)}_{\alpha_1\alpha_2} := D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2} - \mathbb{E}\left[D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}\right]. \tag{38}$$

Hence, Theorem 3 is a special case of the following result.

Theorem 16 *Fix $k, m \geq 1$. Consider any m-tuple $F = (f_1, \ldots, f_m)$ consisting of measurable, functions*

$$f_i : \mathbb{R}^{|\mathcal{A}| \times N(n_0, r)} \to \mathbb{R}, \quad i = 1, \ldots, m$$

that are polynomially bounded and satisfy

$$\mathbb{E}\left[\mathcal{O}_{f_i}^{(\ell)}\right] = \mathbb{E}\left[f_i\left(z_{1;\mathcal{A}}^{(\ell)}\right)\right] = 0, \qquad i = 1, \ldots, m.$$

Define the $m-$tuple of collective observables

$$\overrightarrow{\mathcal{O}}_F^{(\ell)} := \left(\mathcal{O}_{f_i}^{(\ell)},\, i = 1, \ldots, m\right).$$

Consider further any measurable polynomially bounded functions

$$g_j : \mathbb{R}^m \to \mathbb{R}, \quad j = 1, \ldots, k.$$

which are smooth in a neighborhood of 0. If f_i and σ are in fact smooth, then, for every $\ell \geq 1$

$$\sup_{n \geq 1} \left| n^{k-1} \kappa\left(g_1\left(\overrightarrow{\mathcal{O}}_F^{(\ell)}\right), \ldots, g_k\left(\overrightarrow{\mathcal{O}}_F^{(\ell)}\right)\right)\right| < \infty \tag{39}$$

Moreover, (39) holds without the assumption that f_i, σ are smooth provided that for each ℓ the vector of iterated directional derivatives $(D_\alpha^J z_{i;\alpha}^{(\ell)},\, |J| \leq r, \alpha \in \mathcal{A})$ of order at most r is non-degenerate in the sense of (7).

Proof Our starting point is a reduction of Theorem 16 to the case when g_j are polynomials. This is related to a technique called the delta method in some parts of the mathematical statistics literature Ver Hoef (2012).

Proposition 17 (Polynomials are Enough for Theorem 16) *Fix $m \geq 1$ and suppose that for each $n \geq 1$ we have an $m-$tuple $X_n = (X_{n,1}, \ldots, X_{n,m})$ of mean 0 random variables that possess bounded moments of all orders:*

$$\sup_{n \geq 1} \left|\mathbb{E}\left[X_{n,1}^{q_1} \cdots X_{n,m}^{q_m}\right]\right| < \infty, \qquad \forall\, q_1, \ldots, q_m \geq 0. \tag{40}$$

Suppose for any given polynomials $p_1, \ldots, p_k$ in m variables we have

$$\sup_{n \geq 1} \left| n^{k-1} \kappa\left(p_1(X_n), \ldots, p_k(X_n)\right)\right| < \infty. \tag{41}$$

Then, for any measurable, polynomially bounded functions $g_j : \mathbb{R}^m \to \mathbb{R}$, $j = 1, \ldots, k$, which are smooth in some fixed neighborhood of 0

$$\sup_{n \geq 1} \left| n^{k-1} \kappa\left(g_1\left(X_n\right), \ldots, g_k\left(X_n\right)\right)\right| < \infty. \tag{42}$$

Proof We begin with the following simple Lemma, which allows us to translate between the cumulants bounds (41) and high probability bounds.

Lemma 18 *For any $q \geq 1$*

$$\sup_{n\geq 1} \sup_{1\leq i\leq m} \left| n^{\lceil \frac{q}{2} \rceil} \mathbb{E}\left[X_{i,n}^q\right] \right| < \infty.$$

Proof We have by property (26) from Proposition 12 that

$$\mathbb{E}\left[X_{i;n}^q\right] = \sum_{\substack{\pi=(\pi_1,\ldots,\pi_B)\\ \pi\in P(m)}} \prod_{b=1}^{B} \kappa\big(\underbrace{X_{i;n},\ldots,X_{i;n}}_{|\pi_b| \text{ times}}\big).$$

Since by assumption $X_{i;n}$ has mean 0, we have

$$\kappa(X_{i;n}) = \mathbb{E}\left[X_{i;n}\right] = 0.$$

Thus, the only partitions $\pi = (\pi_1, \ldots, \pi_B) \in S(m)$ that give rise to non-zero terms in the expression above must have $B \leq \lfloor \frac{q}{2} \rfloor$. Moreover, for any such partition, we have

$$\left\lceil \frac{q}{2} \right\rceil = q - \left\lfloor \frac{q}{2} \right\rfloor = -\left\lfloor \frac{q}{2} \right\rfloor + \sum_{b=1}^{B} |\pi_b| \leq \sum_{b=1}^{B} \left(|\pi_b| - 1\right).$$

Hence, we find

$$\sup_{n\geq 1} \left| n^{\lceil \frac{q}{2} \rceil} \mathbb{E}\left[X_{i;n}^q\right] \right| \leq \sum_{\substack{\pi=(\pi_1,\ldots,\pi_B)\\ \pi\in P(m),\, |\pi_b|\geq 2}} \prod_{b=1}^{B} \sup_{n\geq 1} \left| n^{|\pi_b|-1} \kappa\big(\underbrace{X_{i;n},\ldots,X_{i;n}}_{|\pi_b| \text{ times}}\big) \right| < \infty,$$

where the final inequality follows from the assumption (41). ∎

Applying Markov's inequality and Lemma 18 shows that for any $q \geq 1$ we have

$$\sup_{n\geq 1} n^q \mathbb{P}\left(S_n^c\right) < \infty, \qquad S_n := \left\{ |X_{i;n}| \leq n^{-1/4}, \qquad i = 1, \ldots, m \right\}. \tag{43}$$

This localization estimate allows us to replace each g_i by its Taylor expansion around 0. Indeed, note that

$$\kappa\left(g_1(X_n), \ldots, g_k(X_n)\right) = P\left(\mathbb{E}\left[g_1(X_n)^{q_1} \cdots g_k(X_n)^{q_k}\right], \quad q_1 + \cdots + q_k \leq k\right)$$

for some universal polynomial P evaluated at the mixed moments of X_n (the formula for this polynomial is given in (27) but is not important). Moreover, using the growth assumption (40) on X and the fact that g_i are polynomially bounded we find that

$$\sup_{n\geq 1} \mathbb{E}\left[g(X_n)^{q_1} \cdots g_k(X_n)^{q_k}\right] < \infty, \qquad \forall\, q_1, \ldots, q_k \geq 1. \tag{44}$$

This, in combination with the localization estimate (43) applied with $q = k - 1$ yields

$$\kappa\left(g_1(X_n), \ldots, g_k(X_n)\right) = P\left(\mathbb{E}\left[\mathbf{1}_{S_n} g_1(X_n)^{q_1} \cdots g_k(X_n)^{q_m}\right], \quad q_1 + \cdots + q_m \leq k\right) + O(n^{-k+1}).$$

Note that for n sufficiently large, on the event S_n, the argument X_n is in any fixed neighborhood of 0. Hence, we may write

$$g_j(X_n) = p_j(X_n) + O(n^{-k+1}),$$

where p_j represents the $j-$th order Taylor expansion of g_j around 0 with j sufficiently large (say bigger than $4k$) and the constant in the error term is uniformly bounded. This yields

$$\kappa\left(g_1(X_n), \ldots, g_k(X_n)\right) = P\left(\mathbb{E}\left[\mathbf{1}_{S_n} p_1(X_n)^{q_1} \cdots p_k(X_n)^{q_m}\right], \quad q_1 + \cdots + q_m \leq k\right) + O(n^{-k+1}).$$

Finally, using the mixed moment estimates (40) and the localization estimate (43), we conclude

$$\begin{aligned}\kappa\left(g_1(X_n), \ldots, g_k(X_n)\right) &= P\left(\mathbb{E}\left[p_1(X_n)^{q_1} \cdots p_k(X_n)^{q_m}\right], \quad q_1 + \cdots + q_m \leq k\right) + O(n^{-k+1}) \\ &= \kappa\left(p_1(X_n), \ldots, p_k(X_n)\right) + O(n^{-k+1}).\end{aligned}$$

Recalling (41) completes the proof. ■

Proposition 17 shows that, in establishing the conclusion (39) of Theorem 16, it is sufficient to assume that g_j are polynomials. The remainder of the proof of Theorem 16 is by induction on ℓ, starting with $\ell = 1$. In view of Proposition 17, the following result establishes the base case.

Proposition 19 (Base Case: Theorem 16 holds for polynomials at $\ell = 1$) *Fix $k, m \geq 1$ and suppose f_i, $i = 1, \ldots, m$ are as in the statement of Theorem 16. Then, if $p_1, \ldots, p_k$ are any polynomials in m variables, we have*

$$\sup_{n \geq 1} \left| n^{k-1} \kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right), \ldots, p_k\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right)\right)\right| < \infty.$$

Proof Since cumulants are multi-linear, we may and shall assume that p_a are monomials:

$$p_a(x) = x^{Q^{(a)}} := x_1^{q_1^{(a)}} \cdots x_m^{q_m^{(a)}}, \qquad x = (x_1, \ldots, x_m), \quad Q^{(a)} = \left(q_1^{(a)}, \ldots, q_m^{(a)}\right). \tag{45}$$

Recall that

$$\mathcal{O}_{f_i}^{(1)} := n_1^{-1} \sum_{j=1}^{n_1} f_i\left(z_{j;\mathcal{A}}^{(1)}\right).$$

Therefore, writing $q^{(a)} := q_1^{(a)} + \cdots + q_m^{(a)}$ we find

$$p_a\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right) = n_1^{-q^{(a)}} \sum_{J^{(a)}} f_{J^{(a)}}, \qquad f_{J^{(a)}} := \prod_{i=1}^{m} \prod_{q=1}^{q_i^{(a)}} f_i\big(z_{j_{q;i}^{(a)};\mathcal{A}}^{(1)}\big),$$

where the sum is over tuples of multi-indices

$$J^{(a)} = \left(J_1^{(a)}, \ldots, J_m^{(a)}\right), \qquad J_i^{(a)} = \left(j_{q;i}^{(a)} \in [n_1],\, i \in [m],\, q \in [q_i^{(a)}]\right). \tag{46}$$

Hence, using that cumulants are multi-linear (see (24)), we obtain

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right),\dots,p_k\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right)\right)=n_1^{-(q^{(1)}+\cdots+q^{(k)})}\sum_{J^{(1)},\dots,J^{(k)}}\kappa\left(f_{J^{(1)}},\dots,f_{J^{(k)}}\right),$$

where the sum extends over ordered collections $\left(J^{(a)},\, 1\le a\le k\right)$ of multi-indices as in (46). The expression on the right hand side can be interpreted as an average. Namely, we can think of the indices $j_{q;i}^{(a)}\in[n_1]$ are chosen uniformly from $[n_1]$ and independently for all i,q,a. Writing $\mathcal{E}$ for the average with respect to this distribution, we obtain

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right),\dots,p_k\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right)\right)=\mathcal{E}\left[\kappa\left(f_{J^{(1)}},\dots,f_{J^{(k)}}\right)\right].$$

Our goal is to show that this average is small. To quantify this, let us associate to each collection $\left(J^{(a)},\, a\in[k]\right)$ a graph

$$\mathcal{G}\left(J^{(a)},\, a\in[k]\right)=\left([k],\, E\left(J^{(a)},\, a\in[k]\right)\right), \tag{47}$$

with vertex set $[k]$ and edge set defined by

$$(a,a')\in\mathcal{E}\left(J^{(a)},\, a\in[k]\right)\quad\Longleftrightarrow\quad\exists i,i'\in[m],\, q\in[q_i^{(a)}],\, q'\in[q_{i'}^{(a')}]\text{ s.t. }j_{q;i}^{(a)}=j_{q';i'}^{(a')}.$$

The key point is that in light of the vanishing property (23) of cumulants and the fact that neurons at layer 1 are independent

$$\mathcal{G}\left(J^{(a)},\, a\in[k]\right)\text{ disconnected}\quad\Longrightarrow\quad\kappa\left(f_{J^{(1)}}^{(1)},\dots,f_{J^{(k)}}^{(1)}\right)=0.$$

Hence,

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right),\dots,p_k\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right)\right)=\mathcal{E}\left[\mathbf{1}_{\left\{\mathcal{G}\left(J^{(a)},\,a\in[k]\right)\text{ connected}\right\}}\kappa\left(f_{J^{(1)}}^{(1)},\dots,f_{J^{(k)}}^{(1)}\right)\right].$$

Since f_i are assumed to be polynomially bounded and the distribution of the neuron pre-activations $z_{i;\alpha}^{(1)}$ is that of centered Gaussians with mean 0 and covariance

$$\mathrm{Cov}\left(z_{i_1;\alpha_1}^{(1)},z_{i_2;\alpha_2}^{(1)}\right)=\delta_{i_1i_2}\left(C_b+\frac{C_W}{n_0}\sum_{j=1^{n_0}}x_{j;\alpha_1}x_{j;\alpha_2}\right),$$

we have for any fixed k that

$$\sup_{n\ge1}\sup_{J^{(1)},\dots,J^{(a)}}\left|\kappa\left(f_{J^{(1)}}^{(1)},\dots,f_{J^{(k)}}^{(1)}\right)\right|<\infty.$$

Hence,

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right),\dots,p_k\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right)\right)=O\left(\mathcal{P}\left(\mathcal{G}\left(J^{(a)},\, a\in[k]\right)\text{ connected}\right)\right),$$

where $\mathcal{P}$ is the probability measure associated to our random draw of $J^{(1)}, \ldots, J^{(k)}$. To complete the proof, note that since $m, q_i^{(a)}$ are fixed, by a simple union bound, we obtain

$$\mathcal{P}\left(\mathcal{G}\left(J^{(a)},\, a \in [k']\right) \text{ connected} \,\middle|\, \mathcal{G}\left(J^{(a)},\, a \in [k'-1]\right) \text{ connected}\right) = O(n^{-1}).$$

Hence,

$$\begin{aligned}
&\mathcal{P}\left(\mathcal{G}\left(J^{(a)},\, a \in [k]\right) \text{ connected}\right) \\
&\quad = \prod_{k'=2}^{k} \mathcal{P}\left(\mathcal{G}\left(J^{(a)},\, a \in [k']\right) \text{ connected} \,\middle|\, \mathcal{G}\left(J^{(a)},\, a \in [k'-1]\right) \text{ connected}\right) \\
&\quad = O(n^{-k+1}). \qquad (48)
\end{aligned}$$

Thus,

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right), \ldots, p_k\left(\overrightarrow{\mathcal{O}}_F^{(1)}\right)\right) = O(n^{-k+1}),$$

as desired. ■

Propositions 17 and 19 together show that the conclusion (39) of Theorem 16 holds at layer 1. In conjunction with Proposition 17, the following result establishes that if the conclusion (39) of Theorem 16 holds at some layer $\ell \geq 1$ then it also holds at layer $\ell + 1$. This will complete the proof by inductive of Theorem 16.

Proposition 20 (Inductive Step: Reducing to smooth cumulants) *Fix $\ell \geq 1$.*

Case 1: Suppose that σ is smooth. Assume that for any collection

$$F' = \left(f_i' : \mathbb{R}^{|\mathcal{A}| \times N(n_0, r)} \to \mathbb{R},\, i = 1, \ldots, m\right)$$

of smooth and polynomially bounded functions and any g_j as in the statement of Theorem 16 the conclusion (39) *of Theorem 16 holds at layer ℓ:*

$$\sup_{n \geq 1} \left| n^{k-1} \kappa\left(g_1\left(\overrightarrow{\mathcal{O}}_{F'}^{(\ell)}\right), \ldots, g_k\left(\overrightarrow{\mathcal{O}}_{F'}^{(\ell)}\right)\right)\right| < \infty.$$

Then, if $p_1, \ldots, p_k$ are any polynomials in m variables, and $F = (f_i,\, i = 1, \ldots, m)$ is an arbitrary collection of smooth and polynomially bounded functions $f_i : \mathbb{R}^{|\mathcal{A}| \times N(n_0, r)} \to \mathbb{R}$, then

$$\sup_{n \geq 1} \left| n^{k-1} \kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_F^{(\ell+1)}\right), \ldots, p_k\left(\overrightarrow{\mathcal{O}}_F^{(\ell+1)}\right)\right)\right| < \infty.$$

Case 2: Suppose σ is not smooth but satisfies Assumption 1 and that $(D_\alpha^J z_{i;\alpha}^{(\ell)},\, \alpha \in \mathcal{A},\, |J| \leq r)$ is non-degenerate in the infinite width in the sense of (7)*. Assume that for any collection*

$$F' = \left(f_i' : \mathbb{R}^{|\mathcal{A}| \times N(n_0, r)} \to \mathbb{R},\, i = 1, \ldots, m\right)$$

of measurable and polynomially bounded functions and any g_j as in the statement of Theorem 16 the conclusion (39) *of Theorem 16 holds at layer ℓ:*

$$\sup_{n\geq 1}\left|n^{k-1}\kappa\left(g_1\left(\overrightarrow{\mathcal{O}}_{F'}^{(\ell)}\right),\ldots,g_k\left(\overrightarrow{\mathcal{O}}_{F'}^{(\ell)}\right)\right)\right|<\infty.$$

Then, if $p_1,\ldots,p_k$ are any polynomials in m variables, and $F=(f_i,\, i=1,\ldots,m)$ is an arbitrary collection of measurable and polynomially bounded functions $f_i:\mathbb{R}^{|\mathcal{A}|\times N(n_0,d)}\to\mathbb{R}$, then

$$\sup_{n\geq 1}\left|n^{k-1}\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)}\right),\ldots,p_k\left(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)}\right)\right)\right|<\infty.$$

Proof The proof of Proposition 20 is similar but somewhat more involved than that of Proposition 19. Moreover, the two cases are proved in essentially the same way, except that we will employ the different cases in Lemma 13. We give the details in the case when σ is smooth and indicate where the proof is modified slightly to handle the non-smooth case.

To start, as in the proof of Proposition 19, note that since cumulants are multi-linear (see (24)), it is enough to assume that p_j are monomials. Thus, borrowing the notation from the proof of Proposition 19 (see starting (45)), we find

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)}\right),\ldots,p_k\left(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)}\right)\right)=n_{\ell+1}^{-(q^{(1)}+\cdots+q^{(a)})}\sum_{J^{(1)},\ldots,J^{(k)}}\kappa\left(f_{J^{(1)}}^{(\ell+1)},\ldots,f_{J^{(k)}}^{(\ell+1)}\right),$$

where

$$f_{J^{(a)}}^{(\ell+1)}:=\prod_{i=1}^{m}\prod_{q=1}^{q_i^{(a)}}f_{j_\alpha^{(a)}}^{(\ell+1)},\qquad f_j^{(\ell+1)}:=f\left(z_{j;\mathcal{A}}^{(\ell+1)}\right).$$

Note that, as in Proposition 21, the polynomially bounded assumption on f_j and the non-linearity σ together with the Gaussianity of weights and biases show that

$$\sup_{n\geq 1}\left|\kappa\left(f_{J^{(1)}}^{(\ell+1)},\ldots,f_{J^{(k)}}^{(\ell+1)}\right)\right|<\infty. \tag{49}$$

Using the law of total cumulance (22), we find that $\kappa\left(p_1(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)}),\ldots,p_k(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)})\right)$ equals

$$\sum_{\pi=(\pi_1,\ldots,\pi_B)}n_{\ell+1}^{-(q^{(1)}+\cdots+q^{(a)})}\sum_{J^{(1)},\ldots,J^{(k)}}\kappa\left(\kappa\left(f_{J^{(\pi_1)}}^{(\ell+1)}\mid\mathcal{F}^{(\ell)}\right),\ldots,\kappa\left(f_{J^{(\pi_B)}}^{(\ell+1)}\mid\mathcal{F}^{(\ell)}\right)\right),$$

where π is any partition of $[k]$ and

$$f_{J^{(\pi_b)}}:=\left(f_{J^{(a)}},\, a\in\pi_b\right).$$

Just as in the proof of Proposition 19, we may interpret the sum over $J^{(1)},\ldots,J^{(k)}$ as an average over the distribution in which $j_{q;i}^{(a)}$ are drawn iid uniformly on $[n_{\ell+1}]$. Writing $\mathcal{E}$ for averages with respect to this distribution yields

$$\kappa\left(p_1(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)}),\ldots,p_k(\overrightarrow{\mathcal{O}}_{F}^{(\ell+1)})\right)=\sum_{\pi=(\pi_1,\ldots,\pi_b)}\mathcal{E}\left[\kappa\left(\kappa\left(f_{J^{(\pi_1)}}^{(\ell+1)}\mid\mathcal{F}^{(\ell)}\right),\ldots,\kappa\left(f_{J^{(\pi_b)}}^{(\ell+1)}\mid\mathcal{F}^{(\ell)}\right)\right)\right]$$

As in (47), we may associate to each collection $J^{(\pi_t)}$ the graph $\mathcal{G}\left(J^{(\pi_t)}\right)$. Recall that by Lemma 14, the neurons pre-activations $D^J_\alpha z^{(\ell+1)}_{i;\alpha}$ in layer $\ell+1$ are independent for different i conditional on $\mathcal{F}^{(\ell)}$. Hence, in view of the vanishing property (23) of cumulants, we obtain that $\kappa\left(p_1(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F), \ldots, p_k(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F)\right)$ equals

$$\sum_{\pi=(\pi_1,\ldots,\pi_B)} \mathcal{E}\left[\mathbf{1}_{\left\{\mathcal{G}\left(J^{(\pi_b)}\right) \text{ connected } \forall b\in[B]\right\}} \kappa\left(\kappa\left(f^{(\ell+1)}_{J^{(\pi_1)}} \mid \mathcal{F}^{(\ell)}\right), \ldots, \kappa\left(f^{(\ell+1)}_{J^{(\pi_B)}} \mid \mathcal{F}^{(\ell)}\right)\right)\right]$$

Since we've assumed that f_i are smooth and polynomially bounded, Lemma 13 shows that for each $b \in [B] = \{1, \ldots, B\}$ the conditional cumulant $\kappa\left(f^{(\ell+1)}_{J^{(\pi_b)}} \mid z^{(\ell)}\right)$ is a smooth function of the centered entries $D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Delta^{(\ell)}_{\alpha_1\alpha_2}$ of the conditional covariance of $\left(D^J_\alpha z^{(\ell+1)}_{i;\mathcal{A}},\ \alpha \in \mathcal{A}, |J| \le r\right)$ given $\mathcal{F}^{(\ell)}$. Thus, since these entries are collective observables at layer ℓ we may apply the inductive hypothesis of Case 1 to find that

$$\kappa\left(p_1(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F), \ldots, p_k(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F)\right) = \sum_{\pi=(\pi_1,\ldots,\pi_B)} \mathcal{P}\left[\mathcal{G}\left(J^{(\pi_b)}\right) \text{ connected } \forall b \in [B]\right] O(n^{-B+1}).$$

Combining this with the estimate (48) shows

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F\right), \ldots, p_k\left(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F\right)\right) \ = \sum_{\pi=(\pi_1,\ldots,\pi_B)} O(n^{-B+1}) \prod_{b=1}^{B} O(n^{-|\pi_b|+1}) = O(n^{-k+1}),$$

as desired. The proof in Case 2 is almost identical. The only difference is that, we must introduce the event

$$S_n = \left\{\left|\Delta^{J_1 J_2,(\ell)}_{\alpha_1\alpha_2}\right| < n^{-1/4}\right\}. \tag{50}$$

Precisely as in the proof of Lemma 18 we find that

$$\mathbb{P}(S_n^c) = O(n^{-\infty}).$$

Hence,

$$\kappa\left(p_1\left(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F\right), \ldots, p_k\left(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F\right)\right) = \kappa\left(p_1\left(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F\right), \ldots, p_k\left(\overrightarrow{\mathcal{O}}^{(\ell+1)}_F\right) \mid S_n\right) + O(n^{-\infty}),$$

where we've implicitly used (49). Moreover, since in Case 2 we assume that the vector $\left(D^J_\alpha z^{(\ell+1)}_{i;\alpha},\ |J| \le r,\ \alpha \in \mathcal{A}\right)$ is non-degenerate in the infinite width limit in the sense of (7), we see that for n sufficiently large the covariance of $\left(D^J_\alpha z^{(\ell+1)}_{i;\alpha},\ |J| \le r,\ \alpha \in \mathcal{A}\right)$ given $\mathcal{F}^{(\ell)}$, which is the matrix with entries

$$\mathbb{E}\left[D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}\right], \qquad \alpha_1, \alpha_2 \in \mathcal{A},\ |J_1|, |J_2| \le r$$

is also non-degenerate. On the event S_n, the conditional covariance of $\left(D^J_\alpha z^{(\ell+1)}_{i;\alpha},\ |J| \le r,\ \alpha \in \mathcal{A}\right)$ given $\mathcal{F}^{(\ell)}$, which is a matrix with entries $D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}$, is also non-degenerate for all n sufficiently large. Hence, we again conclude by Lemma 13 that conditional on S_n (which is

measurable with respect to $\mathcal{F}^{(\ell)}$) for each $b \in [B]$ the conditional cumulant $\kappa\left(f^{(\ell+1)}_{J^{(\pi_b)}} \mid \mathcal{F}^{(\ell)}\right)$ is a smooth function of $D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}$, which are collective observables at layer ℓ. The remainder of the proof now proceeds in the same way as for Case 1. ■

■

6. Proof of Theorem 4

Let us recall the notation. We consider a random depth L neural network with layer widths $n_0, \ldots, n_{L+1}$ with

$$\exists c, C > 0 \text{ s.t.} \qquad cn \leq n_1, \ldots, n_L \leq Cn,$$

and a non-linearity σ that satisfies (1) for some $r \geq 1$. We also fix $p \geq 1$ directional derivatives $d_1, \ldots, d_p$ as in (6) and the corresponding vectors of iterated directional derivatives

$$D^{\leq r} z^{(\ell+1)}_{i,\mathcal{A}} := \left(D^J_\alpha z^{(\ell+1)}_{i;\alpha},\, \alpha \in \mathcal{A},\, J = (j_1, \ldots, j_p) \in \mathbb{N}^p,\, |J| \leq r \right).$$

Theorem 4 concerns, for each fixed $m, \ell \geq 1$, the expectation of a function f of of the form

$$f\left(D^{\leq r} z^{(\ell+1)}_{1,\mathcal{A}}, \ldots, D^{\leq r} z^{(\ell+1)}_{m,\mathcal{A}} \right),$$

which depends on all directional derivatives in d_i of order at most r in any m neuron preactivations at layer $\ell + 1$. We seek to show that if f is both continuous and a tempered distribution, then for all $q_* \geq 1$ we have

$$\mathbb{E}\left[f\left(D^{\leq r} z^{(\ell+1)}_{1,\mathcal{A}}, \ldots, D^{\leq r} z^{(\ell+1)}_{m,\mathcal{A}} \right) \right] = O(n^{-q_*-1}) + \tag{51}$$
$$+ \sum_{q=0}^{2q_*} \frac{(-1)^q}{2^q q!} \mathbb{E}\left[\left\langle \left(\sum_{\substack{|J|,|J'| \leq r \\ \alpha,\alpha' \in \mathcal{A}}} \Delta^{JJ',(\ell)}_{\alpha\alpha'} \sum_{j=1}^{m} \partial_{D^J_\alpha z_{j;\alpha}} \partial_{D^{J'}_{\alpha'} z_{j;\alpha'}} \right)^q f\left(D^{\leq r}_{\mathcal{A}} z_1, \ldots, D^{\leq r}_{\mathcal{A}} z_m \right) \right\rangle_{\kappa^{(\ell)}} \right].$$

We remind the reader the notation in this formula. First, we continue to denoted by $\langle \cdot \rangle_{\kappa^{(\ell)}}$ the expectation with respect to a collection of centered jointly Gaussian random vectors

$$D^{\leq r}_{\mathcal{A}} z_i = \left(D^J_\alpha z_{i;\alpha},\, \alpha \in \mathcal{A},\, |J| \leq r \right)$$

with the same covariance

$$\mathrm{Cov}\left(D^{J_1}_{\alpha_1} z_{i_1;\alpha_1},\, D^{J_2}_{\alpha_2} z_{i_2;\alpha_2} \right) = \mathrm{Cov}\left(D^{J_1}_{\alpha_1} z^{(\ell)}_{i_1;\alpha_1},\, D^{J_2}_{\alpha_2} z^{(\ell)}_{i_2;\alpha_2} \right) = \delta_{i_1 i_2} \kappa^{J_1 J_2,(\ell)}_{\alpha_1\alpha_2}$$

as the true vectors of derivatives $D^{\leq r}_{\mathcal{A}} z^{(\ell)}_{i;\mathcal{A}}$ in each component separately but zero covariance for different i. Second,

$$\kappa^{(\ell)} = \mathbb{E}\left[\Sigma^{\leq r,(\ell)} \right], \qquad \Sigma^{\leq r,(\ell)} = \left(D^J_\alpha D^{J'}_{\alpha'} \Sigma^{(\ell)}_{\alpha\alpha'} \right)_{\substack{|J|,|J'| \leq r \\ \alpha,\alpha' \in \mathcal{A}}}, \tag{52}$$

is an average of the conditional covariances

$$D^J_\alpha D^{J'}_{\alpha'} \Sigma^{(\ell)}_{\alpha\alpha'} := \mathrm{Cov}\left(D^J_\alpha z^{(\ell+1)}_{1;\alpha},\, D^{J'}_{\alpha'} z^{(\ell+1)}_{1;\alpha'} \mid \mathcal{F}^{(\ell)} \right) = D^J_\alpha D^{J'}_{\alpha'} \left\{ C_b + \frac{C_W}{n_\ell} \sum_{j=1}^{n_\ell} \sigma\left(z^{(\ell)}_{j;\alpha}\right) \sigma\left(z^{(\ell)}_{j;\alpha'}\right) \right\}.$$

Finally, $\Delta^{JJ',(\ell)}_{\alpha\alpha'}$ measures the corresponding fluctuations:

$$\Delta^{JJ',(\ell)}_{\alpha\alpha'} := D^J_\alpha D^{J'}_{\alpha'} \Sigma^{(\ell)}_{\alpha\alpha'} - \mathbb{E}\left[D^J_\alpha D^{J'}_{\alpha'} \Sigma^{(\ell)}_{\alpha\alpha'} \right],$$

and we collect $\Delta^{JJ',(\ell)}_{\alpha\alpha'}$ into a matrix as follows:

$$\Delta^{\leq r,(\ell)} := \left(\Delta^{JJ',(\ell)}_{\alpha\alpha'} \right)_{\substack{|J|,|J'|\leq r \\ \alpha,\alpha' \in \mathcal{A}}}.$$

Our first step is to note that since the weights and biases in layer $\ell+1$ are Gaussian, independent of one another, and independent of the sigma algebra $\mathcal{F}^{(\ell)}$ generated by all prior weights and biases, we may write

$$\mathbb{E}\left[f\left(D^{\leq r} z^{(\ell+1)}_{1;\mathcal{A}}, \ldots, D^{\leq r} z^{(\ell+1)}_{m;\mathcal{A}} \right) \right] = \mathbb{E}\left[f\left(\left(\Sigma^{\leq r,(\ell)}\right)^{1/2} Z_1, \ldots, \left(\Sigma^{\leq r,(\ell)}\right)^{1/2} Z_m \right) \right],$$

where $Z_1, \ldots, Z_m$ are standard Gaussians which are independent of one another and of $\Sigma^{\leq r,(\ell)}$. Moreover, because $\Sigma^{\leq r,(\ell)}$ is PSD the relation (52) ensures

$$\ker(\kappa^{(\ell)}) \subseteq \ker(\Sigma^{\leq r,(\ell)}) \text{ a.s.},$$

where we recall our standing notation that $\kappa^{(\ell)} = \mathbb{E}\left[\Sigma^{\leq r}, (\ell)\right]$. By decomposing

$$Z_i = Z_{i;||} + Z_{i;\perp}, \qquad Z_{i;||} \in \ker(\kappa^{(\ell)}),\, Z_{i;\perp} \in \ker(\kappa^{(\ell)})^\perp$$

and writing $\Sigma^{\leq r,(\ell)}_\perp$ for the compression of $\Sigma^{\leq r,(\ell)}$ onto $\ker(\kappa^{(\ell)})^\perp$ we obtain by a slight abuse of notation that

$$\mathbb{E}\left[f\left(D^{\leq r} z^{(\ell+1)}_{1;\mathcal{A}}, \ldots, D^{\leq r} z^{(\ell+1)}_{m;\mathcal{A}} \right) \right] = \mathbb{E}\left[f\left(\left(\Sigma^{\leq r,(\ell)}_\perp\right)^{1/2} Z_{1,\perp}, \ldots, \left(\Sigma^{\leq r,(\ell)}_\perp\right)^{1/2} Z_{m,\perp} \right) \right].$$

The key point is now that $Z_{i;\perp}$ are standard Gaussian vectors supported on a subspace on which $\kappa^{(\ell)}$ is strictly positive definite and that $\Sigma^{\leq r,(\ell)}_\perp$ maps this subspace into itself. Consider the event

$$S_n = \left\{ \left|\Delta^{JJ',(\ell)}_{\alpha\alpha'}\right| < n^{-1/4},\, \alpha,\alpha' \in \mathcal{A},\, |J|, \left|J'\right| \leq r \right\} = \left\{ \left|\left| \kappa^{(\ell)} - \Sigma^{\leq r,(\ell)} \right|\right|_\infty < n^{-1/4} \right\}.$$

Note that, by applying Theorem 16 and arguing exactly as in Lemma 18, we find that since $\Delta^{JJ',(\ell)}_{\alpha\alpha'}$ are centered collective observables,

$$\forall q \geq 1\ \exists C_q > 0 \text{ s.t. } \mathbb{P}(S_n^c) \leq C_q \cdot n^{-q},$$

which we summarize by writing that $\mathbb{P}(S_n^c) = O(n^{-\infty})$. Since f is a tempered distribution and a continuous function, its expectation against any Gaussian is finite and we therefore have

$$\mathbb{E}\left[f\left(D^{\leq r} z_{1;\mathcal{A}}^{(\ell+1)}, \ldots, D^{\leq r} z_{m;\mathcal{A}}^{(\ell+1)}\right)\right] = \mathbb{E}\left[\mathbf{1}_{S_n} f\left(\left(\Sigma_{\perp}^{\leq r,(\ell)}\right)^{1/2} Z_{1,\perp}, \ldots, \left(\Sigma_{\perp}^{\leq r,(\ell)}\right)^{1/2} Z_{m,\perp}\right)\right],$$

plus an error of size $O(n^{-\infty})$. Let us denote by $\widehat{f}(\xi_1, \ldots, \xi_m)$ the Fourier transform of f and abbreviate

$$\xi = (\xi_1, \ldots, \xi_m), \qquad ||\xi||^2 := \sum_{i=1}^m ||\xi_i||^2, \qquad d\xi := d\xi_1 \cdots d\xi_m$$

For a $C > 0$ that we will choose later let us write

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{S_n} f\left(\left(\Sigma_{\perp}^{\leq r,(\ell)}\right)^{1/2} Z_{1,\perp}, \ldots, \left(\Sigma_{\perp}^{\leq r,(\ell)}\right)^{1/2} Z_{m,\perp}\right)\right] \\
&= \int \widehat{f}(\xi)\mathbb{E}\left[\mathbf{1}_{S_n} \exp\left[-\frac{1}{2}\sum_{i=1}^m \xi_i^T \Sigma_{\perp}^{\leq r,(\ell)} \xi_i\right]\right] d\xi \\
&= \int_{||\xi||^2 > C\log(n)} \widehat{f}(\xi)\mathbb{E}\left[\mathbf{1}_{S_n} \exp\left[-\frac{1}{2}\sum_{i=1}^m \xi_i^T \Sigma_{\perp}^{\leq r,(\ell)} \xi_i\right]\right] d\xi \\
&+ \int_{||\xi||^2 \leq C\log(n)} \widehat{f}(\xi)\mathbb{E}\left[\mathbf{1}_{S_n} \exp\left[-\frac{1}{2}\sum_{i=1}^m \xi_i^T \Sigma_{\perp}^{\leq r,(\ell)} \xi_i\right]\right] d\xi \\
&=: I_C + II_C.
\end{aligned}$$

Let us now check that

$$\forall q \geq 1 \ \exists C = C(q) \text{ s.t. } I_C = O(n^{-q}). \tag{53}$$

By the fundamental structure theorem of tempered distributions (see e.g. Friedlander et al. (1998)), there exist bounded continuous function $u_{I,J}$ and an integer $o(f)$, called the order of f, such that

$$\widehat{f}(\xi) = \sum_{\substack{I,J \\ |I,J| \leq o(f)}} \xi^I D^J u_{I,J}(\xi), \tag{54}$$

where where the derivatives D^J with respect to $\xi_1, \ldots, \xi_m$ are defined in the weak sense and ξ raised to a multi-index I denotes the corresponding monomial. Thus, we may use (54) to write

$$\begin{aligned}
|I_C| &= \left| \sum_{\substack{I,J \\ |I|,|J| \leq o(f)}} \int_{||\xi||^2 > C\log(n)} u_{I,J}(\xi) D^J \left(\xi^I \mathbb{E}\left[\mathbf{1}_{S_n} \exp\left[-\frac{1}{2}\sum_{i=1}^m \xi_i^T \Sigma_{\perp}^{\leq r,(\ell)} \xi_i\right]\right]\right) d\xi \right| \\
&\leq \sum_{\substack{I,J \\ |I|,|J| \leq o(f)}} ||u_{I,J}||_\infty \int_{||\xi||^2 > C\log(n)} \mathbb{E}\left[\mathbf{1}_{S_n} \left|p_{o(f)}(\xi)\right| \exp\left[-\frac{1}{2}\sum_{i=1}^m \xi_i^T \Sigma_{\perp}^{\leq r,(\ell)} \xi_i\right]\right] d\xi,
\end{aligned}$$

where $p_{o(f)}$ is some polynomial of degree at most $2o(f)$ in the variables $\xi_1, \ldots, \xi_m$ in which the coefficients are themselves polynomials the entries of $\Sigma_{\perp}^{\leq r,(\ell)}$. On the event S_n, entries of $\Sigma_{\perp}^{\leq r,(\ell)}$ are uniformly bounded in n since by Theorem 16 the entries of $\kappa^{(\ell)}$ are uniformly bounded in n and the event S_n guarantees that the difference $\kappa^{(\ell)} - \Sigma^{\leq r,(\ell)}$ is small for all large n. In particular, for some $T > 0$ we may write

$$\mathbf{1}_{S_n} \left| p_{o(f)} \left(\xi_1, \ldots, \xi_m \right) \right| \leq \mathbf{1}_{S_n} T \left(1 + ||\xi||^2 \right)^{o(f)}.$$

Note moreover that for all n sufficiently large, on the event S_n, we have that for some $\lambda_0 > 0$ and any $\xi \in \ker(\kappa^{(\ell)})^{\perp}$ that

$$\frac{1}{2} \xi^T \Sigma_{\perp}^{\leq r,(\ell)} \xi \geq \lambda_0 \left| |\xi| \right|^2.$$

Hence, passing to polar coordinates, we find that

$$I_C \leq T \sum_{\substack{I,J \\ |I|,|J| \leq o(f)}} ||u_{I,J}||_{\infty} \int_{r^2 > C \log(n)} (1 + r^2)^{o(f) + mN(r,p)|\mathcal{A}|-1} e^{-\lambda_0 r^2} dr,$$

where we recall that $N(r,p)$ is the number of derivatives of order at most r in the p vector fields $d_1, \ldots, d_p$. Thus, we conclude that that for any $q \geq 1$ there indeed exists $C = C(q), C' = C'(q)$ such that

$$I_C \leq C' n^{-q},$$

confirming (53). We therefore define $C := C(q_* + 1)$ and rewrite II_C as follows:

$$\begin{aligned} II_C &= \int_{||\xi||^2 \leq C \log(n)} \widehat{f}(\xi) \mathbb{E} \left[\mathbf{1}_{S_n} \exp \left[-\frac{1}{2} \sum_{i=1}^m \xi_i^T \Sigma_{\perp}^{\leq r,(\ell)} \xi_i \right] \right] d\xi \\ &= \int_{||\xi||^2 \leq C \log(n)} \widehat{f}(\xi) \exp \left[-\frac{1}{2} \sum_{i=1}^m \xi_i^T \kappa^{(\ell)} \xi_i \right] \mathbb{E} \left[\mathbf{1}_{S_n} \exp \left[-\frac{1}{2} \sum_{i=1}^m \xi_i^T \Delta^{\leq r,(\ell)} \xi_i \right] \right] d\xi. \end{aligned}$$

Note that on the event S_n there exists $T > 0$ so that

$$\sup_{||\xi||^2 \leq C \log(n)} \sum_{i=1}^m \xi_i^T \Delta_{\mathcal{A}}^{\leq r,(\ell)} \xi_i \leq CTm \left| \mathcal{A} \right|^2 \frac{\log(n)}{n^{1/4}}.$$

Hence, we may choose $Q^* = Q^*(q^*, C, |\mathcal{A}|) \geq 1$ so that

$$\mathbb{E} \left[\mathbf{1}_{S_n} \exp \left[-\frac{1}{2} \sum_{i=1}^m \xi_i^T \Delta_{\mathcal{A}}^{\leq r,(\ell)} \xi_i \right] \right] = \mathbb{E} \left[\mathbf{1}_{S_n} \sum_{q=0}^{Q_*} \frac{(-1)^q}{2^q q!} \left(\sum_{i=1}^m \xi_i^T \Delta_{\mathcal{A}}^{\leq r,(\ell)} \xi_i \right)^q \right] + O(n^{-q_* - 1}). \tag{55}$$

We thus conclude that II_C equals

$$\sum_{q=0}^{Q_*} \frac{(-1)^q}{2^q q!} \int_{||\xi||^2 \leq C \log(n)} \mathbb{E} \left[\left(\sum_{i=1}^m \xi_i^T \Delta_{\mathcal{A}}^{\leq r,(\ell)} \xi_i \right)^q \right] \widehat{f}(\xi) \exp \left[-\frac{1}{2} \sum_{i=1}^m \xi_i^T \kappa^{(\ell)} \xi_i \right] d\xi$$

plus an error of size $O(n^{-q_*-1})$. Note also that by applying Lemma 18 we have

$$\mathbb{E}\left[\mathbf{1}_{S_n}\left(\sum_{i=1}^{m}\xi_i^T\Delta_{\mathcal{A}}^{(\ell)}\xi_i\right)^q\right] = O\left(||\xi||^{2q}\, n^{-\lceil\frac{q}{2}\rceil}\right).$$

Hence, since $\widehat{f}$ is a tempered distribution and $\kappa^{(\ell)}$ is strictly positive definite on $\ker(\kappa^{(\ell)})^{\perp}$, the terms corresponding to $2q^*+1 \leq q \leq Q^*$ in (55) are of size $O(n^{-q_*-1})$. Moreover, applying the same reasoning as we used to bound I_C, by incurring another error of order $O(n^{-q_*-1})$ we may drop the restriction in II_C that $||\xi||^2 \leq C\sqrt{\log(n)}$. All together, II_C therefore equals

$$\sum_{q=0}^{2q_*}\frac{(-1)^q}{2^q q!}\int \mathbb{E}\left[\left(\sum_{i=1}^{m}\xi_i^T\Delta_{\mathcal{A}}^{\leq r,(\ell)}\xi_i\right)^q\right]\widehat{f}(\xi)\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\xi_i^T\kappa^{(\ell)}\xi_i\right]d\xi.$$

plus an error of size $O(n^{-q_*-1})$. Using that multiplication by components of ξ_i acting on the Fourier transform corresponds to differentiation of with respect to the variables $\{D_\alpha^J z_{i;\alpha},\, \alpha\in\mathcal{A},\, |J|\leq r\}$, yields the desired expression (51) and completes the proof of Theorem 4. □

7. Proof of Corollary 6

The goal of this section is to derive recursions for

$$\kappa_{2k;\alpha}^{(\ell+1)} = \kappa_k\big(\underbrace{\Delta_{\alpha\alpha}^{(\ell)},\ldots,\Delta_{\alpha\alpha}^{(\ell)}}_{k\text{ times}}\big),$$

where we defined $\Delta_{\alpha\alpha}^{(\ell)}$ in (11). Let us write

$$X_j := \sigma\left(z_{j;\alpha}^{(\ell)}\right)^2 - \mathbb{E}\left[\sigma\left(z_{j;\alpha}^{(\ell)}\right)^2\right]$$

so that

$$\Delta_{\alpha\alpha}^{(\ell)} = \frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}X_j.$$

By symmetry, we then have

$$\kappa_{4;\alpha}^{(\ell+1)} = \mathbb{E}\left[\left(\Delta_{\alpha\alpha}^{(\ell)}\right)^2\right] = \frac{C_W^2}{n_\ell}\mathbb{E}\left[X_1^2\right] + C_W^2\left(1-n_\ell^{-1}\right)\mathbb{E}\left[X_1X_2\right] \tag{56}$$

$$\begin{aligned}\kappa_{6;\alpha}^{(\ell+1)} &= \mathbb{E}\left[\left(\Delta_{\alpha\alpha}^{(\ell)}\right)^3\right] \\ &= \frac{C_W^3}{n_\ell^2}\mathbb{E}\left[X_1^3\right] + \frac{3C_W^3}{n_\ell}\left(1-\frac{1}{n_\ell}\right)\mathbb{E}\left[X_1^2X_2\right] + C_W^3\left(1-\frac{1}{n_\ell}\right)\left(1-\frac{2}{n_\ell}\right)\mathbb{E}\left[X_1X_2X_3\right]\end{aligned} \tag{57}$$

$$\begin{aligned}\kappa_{8;\alpha}^{(\ell+1)} &= \mathbb{E}\left[\left(\Delta_{\alpha\alpha}^{(\ell)}\right)^4\right] - 3\mathbb{E}\left[\left(\Delta_{\alpha\alpha}^{(\ell)}\right)^2\right]^2 = \frac{C_W^4}{n_\ell^3}\left(\mathbb{E}\left[X_1^4\right] - 3\mathbb{E}\left[X_1^2\right]^2\right) \\ &\quad + \frac{C_W^4}{n_\ell^2}\left(1-\frac{1}{n_\ell}\right)\left[\binom{4}{2}\left\{\mathbb{E}\left[X_1^2X_2\right]^2 - \mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_2^2\right] - 2\mathbb{E}\left[X_1X_2\right]^2\right\}\right. \\ &\qquad\qquad \left. + \binom{4}{1}\left\{\mathbb{E}\left[X_1^3X_2\right] - \mathbb{E}\left[X_1^3\right]\mathbb{E}\left[X_2\right] - 2\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1X_2\right]\right\}\right] \\ &\quad + \frac{C_W^4}{n_\ell}\left(1-\frac{1}{n_\ell}\right)\left(1-\frac{2}{n_\ell}\right)\binom{4}{2}\left\{\mathbb{E}\left[X_1^2X_2X_3\right] - \mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1X_2\right] - 2\mathbb{E}\left[X_1X_2\right]^2\right\} \\ &\quad + C_W^4\left(1-\frac{1}{n_\ell}\right)\left(1-\frac{2}{n_\ell}\right)\left(1-\frac{3}{n_\ell}\right)\mathbb{E}\left[X_1X_2X_3X_4\right].\end{aligned} \tag{58}$$

To evaluate the mixed moments of X_i that appear in (56)-(58), we use Theorem 4 in the case $g \equiv 1, r = 0, q_* = 1$. In this setting, if f is a continuous function and a tempered distribution, we find

$$\begin{aligned}\mathbb{E}\left[f\left(z_{1;\alpha}^{(\ell)},\ldots,z_{m;\alpha}^{(\ell)}\right)\right] &= \left\langle f\left(z_{1;\alpha},\ldots,z_{m;\alpha}\right)\right\rangle_{\kappa^{(\ell)}} \\ &\quad + \frac{\kappa_{4;\alpha}^{(\ell)}}{2^2\cdot 2!}\left\langle\left(\sum_{j=1}^m \partial_{z_{j;\alpha}}^2\right)^2 f\left(z_{1;\alpha},\ldots,z_{m;\alpha}\right)\right\rangle_{\kappa^{(\ell)}} \\ &\quad + \frac{\kappa_{6;\alpha}^{(\ell)}}{2^3\cdot 3!}\left\langle\left(\sum_{j=1}^m \partial_{z_{j;\alpha}}^2\right)^3 f\left(z_{1;\alpha},\ldots,z_{m;\alpha}\right)\right\rangle_{\kappa^{(\ell)}} \\ &\quad + \frac{\kappa_{8;\alpha}^{(\ell)} + 3\left(\kappa_{4;\alpha}^{(\ell)}\right)^2}{2^4\cdot 4!}\left\langle\left(\sum_{j=1}^m \partial_{z_{j;\alpha}}^2\right)^4 f\left(z_{1;\alpha},\ldots,z_{m;\alpha}\right)\right\rangle_{\kappa^{(\ell)}} + O(n^{-4}).\end{aligned} \tag{59}$$

We remind the reader that, by definition, $z_{1;\alpha},\ldots,z_{m;\alpha}$ are iid centered Gaussians with variance $\kappa_{\alpha\alpha}^{(\ell)}$. Since the derivations of (16)-(18) are very similar, let us give the details for only cases of $\kappa_{4;\alpha}^{(\ell)}$ and $\kappa_{6;\alpha}^{(\ell)}$. We have, using (59), that

$$\begin{aligned}\kappa_{\alpha\alpha}^{(\ell+1)} = \mathbb{E}\left[\left(\Delta_{\alpha\alpha}^{(\ell)}\right)^2\right] &= \frac{C_W^2}{n_\ell}\left(\left\langle\sigma^4\right\rangle_{\kappa_{\alpha\alpha}^{(\ell)}} - \left\langle\sigma^2\right\rangle_{\kappa_{\alpha\alpha}^{(\ell)}}^2\right) \\ &\quad + C_W^2\left(1-n_\ell^{-1}\right)\left(\left\langle X_1\right\rangle_{\kappa_{\alpha\alpha}^{(\ell)}}^2 + \frac{1}{4}\left\langle\partial^2\sigma^2\right\rangle_{\kappa_{\alpha\alpha}^{(\ell)}}^2\kappa_{4;\alpha}^{(\ell)}\right) + O(n^{-2}).\end{aligned} \tag{60}$$

Next, note that by Theorem 4 we have

$$\kappa^{(\ell)}_{\alpha\alpha} = K^{(\ell)}_{\alpha\alpha} + O(n^{-1}).$$

Moreover, note that since $x_\alpha \neq 0$, we have $K^{(\ell)}_{\alpha\alpha}$ is non-zero. Hence, Gaussian integration by parts yields that for any measureable polynomially bounded f we have

$$\langle f \rangle_{\kappa^{(\ell)}_{\alpha\alpha}} = \langle f \rangle_{K^{(\ell)}_{\alpha\alpha}} + O(n^{-1}). \tag{61}$$

Note also that for all i by applying (59) we have

$$\langle X_i \rangle_{\kappa^{(\ell)}} = -\frac{1}{8}\kappa^{(\ell)}_{4;\alpha} \left\langle \partial^4 X_1 \right\rangle_{K^{(\ell)}} + O(n^{-2}). \tag{62}$$

Thus, recalling the definition (72) of $\chi^{(\ell)}_{||;\alpha}$ the estimate (60) immediately yields (16). Next, using (59) as well as (61) and (62) that

$$C_W^3 \mathbb{E}\left[X_1^3\right] = C_W^3 \left\langle X_1^3 \right\rangle_{K^{(\ell)}_{\alpha\alpha}} + O(n^{-1}) = T^{(\ell)}_{0,3;\alpha} + O(n^{-1}).$$

Further, we seek to evaluate $\mathbb{E}\left[X_1^2 X_2\right]$ up to errors of size $O(n^{-2})$. We apply (59) as well as (61) and (62) to obtain

$$\begin{aligned} C_W^3 \mathbb{E}\left[X_1^2 X_2\right] &= C_W^3 \left\langle X_1^2 \right\rangle_{\kappa^{(\ell)}} \langle X_2 \rangle_{\kappa^{(\ell)}} + O(n^{-2}) \\ &\quad + \frac{C_W^3}{8}\kappa^{(\ell)}_{4;\alpha} \left[\left\langle X_1^2 \right\rangle_{K^{(\ell)}} \left\langle \partial^4 X_2 \right\rangle_{K^{(\ell)}} + 2 \left\langle \partial^2 X_1^2 \right\rangle_{K^{(\ell)}} \left\langle \partial^2 X_2 \right\rangle_{K^{(\ell)}} \right] \\ &\quad + \frac{1}{2} T^{(\ell)}_{2,2;\alpha} \chi^{(\ell)}_{||;\alpha} \kappa^{(\ell)}_{4;\alpha} + O(n^{-2}). \end{aligned}$$

Finally, we must evaluate $\mathbb{E}\left[X_1 X_2 X_3\right]$ up to errors of size $O(n^{-3})$. Again using (62), we find

$$\begin{aligned} C_W^3 \mathbb{E}\left[X_1 X_2 X_3\right] &= C_W^3 \langle X_1 \rangle_{\kappa^{(\ell)}} \langle X_2 \rangle_{\kappa^{(\ell)}} \langle X_3 \rangle_{\kappa^{(\ell)}} \\ &\quad + \frac{C_W^3}{8}\kappa^{(\ell)}_{4;\alpha} \left[6 \langle X_1 \rangle_{\kappa^{(\ell)}} \left\langle \partial^2 X_2 \right\rangle^2_{\kappa^{(\ell)}} + O(n^{-2}) \right] \\ &\quad + \frac{C_W^3}{48}\kappa^{(\ell)}_{6;\alpha} \left[6 \left\langle \partial^2 X_1 \right\rangle^3_{\kappa^{(\ell)}} + O(n^{-1}) \right] \\ &\quad + \frac{9 C_W^3}{8} \left(\chi^{(\ell)}_{||;\alpha} \kappa^{(\ell)}_{4;\alpha} \right)^2 + O(n^{-3}) \\ &= \frac{3}{4} T^{(\ell)}_{4,1;\alpha} \left(\chi^{(\ell)}_{||;\alpha} \kappa^{(\ell)}_{4;\alpha} \right)^2 \left\langle \partial^4 X_1 \right\rangle_{K^{(\ell)}} + \left(\chi^{(\ell)}_{||;\alpha} \right)^3 \kappa^{(\ell)}_{6;\alpha} + O(n^{-3}). \end{aligned}$$

This completes the derivation of the recursion for $\kappa^{(\ell)}_{6;\alpha}$. □

7.1 Proof of Corollary 8: 2nd, 4th Cumulants at Large Depth for the $K_* = 0$ Universality Class

In this section, we complete the proof of Corollary 8. The results when σ is the ReLU follow directly form the exact formula (76). For non-linearities such as tanh in the $K_* = 0$

universality class, our starting point is to observe that for every $x_\alpha \neq 0$ and $\delta \in (0,1)$ we have

$$K_{\alpha\alpha}^{(\ell+1)} = \frac{1}{a\ell} + O_\delta(\ell^{-2+\delta}), \tag{63}$$

where the implicit constant depends on δ, x_α and the constant a is defined in (79). This result was already derived in (5.93) of Roberts et al. (2022) (see Proposition 22 for a mathematically complete proof). In order to prove Corollary (8) for $k = 4$ we must therefore show that

$$\kappa_{4;\alpha}^{(\ell)} = \frac{2}{3n\ell a^2}(1 + O(\ell^{-1})) = \frac{2\ell}{3n}\left(K_{\alpha\alpha}^{(\ell)}\right)^2 (1 + O(\ell^{-1})). \tag{64}$$

The proof of this estimate is a straightforward calculation using Theorem 4 and the technical Lemma 23. Indeed, Theorem 4 shows

$$\kappa_{4;\alpha}^{(\ell+1)} = \frac{C_W^2}{n_\ell}\left(\left\langle \sigma^4 \right\rangle_{\kappa^{(\ell)}} - \left\langle \sigma^2 \right\rangle_{\kappa^{(\ell)}}^2\right) + \left(\chi_{||;\alpha}^{(\ell)}\right)^2 \kappa_{4;\alpha}^{(\ell)}$$

plus errors of size $O(n^{-2})$. A direction computation then yields

$$\frac{C_W^2}{n_\ell}\left(\left\langle \sigma^4 \right\rangle_{\kappa^{(\ell)}} - \left\langle \sigma^2 \right\rangle_{\kappa^{(\ell)}}^2\right) = \frac{2}{n_\ell}\left(\kappa_{\alpha\alpha}^{(\ell)}\right)^2 \left(1 + O(\ell^{-1})\right).$$

Hence, setting $n_\ell = n$ we may apply Lemma 23 to obtain

$$\kappa_{4;\alpha}^{(\ell)} = \frac{2}{3n\ell a^2}(1 + O(\ell^{-1})).$$

The proof of Corollary 8 for $k = 6, 8$ is similar and is left to the reader.

8. Acknowledgements

The gratefully acknowledges support from NSF CAREER grant DMS-2143754, NSF grants DMS-1855684, DMS-2133806, and an ONR MURI on Foundations of Deep Learning.

References

Gernot Akemann and Zdzislaw Burda. Universal microscopic correlation functions for products of independent ginibre matrices. *Journal of Physics A: Mathematical and Theoretical*, 45(46):465201, 2012.

Gernot Akemann, Mario Kieburg, Adam Mielke, and Tomaž Prosen. Universal signature from integrability to chaos in dissipative open quantum systems. *Physical review letters*, 123(25):254101, 2019.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*, 2021.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *arXiv preprint arXiv:2304.03408*, 2023.

David R Brillinger. The calculation of cumulants via conditioning. *Annals of the Institute of Statistical Mathematics*, 21(1):215–218, 1969.

David R Brillinger. *Time series: data analysis and theory*. SIAM, 2001.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

Hugo Cui, Florent Krzakala, and Lenka Zdeborová. Optimal learning of deep random networks of extensive-width. *arXiv preprint arXiv:2302.00375*, 2023.

Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *ICML*, 2018.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Friedrich Gerard Friedlander, Mark Suresh Joshi, and M Joshi. *Introduction to the Theory of Distributions*. Cambridge University Press, 1998.

Vadim Gorin and Yi Sun. Gaussian fluctuations for products of random matrices. *arXiv preprint arXiv:1812.06532*, 2018.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, 2018.

Boris Hanin. Random neural networks in the infinite width limit as gaussian processes. *arXiv preprint arXiv:2107.01562*, 2021.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *ICLR 2020*, 2020a.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020b.

Boris Hanin and Grigoris Paouris. Non-asymptotic results for singular values of gaussian matrix products. *Geometric and Functional Analysis*, 31(2):268–324, 2021.

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.

Boris Hanin and Alexander Zlokapa. Bayesian interpolation with deep linear networks. *arXiv preprint arXiv:2212.14457*, 2022.

Boris Hanin, Ryan Jeong, and David Rolnick. Deep relu networks preserve expected length. *ICLR*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Donald O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 1949.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning*, pages 4542–4551. PMLR, 2020.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in neural information processing systems*, pages 1945–1953, 2017.

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.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *ICML 2018 andarXiv:1711.00165*, 2018.

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.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Mufan Li, Mihai Nica, and Dan Roy. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization. *Advances in Neural Information Processing Systems*, 34, 2021.

Mufan Bill Li, Mihai Nica, and Daniel M Roy. The neural covariance sde: Shaped infinite depth-and-width networks at initialization. *NeurIPS 2022*, 2022.

Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Dang-Zheng Liu, Dong Wang, and Lun Zhang. Bulk and soft-edge universality for singular values of products of Ginibre random matrices. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1734 – 1762, 2016.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

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

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

Daniel S Park, Jascha Sohl-Dickstein, Quoc V Le, and Samuel L Smith. The effect of network width on stochastic gradient descent and generalization: an empirical study. *arXiv preprint arXiv:1905.03776*, 2019.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

Maithra Raghu, Ben Poole, Jon M. Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 2847–2854, 2017.

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

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Advances in neural information processing systems*, 31, 2018.

Inbar Seroussi and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *arXiv preprint arXiv:2112.15383*, 2021.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 2021.

Jay M Ver Hoef. Who invented the delta method? *The American Statistician*, 66(2): 124–127, 2012.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

Sho Yaida. Non-gaussian processes and neural networks at finite widths. *MSML*, 2020.

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

Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *arXiv preprint arXiv:1910.12478*, 2019b.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

Greg Yang and Sam S Schoenholz. Deep mean field theory: Layerwise variance and width variation as methods to control gradient explosion. 2018.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Jacob Zavatone-Veth and Cengiz Pehlevan. Exact marginal prior distributions of finite bayesian neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Quadratic models for understanding neural network dynamics. *arXiv preprint arXiv:2205.11787*, 2022.

Appendix A. Proof of Theorem 2

Our proof of Theorem 2 closely follows the proof of Theorem 1.2 in Hanin (2021). Let us first recall the notation and assumptions. We fix $\sigma : \mathbb{R} \to \mathbb{R}$ such that

- There exists $r \geq 1$ so that the r-th derivative of σ belongs to L^∞.
- There exist $c, c' > 0$ so that

$$\left|\left| e^{-cx^{2-c'}} \frac{d^r}{dx^r}\sigma(x) \right|\right|_{L^\infty} < \infty.$$

We take $C_b \geq$ and $C_W > 0$ and consider a random depth L neural network with input dimension n_0, output dimension n_{L+1}, hidden layer widths satisfying

$$\exists c, C > 0 \text{ s.t. } cn \leq n_1, \ldots, n_L \leq Cn, \qquad n \gg 1,$$

non-linearity σ and random weights and biases as in (2). We also fix a finite collection

$$x_{\mathcal{A}} := \{x_\alpha, \quad \alpha \in \mathcal{A}\}$$

of distinct network inputs as well as an integer m and study for each ℓ the random vectors

$$D^{\leq r} z^{(\ell)}_{\mathcal{A}} := \left(D^{\leq r} z_{i;\mathcal{A}}, \quad i = 1, \ldots, m\right),$$

where

$$D^{\leq r} z_{i;\mathcal{A}} := \left(D^J_\alpha z^{(\ell)}_{i;\alpha}, \quad \alpha \in \mathcal{A},\, i = 1, \ldots, m, |J| \leq r\right)$$

are the derivatives of $z^{(\ell)}_{i;\mathcal{A}}$ of order at most r. Our goal is to show that, as $n \to \infty$, the joint distribution of the random vectors $D^{\leq r} z^{(\ell)}_{i;\mathcal{A}}$ converges to that of of centered jointly Gaussian vectors that are independent for different i and satisfy

$$\lim_{n\to\infty} \text{Cov}\left(D^{J_1}_{\alpha_1} z^{(\ell)}_{i;\alpha_1},\, D^{J_2}_{\alpha_2} z^{(\ell)}_{i;\alpha_2}\right) = D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} K^{(\ell)}_{\alpha_1\alpha_2},$$

where

$$K^{(\ell+1)}_{\alpha_1\alpha_2} = C_b + C_W \left\langle \sigma(z_\alpha)\sigma(z_\beta)\right\rangle_{K^{(\ell)}}, \quad K^{(1)}_{\alpha_1\alpha_2} = C_b + C_W \sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2}$$

is the infinite width covariance from Theorem 2. To prove this, let us denote by $\mathcal{F}^{(\ell)}$ the sigma algebra generated by the weigts and biases in layer up to and including ℓ. Observe that, conditional on $\mathcal{F}^{(\ell)}$, we have that $D^{\leq r} z^{(\ell)}_{i;\mathcal{A}}$ are already independent for different i and that, since the weights and biases are Gaussian, each is a centered Gaussian with conditional covariance

$$\text{Cov}\left(D^{J_1}_{\alpha_1} z^{(\ell+1)}_{i;\alpha_1},\, D^{J_2}_{\alpha_2} z^{(\ell+1)}_{i;\alpha_2} \;\middle|\; \mathcal{F}^{(\ell)}\right) = D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2},$$

where

$$\Sigma^{(\ell)}_{\alpha_1\alpha_2} = C_b + \frac{C_W}{n_\ell} \sum_{j=1}^{n_\ell} \sigma\left(z^{(\ell)}_{j;\alpha_1}\right) \sigma\left(z^{(\ell)}_{j;\alpha_2}\right).$$

Thus, by the continuous mapping theorem, it suffices to show that for any multi-indices J_1, J_2 with $|J_i| \leq r$ and any $\alpha_1, \alpha_2 \in \mathcal{A}$ we have

$$\lim_{n\to\infty} \mathbb{E}\left[D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}\right] \quad \text{exists and is finite} \tag{65}$$

and

$$\lim_{n\to\infty} \mathrm{Var}\left[D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}\right] = 0. \tag{66}$$

We establish (65) and (66) by induction on ℓ the following more general statement.

Proposition 21 *Denote by $N(n_0, r)$ the number of derivatives of order at most r in n_0 variables. Consider any measureable function $f : \mathbb{R}^{N(n_0,r)\times|\mathcal{A}|} \to \mathbb{R}$ that is polynomially bounded, and define*

$$\mathcal{O}^{(\ell)}_f := \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} f\left(D^{\leq r} z_{j;\mathcal{A}}\right).$$

Then,

$$\lim_{n\to\infty} \mathbb{E}\left[\mathcal{O}^{(\ell)}_f\right] \quad \textit{exists and is finite} \tag{67}$$

and

$$\lim_{n\to\infty} \mathrm{Var}\left[\mathcal{O}^{(\ell)}_f\right] = 0. \tag{68}$$

Proof We proceed by induction, starting with $\ell = 1$. Since weights and biases are Gaussian, the vectors $D^{\leq r} z^{(1)}_{i;\mathcal{A}}$ are independent for all i and jointly Gaussian. The polynomial growth assumption on f show the moments of $f(x)$ are finite if x is Gaussian. This allows us to apply the SLLN to conclude both (67) and (68).

Let us now assume we have proved (65) and (66) for layers $1, \ldots, \ell$. We start by fixing any polynomially bounded f and establishing (65) at layer $\ell + 1$. We have

$$\mathbb{E}\left[\mathcal{O}^{(\ell+1)}_f\right] = \mathbb{E}\left[f\left(D^{\leq r} z_{1;\mathcal{A}}\right)\right].$$

As above, conditionl on $\mathcal{F}^{(\ell)}$, we have the following equality in distribution:

$$D^{\leq r} z_{1;\mathcal{A}} \overset{d}{=} \left(\Sigma^{\leq r,(\ell)}\right)^{1/2} Z, \qquad Z \sim \mathcal{N}\left(0, I_{N(n_0,r)\times\mathcal{A}}\right) \tag{69}$$

where Z is independent of the conditional covariance matrix

$$\Sigma^{\leq r,(\ell)} = \left(D^{J_1}_{\alpha_1} D^{J_2}_{\alpha_2} \Sigma^{(\ell)}_{\alpha_1\alpha_2}\right)_{\alpha_1,\alpha_2\in\mathcal{A}}.$$

The key observation is that each entry of $\Sigma^{\leq r,(\ell)}$ is of the form $\mathcal{O}^{(\ell)}_f$ for polynomially bounded f. Hence, we may apply the inductive hypothesis to conclude that there exists a matrix $\overline{\Sigma}^{\leq r,(\ell)}$ such that the following convergence in distribution holds

$$\Sigma^{\leq r,(\ell)} \quad \xrightarrow{d} \quad \overline{\Sigma}^{\leq r,(\ell)} \qquad \text{as} \quad n \to \infty.$$

The polynomial growth assumption on f together with the Skorohod representation theorem and dominated convergence show that

$$\lim_{n\to\infty} \mathbb{E}\left[f\left(D^{\le r} z_{1;\mathcal{A}}\right)\right] = \mathbb{E}\left[f\left(\left(\overline{\Sigma}^{\le r,(\ell)}\right)^{1/2} Z\right)\right] =: \overline{\mathcal{O}}_f^{(\ell+1)} \quad \text{exists and is finite.}$$

This proves (67). To show (68), we proceed similarly. Namely, we have

$$\mathrm{Var}\left[\mathcal{O}_f^{(\ell+1)}\right] = \frac{1}{n_{\ell+1}}\mathrm{Var}\left[f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right)\right] + \left(1 - \frac{1}{n_{\ell+1}}\right)\mathrm{Cov}\left(f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right),\, f\left(D^{\le r} z_{2;\mathcal{A}}^{(\ell+1)}\right)\right).$$

Note that

$$\mathrm{Var}\left[f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right)\right] \le \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n_\ell}\left[f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right)\right]^2\right].$$

Hence, since f^2 is also polynomially bounded we have already shown that (67) holds at layer $\ell+1$, we see that

$$\mathrm{Var}\left[\mathcal{O}_f^{(\ell+1)}\right] = \mathrm{Cov}\left(f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right),\, f\left(D^{\le r} z_{2;\mathcal{A}}^{(\ell+1)}\right)\right) + O(n^{-1}).$$

Next, using that conditional on $\mathcal{F}^{(\ell)}$ the vectors $D^{\le r} z_{i;\mathcal{A}}^{(\ell+1)}$ are independent for different i we conclude from the law of total covariance that

$$\mathrm{Cov}\left(f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right),\, f\left(D^{\le r} z_{2;\mathcal{A}}^{(\ell+1)}\right)\right) \le \mathrm{Var}\left[\mathbb{E}\left[f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right) \;\middle|\; \mathcal{F}^{(\ell)}\right]\right].$$

Combining the equality in distribution (69) with the polynomial growth condition on f and the dominated convergence theorem we find

$$\lim_{n\to\infty} \mathrm{Var}\left[\mathbb{E}\left[f\left(D^{\le r} z_{1;\mathcal{A}}^{(\ell+1)}\right) \;\middle|\; \mathcal{F}^{(\ell)}\right]\right] = \mathrm{Var}\left[\mathbb{E}\left[f\left(\left(\overline{\Sigma}^{\le r,(\ell)}\right)^{1/2} Z\right)\right]\right] = 0.$$

This completes the proof that (68) holds at infinite width, establishing Proposition 21. ∎

Appendix B. Criticality and Universality in Wide and Deep Networks

In the main body we presented two kinds of results about the structure of random neural networks at large but finite width. The first, Theorem 3, concerned the order of magnitude for cumulants of the output of such a random network and its derivatives. The second, Theorem 4 and Corollary 6, spelled out recursions with respect to the layer index ℓ that describe, to leading order in $1/n$, network cumulants at layer $\ell+1$ in terms of those at layer ℓ. Our purpose in forthcoming sections is to analyze these recursions at large ℓ and to apply this analysis to obtain results about the structure of gradients in deep fully connected networks. Before doing this, we must take a step back and ask: for which σ, C_b, C_W are the recursions (4) describing the infinite width covariance $K^{(\ell)}$ well-behaved at large ℓ?

In section §B.1, we recall a more or less canonical answer to this question whose roots are in the early articles Poole et al. (2016); Raghu et al. (2017) and that was recently spelled out in the generality presented here in Roberts et al. (2022). This procedure, called tuning to criticality, prescribes combinations of σ, C_b, C_W for which $K^{(\ell)}$ is indeed well-behaved at large ℓ. As we shall see below, the term criticality is meant to be evocative of it's use in the analysis of 2d systems in statistical mechanics in that tuning to criticality consists of choosing C_b, C_W so that the infinite width covariance function $K^{(\ell)}$ is as close to constant as a function ℓ as possible.

At a high level, there are two reasons to ask that $K^{(\ell)}$ be slowly varying as a function of ℓ. First, it arguably only makes sense to study perturbative corrections in $1/n$ recursively in ℓ if the limiting $n \to \infty$ covariance structure does not change too rapidly between consecutive layers. Second, and perhaps more importantly, as explained and thoroughly validated in Park et al. (2019); Raghu et al. (2017), deep fully connected networks (without residual connections He et al. (2015), batch normalization Ioffe (2017), etc) are numerically stable enough for gradient-based optimization to succeed only if they are tuned to criticality.

We discuss in §B.2 how considerations underlying criticality naturally give rise to a notion of universality classes for random neural networks. Even the correct definition of universality is still not fully understood. Unlike in random matrix theory, universality for random neural networks depends not on the statistics of the individual weights and biases (though this is also an interesting direction to consider e.g. Hanin et al. (2022)) but rather on the effect of the non-linearity σ on the behavior of the infinite width covariance $K^{(\ell)}$ at large values of the depth ℓ.

Before giving the details, we take this opportunity to emphasize, as we have elsewhere, that the definitions of criticality and universality, the approach to solving the recursions for $\kappa^{(\ell)}_{2k;\alpha}$ from Corollary 6, and the resulting lessons learned about the role of the effective network depth L/n closely follow the ideas developed in the monograph Roberts et al. (2022). Though we pursue them in a somewhat different way, the author would nonetheless like to acknowledge that his co-authors Dan Roberts and Sho Yaida in the book deserve significant credit.

B.1 Tuning to Criticality

As originally explained in Poole et al. (2016); Raghu et al. (2017) and recently spelled out in a definitive way in Roberts et al. (2022), tuning a neural network to criticality means seeking choices of (C_b, C_W) that lead to critical fixed points of the form (K_*, K_*, K_*) for the recursion (4), viewed as a dynamical system describing $(K^{(\ell)}_{\alpha\alpha}, K^{(\ell)}_{\beta\beta}, K^{(\ell)}_{\alpha\beta})$ with time parameter ℓ. Specifically, criticality requires

$$\exists K_* \geq 0 \qquad \text{s.t.} \qquad K_* = C_b + C_W \left\langle \sigma^2(z) \right\rangle_{K_*} \tag{$*$}$$

$$\forall \ell \geq 1 \qquad \left. \frac{\partial K^{(\ell)}_{\alpha\alpha}}{\partial K^{(1)}_{\alpha\alpha}} \right|_{K^{(1)}_{\alpha\alpha} = K_*} = 1 \tag{$||$}$$

$$\forall \ell \geq 1 \qquad \left. \frac{\partial K^{(\ell)}_{\alpha\beta}}{\partial K^{(1)}_{\alpha\beta}} \right|_{K^{(1)}_{\alpha\alpha} = K^{(1)}_{\alpha\alpha} = K^{(1)}_{\alpha\beta} = K_*} = 1, \tag{$\perp$}$$

where

$$K^{(1)}_{\alpha\alpha} = C_b + C_W K^{(0)}_{\alpha\beta}, \qquad K^{(0)}_{\alpha\beta} := \frac{1}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha}x_{j;\beta}, \qquad x_\alpha, x_\beta \in \mathbb{R}^{n_0}.$$

The intuitive meaning of these conditions is as follows. Due to Theorem 2, the first guarantees the existence of a fixed point K_* for of the recursion

$$K^{(\ell+1)}_{\alpha\alpha} = C_b + C_W \left\langle \sigma^2(z) \right\rangle_{K^{(\ell)}_{\alpha\alpha}} \tag{70}$$

of the infinite width variance. In particular, $(*)$ implies

$$K^{(1)}_{\alpha\alpha} = C_b + \frac{C_W}{n_0}\,||x_\alpha||^2 = K_* \qquad \Longrightarrow \qquad K^{(\ell)}_{\alpha\alpha} = \lim_{n\to\infty} \mathrm{Var}\left[z^{(\ell)}_{i;\alpha}\right] = K_* \quad \forall \ell \geq 1.$$

Thus, if a network is tuned to criticality, there is a critical radius

$$K^2_{\text{crit}} := n_0 C_W^{-1}(K_* - C_b)$$

such that for inputs x_α on the sphere of radius K_{crit} the variance of $z^{(\ell)}_{i;\alpha}$ is independent of ℓ in the infinite width limit. In non-critical networks, we expect this variance to either grow or decay exponentially in ℓ, leading to numerical instabilities. The second condition $(||)$ considers the infinite width limit of the variance of $z^{(\ell)}_{i;\alpha}$ for an input x_α for which $K^{(1)}_{\alpha\alpha}$ is close to K_*. Specifically, condition $(||)$ requires for all $\ell \geq 1$ that

$$K^{(1)}_{\alpha\alpha} = \mathrm{Var}[z^{(1)}_{i;\alpha}] = K_* + \delta K \qquad \Longrightarrow \qquad K^{(\ell)}_{\alpha\alpha} = \lim_{n\to\infty} \mathrm{Var}\left[z^{(\ell)}_{i;\alpha}\right] = K_* + \delta K + o(\delta K).$$

This guarantees that the fixed point K_* of the recursion (70) is critical and hence that for inputs near the sphere of radius K_{crit} the variance of the resulting pre-activations $z^{(\ell)}_{i;\alpha}$ is approximately constant in ℓ at large n. The final condition $(\perp)$ considers instead the covariance between two inputs on the sphere of radius K_{crit}. Namely, given two nearby network inputs $x_\alpha, x_\beta \in \mathbb{R}^{n_0}$ with

$$K^{(1)}_{\alpha\alpha} = K^{(1)}_{\beta\beta} = K_*, \qquad K^{(1)}_{\alpha\beta} = C_b + \frac{C_W}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha}x_{j;\beta} = K_* - \delta K,$$

the third condition asks that

$$K^{(\ell)}_{\alpha\beta} = \lim_{n\to\infty} \mathrm{Cov}\left(z^{(\ell)}_{i;\alpha}, z^{(\ell)}_{i;\beta}\right) = K_* - \delta K + o(\delta K), \quad \forall \ell.$$

This ensures that the covariance between pre-activations $z^{(\ell)}_{i;\alpha}$ and $z^{(\ell)}_{i;\beta}$ corresponding to two nearby inputs on the K_{crit}-sphere are approximately independent of ℓ at large n. A simple computation directly from the recursion (4) shows that

$$\chi_{||}(K) := \left.\frac{\partial K^{(\ell+1)}_{\alpha\alpha}}{\partial K^{(\ell)}_{\alpha\alpha}}\right|_{K^{(\ell)}_{\alpha\alpha}=K} = \frac{C_W}{2}\left\langle \partial_z^2(\sigma^2(z))\right\rangle_K \tag{71}$$

$$\chi_{\perp}(K) := \left.\frac{\partial K^{(\ell+1)}_{\alpha\beta}}{\partial K^{(\ell)}_{\alpha\beta}}\right|_{K^{(\ell)}_{\alpha\alpha}=K^{(\ell)}_{\beta\beta}=K^{(\ell)}_{\alpha\beta}=K} = C_W\left\langle (\partial_z\sigma(z))^2\right\rangle_K. \tag{72}$$

Hence, all together, tuning to criticality requires

$$\boxed{K_* \geq 0 \text{ s.t. } K_* = C_b + C_W\left\langle \sigma^2(z)\right\rangle_{K_*} \qquad \text{and} \qquad \chi_{||}(K_*) = \chi_{\perp}(K_*) = 1.} \tag{73}$$

B.2 Universality Classes of Random Neural Networks: Two Examples

We now turn to discussing the notion of universality classes for random neural networks. To start, recall from Theorem 2 that the behavior at large depth ℓ of random fully connected neural networks at infinite width is completely specified by the asymptotics of the limiting covariance function $K^{(\ell)}$. Observe, moreover, that the coefficients in the recursions for $k = 2, 3, 4$ of the cumulants $\kappa^{(\ell+1)}_{2k;\alpha}$ from Corollary 6, which by Theorem 3 determine the behavior of random neural networks at finite width to the first four orders in $1/n$, are completely determined by σ, the infinite width covariance $K^{(\ell)}$, and cumulants $\kappa^{(\ell)}_{2j;\alpha}$, $j \leq k$. It is therefore in terms of the large ℓ behavior of $K^{(\ell)}$ that we should hope to define universality classes of random neural networks at large depth. At present it is not clear what the correct general definition of such a universality class should be. We content ourselves instead with studying two important classes of examples.

B.2.1 The Universality Class of ReLU

The most popular non-linearities used in practice are positively homogeneous of degree 1, i.e. have the form

$$\sigma(t) = (a_- \mathbf{1}_{\{t<0\}} + a_+ \mathbf{1}_{\{t>0\}})t, \qquad a_-, a_+ \in \mathbb{R}, \quad a_- \neq a_+, \quad a_-^2 + a_+^2 \neq 0. \tag{74}$$

Such non-linearities include the ReLU ($a_- = 0, a_+ = 1$) and the leaky ReLU ($a_- \in (0,1), a_+ = 1$). A direct computation, left to the reader, shows that criticality is achieved if and only if

$$K_* \geq 0 \text{ is arbitrary} \quad \text{and} \quad C_b = 0,\, C_W = \frac{2}{a_+^2 + a_-^2}.$$

Thus, the first property of the ReLU universality class is that setting $(C_b, C_W) = (0, 2/(a_+^2 + a_-^2))$ allows all non-negative K_* to satisfy $(*)$. In fact, at criticality, a simple symmetrization argument shows that the variance of neuron pre-activations is preserved exactly *even at finite width*

$$\mathrm{Var}\left[z^{(\ell)}_{i;\alpha}\right] = \mathrm{Var}\left[z^{(1)}_{i;\alpha}\right] = \frac{C_W}{n_0}\,||x_\alpha||^2 \qquad \forall \ell, n_0, \ldots, n_\ell \geq 1,\, x_\alpha \in \mathbb{R}^{n_0} \tag{75}$$

and, relatedly, that we have

$$\chi^{(\ell)}_{||;\alpha} := \chi_{||}(K^{(\ell)}_{\alpha\alpha}) = 1 = \chi_\perp(K^{(\ell)}_{\alpha\alpha}) =: \chi^{(\ell)}_{\perp;\alpha}, \qquad \forall \ell \geq 1,\, x_\alpha \in \mathbb{R}^{n_0}.$$

The remarkable property (75) is much stronger than the criticality condition $(*)$, which requires only that this condition holds for *some* value of $n_0^{-1}\,||x_\alpha||^2$ and only in the limit when $n \to \infty$. It implies that the cumulant recursions from Corollary 6 for $1-$homogeneous non-linearities have constant coefficients and are therefore particularly simple to solve. For instance, we find at leading order in $1/n$

$$\begin{aligned}\kappa^{(\ell+1)}_{4;\alpha} &= \frac{C_W^2}{n_\ell}\left[\left\langle \sigma(z)^4 \right\rangle_{K^{(\ell)}_{\alpha\alpha}} - \left\langle \sigma(z)^2 \right\rangle^2_{K^{(\ell)}_{\alpha\alpha}}\right] + \left(\chi^{(\ell)}_{||;\alpha}\right)^2 \kappa^{(\ell)}_{4;\alpha} \\ &= \left(\frac{2}{(a_+^2 + a_-^2)n_0}\,||x_\alpha||^2\right)^2 \left(6\frac{a_+^4 + a_-^4}{(a_+^2 + a_-^2)^2} - 1\right) \sum_{\ell'=1}^{\ell} \frac{1}{n_{\ell'}},\end{aligned}$$

which shows that while $\kappa_{4,\alpha}^{(\ell)}$ is suppressed by one power of $1/n$ relative to the infinite width variance $K_{\alpha\alpha}^{(\ell)}$, it also grows one order faster in ℓ. This illustrates an important and general theme: depth amplifies finite width effects. It is the effective depth ℓ/n of neurons at layer ℓ that measures the distance to the infinite width Gaussian regime.

Moreover, in the special setting of 1-homogeneous non-linearities there is a simple method for obtaining the full distribution of the pre-activation vector $z_\alpha^{(\ell)}$ at a single input at any finite values of $n_0, \ldots, n_\ell$. This was first pointed out in Hanin (2018); Hanin and Nica (2020b); Zavatone-Veth and Pehlevan (2021) and is briefly reviewed in Appendix D. A key takeaway is that if we take the hidden layer widths $n_1 = \cdots = n_L = n$, then we have following convergence in distribution to product of independent normal and log-normal random variables:

$$\lim_{\substack{n,L\to\infty \\ L/n\to\xi\in[0,\infty)}} z_{i;\alpha}^{(L)} \quad \stackrel{d}{=} \quad \left(\frac{2\,||x_\alpha||^2}{(a_+^2+a_-^2)n_0}\right)^{1/2} Z_1 \exp\left[-\mu(\xi,a_+,a_-)+\sigma(\xi,a_+,a_-)Z_2\right], \quad (76)$$

where

$$\mu(\xi,a_+,a_-)=\sigma^2(\xi,a_+,a_-) := \frac{\xi}{4}\left(6\frac{a_+^4+a_-^4}{(a_+^2+a_+^2)^2}-1\right), \quad Z_1, Z_2 \sim \mathcal{N}(0,1) \text{ iid.}$$

The convergence (76) reveals that for a fixed input the distribution of the output of a random with $1-$homogeneous non-linearities at large depth and width depends in a simple way on the limiting effective network depth ξ. This bolsters the claim that they are all part of the same universality class. It also means that increasing the network depth L drives it away from the infinite width Gaussian behavior observed at $\xi = 0$ and that the outputs of *deep and wide* networks are not well-approximated by a Gaussian at all, unless ξ is infinitesimal, in which case the log-normal term $\exp\left[-\mu(\xi,a_+,a_-)+\sigma(\xi,a_+,a_-)Z_2\right]$ is negligible.

Prior work Hanin and Nica (2020a,b); Hanin and Rolnick (2018) of the author shows that when $\sigma = \mathrm{ReLU}$ (or any other 1-homogeneous non-linearity), the distribution at large n, L of not only the network output $z_{i;\alpha}^{(L+1)}$ but also is derivatives with respect to inputs x_α and model parameters (e.g. weights and biases) depends only on the effective depth L/n. We further note that it has also been observed that log-normal random variables describe the structure of gradients in residual networks, even during/after training Li et al. (2021).

To complete our discussion of the ReLU universality class, we make two final remarks. First, a direct computation (reviewed briefly in Proposition 26 of Appendix D) shows that at criticality for any non-zero inputs $x_{\alpha_1}, x_{\alpha_2} \in \mathbb{R}^{n_0}$ with the same norm we have

$$\lim_{n\to\infty} \mathrm{Corr}\left(z_{i;\alpha_1}^{(\ell)}, z_{i;\alpha_2}^{(\ell)}\right) = 1 - \frac{2(a_+-a_-)^2}{3\pi(a_+^2+a_-^2)}\ell^{-2}(1+o(1)). \quad (77)$$

The power law exponent 2 that appears in this estimate is common to all $1-$homogeneous non-linearities and is another reason to believe they fall into the same universality class. In contrast, this exponent equals one for non-linearities in the $K_* = 0$ universality class presented below. The estimate (77) suggests that in order to define a double scaling limit $n, L \to \infty$ and $L/n \to \xi$ in which the entire field $x_\alpha \mapsto z_\alpha^{(L+1)}$ is non-degenerate (rather than

just its value at a single input) one must rescale distances in the input space to prevent the collapse of correlations otherwise guaranteed in (77). We leave this as an interesting direction for future work.

B.2.2 The Universality Class of Hyperbolic Tangent

The second class of non-linearities we study is what Roberts et al. (2022) termed the $K_* = 0$ universality class, which we take to mean non-linearities σ such that

- σ is a smooth, odd function satisfying Assumption 1.
- σ satisfies
$$\sigma_1\sigma_3 < 0, \qquad \sigma_j := \frac{1}{j!}\frac{d^j}{dt^j}\bigg|_{t=0}\sigma(t). \tag{78}$$
- $K_* = 0$ is the unique fixed point of equation $(*)$.
- At criticality, for every non-zero network input $x_\alpha \in \mathbb{R}^{n_0}$ and each $\delta \in (0,1)$ we have as $L \to \infty$ that
$$K^{(L)}_{\alpha\alpha} = \frac{1}{aL}\left(1 + O(L^{-1+\delta})\right), \tag{79}$$
where the implicit constant depends on δ and x_α and we've set
$$a := -6\frac{\sigma_3}{\sigma_1}.$$
This specific value of a, which is positive by (78), is the only possible candidate for decay of the form (79) that is consistent with the recursion (4).

Some remarks are in order. First, if $K_* = 0$ is the unique fixed point for $(*)$, then a simple computation shows that criticality is achieved if and only if

$$K_* = 0, \qquad C_b = 0, \qquad C_W = \sigma_1^{-2}. \tag{80}$$

Next, our definition of the $K_* = 0$ universality class does not make apparent whether it is empty. As we will see in Proposition 22, however, the $K_* = 0$ universality class is in fact quite large and contains for example the hyperbolic tangent and more generally any non-linearity that is tanh-like in the sense that is smooth with $\sigma_1 \neq 0$, has the opposite sign as its second derivative

$$\text{for almost every } z,\ \operatorname{sgn}\left(\sigma(z)\sigma''(z)\right) = -1,$$

is sub-linear

$$\exists C > 0 \text{ s.t. } \forall z \in \mathbb{R}\ |\sigma(z)| \leq |\sigma_1 z|,$$

and is controlled by its first few non-zero Taylor series coefficients at 0:

$$\exists C \geq 0 \text{ s.t. } \forall z \geq 0, \quad \sigma_1 z + \sigma_3 z^3 \leq \sigma(z) \leq \sigma_1 z + \sigma_3 z^3 + Cz^4.$$

Further, by definition, for the $K_* = 0$ universality class, the infinite width variance $K^{(\ell)}_{\alpha\alpha}$ of neuron pre-activations $z^{(\ell)}_{i;\alpha}$ is qualitatively different from that of $1-$homogeneous non-linearities. Indeed, $K^{(L)}_{\alpha\alpha}$ depends on L, decaying polynomially to 0. Moreover, at large L,

the value of $K^{(L)}_{\alpha\alpha}$ is independent of the initial condition $K^{(0)}_{\alpha\alpha}$ to leading order in L. As a final remark let us point out that searching for non-linearities σ so that $K_* = 0$ at criticality is quite natural. Indeed, for any σ that is twice differentiable, we have

$$\chi_{||}(K) = \chi_{\perp}(K) + C_W \left\langle \sigma(z)\sigma''(z) \right\rangle_K$$

Hence, if $K > 0$, then

$$\chi_{||}(K) = 1,\ \chi_{\perp}(K) = 1 \qquad \Longrightarrow \qquad \left\langle \sigma(z)\sigma''(z) \right\rangle_K = 0.$$

But if σ is a sigmoidal function such as tanh, then $\sigma(z)\sigma''(z) < 0$ for all $z \neq 0$. Hence, $\langle \sigma(z)\sigma''(z) \rangle_K = 0$ can only occur when $K = 0$.

As in the monograph Roberts et al. (2022), let us now probe the role of network depth by studying the large L behavior of the cumulants $\kappa^{(L)}_{2k;\alpha}$, $k = 2, 3, 4$, in networks with non-linearities from the $K_* = 0$ universality class tuned to criticality. Note that in (79) the limiting behavior of the variance $K^{(L)}_{\alpha\alpha}$ depends (mildly) on the non-linearity σ in terms of is first few Taylor coefficients at 0. As we are about to see, however, the behavior of the higher cumulants $\kappa^{(L)}_{2k;\alpha}$, $k = 2, 3, 4$, when normalized by the appropriate power of $K^{(L)}_{\alpha\alpha}$, is independent of σ at leading order in n and L and depends only on universal constants and the effective network depth L/n.

Appendix C. Infinite Width Analysis of Tanh-like Non-linearities

The purpose of this section is to derive some basic properties of the infinite width variance recursion

$$\kappa^{(\ell+1)}_{\alpha\alpha} = C_b + C_W \left\langle \sigma(z)^2 \right\rangle_{\kappa^{(\ell)}_{\alpha\alpha}}. \tag{81}$$

We abbreviate

$$\sigma_j := \frac{1}{j!}\frac{d^j}{dx^j}\bigg|_{x=0} \sigma(x)$$

and consider here the case when σ that is a tanh-like non-linearity in the sense that σ satisfies:

- σ is smooth at 0 with $\sigma_1 \neq 0$
- σ has the opposite sign as its second derivative

$$\text{for almost every } z,\ \operatorname{sgn}\left(\sigma(z)\sigma''(z)\right) = -1. \tag{82}$$

  Note that this forces $\sigma_2 = 0$ and

$$a := -\frac{6\sigma_3}{\sigma_1} > 0.$$

- σ is sub-linear:

$$\exists C > 0 \text{ s.t. } \forall z \in \mathbb{R}\ |\sigma(z)| \leq |\sigma_1 z|, \tag{83}$$

- σ is controlled by its first few non-zero Taylor series coefficients at 0:

$$\exists C \geq 0 \text{ s.t. } \forall z \geq 0, \quad \sigma_1 z + \sigma_3 z^3 \leq \sigma(z) \leq \sigma_1 z + \sigma_3 z^3 + C z^4 \tag{84}$$

We will be interested in understanding the recursion (81) at criticality in the sense defined in §B.1. Specifically, we remind the reader that this means we choose C_b, C_W so that

$$\begin{aligned} \exists K_* \geq 0 \quad \text{s.t.} \quad K_* &= C_b + C_W \left\langle \sigma^2(z) \right\rangle_{K_*} \\ \chi_{||}(K_*) &= \frac{C_W}{2} \left\langle \partial^2(\sigma(z)^2) \right\rangle_{K_*} = 1 \\ \chi_{\perp}(K_*) &= C_W \left\langle (\sigma'(z))^2 \right\rangle_{K_*} = 1. \end{aligned}$$

Before stating our main result (Proposition 22), let us explain intuitively what we expect. First of all, as we shall see in Proposition 22, tanh-like non-linearities requires $K_* = 0$ for criticality. Second, by Taylor expanding the recursion (4) around small values of $K^{(\ell)}_{\alpha\alpha}$ we find

$$K^{(\ell+1)}_{\alpha\alpha} = K^{(\ell)}_{\alpha\alpha} - a \left(K^{(\ell)}_{\alpha\alpha} \right)^2 + O\left(\left(K^{(\ell)}_{\alpha\alpha} \right)^3 \right).$$

This is well-approximated by the ODE

$$\frac{d}{dt} K(t) = -a K(t)^2,$$

whose solution is

$$K(t) = \left(at + \frac{1}{K(0)} \right)^{-1}.$$

This form for the solution has two important properties that we will check in Proposition 22 hold for the actual solution $K^{(\ell)}_{\alpha\alpha}$ to the discrete difference equation (4):

- At large t, $K(t)$ tends to zero like $1/at$ plus an error of size roughly $O(t^{-2})$.
- The leading order behavior of $K(t)$ at large t is independent of the initial condition.

Proposition 22 *If σ is a tanh-like non-linearity in the sense defined above then criticality is achieved for σ only with*

$$K_* = 0, \qquad C_b = 0, \qquad \textit{and} \qquad C_W = \sigma_1^{-2}. \tag{85}$$

Moreover, for every $\delta \in (0,1)$ we have

$$K^{(0)}_{\alpha\alpha} > 0 \quad \Rightarrow \quad \sup_{\ell \geq 1} \ell^{2-\delta} \left| K^{(\ell)}_{\alpha\alpha} - \frac{1}{a\ell} \right| < \infty. \tag{86}$$

Proof The proof relies on the following estimate

Lemma 23 *Fix $C_1, C_2, \psi > 0$ satisfying*

$$C_2 \geq 1, \quad \psi \neq C_2 + 1$$

as well as $* \in \{\leq, \geq\}$. Suppose also that for each $\ell \geq 0$ we have

$$a_{\ell+1} \ * \ \xi_\ell + (1 - \zeta_\ell) a_\ell, \qquad \zeta_\ell \in [0,1] \tag{87}$$

with $a_0 \in \mathbb{R}$ given and that there exist $C_1', C_2' > 0$ so that

$$\left|\xi_\ell - C_1 \ell^{-\psi}\right| \le C_1' \ell^{-1-\psi}, \qquad \left|\zeta_\ell - C_2 \ell^{-1}\right| \le C_2' \ell^{-2}.$$

Then

$$a_{\ell+1} * \frac{\ell^{1-\psi}}{1-\psi+C_2}\left(1+O(\ell^{-1})\right) + e^{-C_2\gamma}\ell^{-C_2} a_0 \left(1+O(\ell^{-1})\right) \tag{88}$$

where γ is the Euler-Mascheroni constant and the implied constants depend only C_1, C_2, C_1', C_2'.

Proof By unfolding the recursion (87) we find

$$a_{\ell+1} \; * \; \sum_{\ell'=1}^{\ell} \xi_{\ell'} \prod_{\ell''=\ell'+1}^{\ell} (1-\zeta_{\ell''}) + a_0 \prod_{\ell''=0}^{\ell} (1-\zeta_{\ell''}).$$

We have

$$\begin{aligned}
\prod_{\ell''=1}^{\ell} (1-\zeta_{\ell''}) &= \exp\left[\sum_{\ell''=1}^{\ell} \log\left(1 - C_2(\ell'')^{-1} + O(\ell^{-2})\right)\right] \\
&= \exp\left[O(\ell^{-1}) + \sum_{\ell''=1}^{\ell} -C_2(\ell'')^{-1}\right] \\
&= \exp\left[O(\ell^{-1}) - C_2\log(\ell) - C_2\gamma\right] \\
&= e^{-C_2\gamma}\ell^{-C_2}\left(1+O(\ell^{-1})\right).
\end{aligned}$$

This gives the second term in (88). For the first term, we write

$$\begin{aligned}
\sum_{\ell'=1}^{\ell} \xi_{\ell'} \prod_{\ell''=\ell'+1}^{\ell} (1-\zeta_\ell) &= \sum_{\ell'=1}^{\ell} \xi_{\ell'} \exp\left[\sum_{\ell''=\ell'+1}^{\ell} \log\left(1-\zeta_\ell\right)\right] \\
&= \sum_{\ell'=1}^{\ell} \xi_{\ell'} \exp\left[\sum_{\ell''=\ell'+1}^{\ell} -C_2(\ell'')^{-1} + O((\ell'')^{-2})\right] \\
&= \sum_{\ell'=1}^{\ell} C_1(\ell')^{-\psi}(1+O(\ell')^{-1}) \exp\left[-C_2\log\left(\frac{\ell}{\ell'}\right) + O((\ell')^{-1})\right] \\
&= \ell^{-C_2} \sum_{\ell'=1}^{\ell} C_1(\ell')^{-\psi+C_2}(1+O(\ell')^{-1}) \\
&= \frac{C_1}{1+C_2-\psi}\ell^{1-\psi}\left(1+O(\ell^{-1})\right).
\end{aligned}$$

This completes the proof of (88). ∎

Note that for any $K \ge 0$ we have

$$\chi_{||}(K) = \chi_\perp(K) + C_W \left\langle \sigma(z)\sigma''(z)\right\rangle_K. \tag{89}$$

Hence, at criticality, we must have

$$\left\langle \sigma(z)\sigma''(z)\right\rangle_{K_*} = 0.$$

But due to assumption (82) we have

$$K > 0 \quad \Longrightarrow \quad \left\langle \sigma(z)\sigma''(z)\right\rangle_{K} < 0.$$

Thus, we indeed find that we must have $K_* = 0$ at criticality. Hence, in light of (89) criticality is equivalent to the system of equations

$$K_* = 0 = C_b + C_W\sigma(0)^2, \qquad \chi_{||}(0) = \chi_{\perp}(0) = C_W\left\langle (\sigma'(z))^2\right\rangle_0 = C_W\sigma_1^2 = 1.$$

This system has a unique solution:

$$C_b = 0, \quad C_W = \sigma_1^{-2},$$

completing the proof of the criticality conditions (85). Let us now establish (86). First note that at criticality the sub-linearity condition (83) guarantees that for all $\delta > 0$ there exists $c_\delta \in (0,1)$ such that

$$K > \delta \quad \Longrightarrow \quad C_W\left\langle \sigma(z)^2\right\rangle_K < (1-c_\delta)\left\langle z^2\right\rangle_K = (1-c_\delta)K.$$

Hence, for all $K, \delta > 0$ there exists $\ell_0 \geq 1$ such that

$$K^{(0)}_{\alpha\alpha} \leq K \qquad \Longrightarrow \qquad K^{(\ell)}_{\alpha\alpha} \leq \delta \quad \forall \ell \geq \ell_0. \tag{90}$$

In particular, $K^{(\ell)}_{\alpha\alpha}$ is monotonically decreasing and converges to $K_* = 0$ as ℓ grows. Let us now define for each $\ell \geq 1$

$$K^{(\ell)}_{\alpha\alpha} =: \frac{1}{a\ell} + \epsilon^{(\ell)}, \qquad a := -6\frac{\sigma_3}{\sigma_1} > 0,$$

where a is positive due to (82). Note that since $K^{(\ell)}_{\alpha\alpha}$ tends to zero with ℓ, so does $\epsilon^{(\ell)}$. Let us agree that for any $t \in \mathbb{R}$ the symbol t^+ (resp. t^-) means that for ℓ sufficiently large we may make the constant t^+ (resp. t^-) arbitrary close to t from above (resp. below). In order to prove (86), we start with the following elementary estimate.

Lemma 24 *For all $\ell \geq 1$, we have*

$$\epsilon^{(\ell+1)} \;\geq\; -\frac{1}{a\ell^2(\ell+1)} + \epsilon^{(\ell)}\left(1 - \frac{2}{\ell} - a\epsilon^{(\ell)}\right). \tag{91}$$

Further, there exists a constant $C > 0$ depending only on σ with the following property. For all $K > 0$ there exists a constant $\ell_0 \geq 1$ so that if $K^{(0)}_{\alpha\alpha} \leq K$, then for all $\delta \in (0,1)$ we have

$$\epsilon^{(\ell+1)} \;\leq\; \frac{C}{\ell^3} + \epsilon^{(\ell)}\left(1 - \frac{2-\delta}{\ell}\right), \qquad \forall\ \ell \geq \ell_\delta := \max\left\{\frac{C}{\delta}, \frac{2C}{a}, \ell_0\right\}. \tag{92}$$

Proof Plugging in the estimates (84) into the recursion (81) yields for some $C > 0$ depending only on σ

$$\epsilon^{(\ell+1)} \leq \frac{C}{\ell^3} + \epsilon^{(\ell)} \left[1 - \frac{2}{\ell} + \frac{C}{\ell^2}\right] + \left(\epsilon^{(\ell)}\right)^2 \left(-a + \frac{C}{\ell}\right) + C(\epsilon^{(\ell)})^3$$

Note that for all $\ell \geq 2C/a$ we have $-a + C/\ell \leq 0$. Hence, for all $\ell \geq \max\{2C/a, C/\delta\}$ we have

$$\epsilon^{(\ell+1)} \leq \frac{C}{\ell^3} + \epsilon^{(\ell)} \left[1 - \frac{2-\delta}{\ell}\right] + C(\epsilon^{(\ell)})^3.$$

Moreover, if $\epsilon^{(\ell)} \leq 0$, then $(\epsilon^{(\ell)})^3 \leq 0$. If on the other hand $\epsilon^{(\ell)} \geq 0$, then from (90) we find that there for all $K > 0$ there exists ℓ_0 so that $(\epsilon^{(\ell)})^3 \leq a(\epsilon^{(\ell)})^2/4$ for all $\ell \geq \ell_0$. Hence, in all cases, for each $\delta \in (0,1)$ if $\ell \geq \max\{2C/a, C/\delta, \ell_0\}$, we find

$$\epsilon^{(\ell+1)} \leq \frac{C}{\ell^3} + \epsilon^{(\ell)} \left[1 - \frac{2-\delta}{\ell}\right],$$

as claimed. The lower bound follows from a similar but simpler computation. ■

Fix $\delta \in (0,1)$. The relation (92), together with Lemma 23, show that for all $K > 0$ there exists some $C' > 0$ depending on δ, σ, K such that if $K^{(0)}_{\alpha\alpha} \leq K$ then

$$\epsilon^{(\ell+1)} \leq \sum_{\ell'=\ell_\delta}^{\ell} \frac{C}{\ell^3} \prod_{\ell''=\ell'+1}^{\ell} \left(1 - \frac{2^-}{\ell}\right) \leq C' \left[\frac{1}{\ell^2} + \epsilon^{(\ell_\delta)} \frac{1}{\ell^{2-\delta}}\right].$$

This shows that

$$\forall \delta \in (0,1)\ \exists \ell_\delta \geq 1 \ \text{ s.t.} \quad \epsilon^{(\ell)} \leq \frac{1}{\ell^{2-\delta}} \quad \forall \ell \geq \ell_\delta. \tag{93}$$

To conclude (86) it therefore remains to deduce that

$$\forall K_1, K_2 > 0,\ \delta \in (0,1)\ \exists \ell_\delta \geq 1 \ \text{ s.t.} \quad K_1 < K^{(0)}_{\alpha\alpha} < K_2 \quad \Longrightarrow \quad \epsilon^{(\ell)} \geq -\frac{1}{\ell^{2-\delta}} \quad \forall \ell \geq \ell_\delta. \tag{94}$$

To aid with this, we will need the following

Lemma 25 *For any $\delta \in (0,1)$ there exists $\ell_\delta \geq 1$ with the property that if $\ell \geq \ell_\delta$ then*

$$\epsilon^{(\ell)} \geq -\ell^{-2+\delta} \quad \Longrightarrow \quad \epsilon^{(\ell+1)} \geq -(\ell+1)^{-2+\delta}.$$

Proof Suppose $\epsilon^{(\ell)} \geq -\ell^{-2+\delta}$. The lower bound in (91) yields for some $C, C' > 0$

$$\begin{aligned} \epsilon^{(\ell+1)} + (\ell+1)^{-2+\delta} &\geq (\ell+1)^{-2+\delta} \left[1 - \left(1 + \ell^{-1}\right)^{2-\delta}\right] - 2\ell^{-3+\delta} - C\ell^{-4+2\delta} \\ &\geq \delta \ell^{-3+\delta} - C'(\ell^{-3} + \ell^{-4+2\delta}), \end{aligned}$$

which is non-negative for all ℓ sufficiently large. ■

We are now in a position to establish (94). In light of the previous Lemma we need only consider the case when

$$\forall \delta \in (0,1) \; \exists \; \ell_\delta \geq 1 \quad \text{s.t.} \quad \epsilon^{(\ell_\delta)} < -\ell^{2-\delta}.$$

Note that in light of the upper bound (93) we find that for all $\delta \in (0,1)$ there exists $\ell_\delta \geq 1$ and $C_\delta > 0$ so that for all $\ell \geq \ell_\delta$ we have

$$K_{\alpha\alpha}^{(\ell+1)} \geq K_{\alpha\alpha}^{(\ell)} \left(1 - aK_{\alpha\alpha}^{(\ell)}\right) \geq K_{\alpha\alpha}^{(\ell)} \left(1 - a\left(-\frac{1}{a\ell} + C_\delta \ell^{-2+\delta}\right)\right) = K_{\alpha\alpha}^{(\ell)} \left(1 - \frac{1}{\ell} - aC_\delta \ell^{-2+\delta}\right).$$

Hence, assuming $K_2 \geq K_{\alpha\alpha}^{(0)} \geq K_1 > 0$, we may iterate this inequality to find that there exists $c > 0$ depending on K_1, K_2 and $\ell_0 \geq 1$ so that

$$K_{\alpha\alpha}^{(\ell)} \geq \frac{c}{a\ell} \qquad \forall \ell \geq \ell_0.$$

Hence, since $\epsilon^{(\ell)} < 0$ for all $\ell \geq \ell_\delta$ we find for all $\ell \geq \max\{\ell_0, \ell_\delta\}$ that

$$-a(\epsilon^{(\ell)})^2 \geq \epsilon^{(\ell)} \frac{1-c}{\ell}$$

Substituting this into the lower bound (91), we find that for all $\ell \geq \max\{\ell_0, \ell_\delta\}$

$$\epsilon^{(\ell+1)} \geq -\frac{C'}{\ell^3} + \epsilon^{(\ell)} \left(1 - \frac{1+c}{\ell}\right).$$

Since $\epsilon^{(\ell_\delta)} < 0$, we see by applying Lemma 23 that there exists $C > 0$ so that for all $\ell \geq \max\{\ell_0, \ell_\delta\}$

$$\epsilon^{(\ell+1)} \geq -\frac{C}{\ell^{1+c}}.$$

But now we can bootstrap this estimate. Indeed, for any $\delta \in (0,1)$ we substitute this into the lower bound (91) to find that for all ℓ sufficiently large

$$\epsilon^{(\ell+1)} \geq -\frac{C'}{\ell^3} + \epsilon^{(\ell)} \left(1 - \frac{2-\delta}{\ell}\right).$$

Again applying Lemma 23 yields that for all ℓ sufficiently large

$$\epsilon^{(\ell+1)} \geq -\frac{C}{\ell^{2-\delta}}.$$

This completes the proof.

■

Appendix D. Exact Solutions for 1-homogeneous activations

In this appendix, we collect several known computations related to the distribution of neuron activations in random fully connected networks with $1-$homogeneous activations. Specifically, we fix a one homogeneous non-linearity

$$\sigma(t) = (a_+ \mathbf{1}_{t>0} + a_- \mathbf{1}_{t<0})t$$

and consider a random fully connected neural network with input dimension n_0, output dimension n_{L+1}, hidden layer widths $n_1, \ldots, n_\ell$, and non-linearity σ that is tuned to criticality in the sense that

$$C_b = 0, \qquad C_W = \frac{2}{a_+^2 + a_-^2}.$$

Our first task is to derive in §D.1 a known exact formula for the infinite width covariance $K_{\alpha\beta}^{(\ell+1)}$ as a function of $K_{\alpha\alpha}^{(\ell)}, K_{\alpha\beta}^{(\ell)}, K_{\beta\beta}^{(\ell)}$. Then, in Section §D.2, we sketch a derivation of the limiting distribution (76) of a neuron pre-activation in the double scaling limit $n, L \to \infty, L/n \to \gamma$.

D.1 Covariance Propagation in Random Fully Connected 1-homogeneous Networks

In this section, we consider two network inputs x_α, x_β of the same norm:

$$K_{\alpha\alpha}^{(0)} = \frac{1}{n_0}\,||x_\alpha||^2 = K = \frac{1}{n_0}\,||x_\beta||^2 = K_{\beta\beta}^{(0)}, \qquad K > 0. \tag{95}$$

Let us define

$$\epsilon_{\alpha\beta}^{(\ell)} := \frac{1 - \mathrm{Corr}_{\alpha\beta}^{(\ell)}}{2}, \qquad \mathrm{Corr}_{\alpha\beta}^{(\ell)} := \frac{K_{\alpha\beta}^{(\ell)}}{\left(K_{\alpha\alpha}^{(\ell)} K_{\beta\beta}^{(\ell)}\right)^{1/2}}$$

Our goal is to derive the following explicit recursion for $\epsilon_{\alpha\beta}^{(\ell+1)}$ in terms of $\epsilon_{\alpha\beta}^{(\ell)}$. This derivation follows the approach in §5.5 Roberts et al. (2022). To the author's knowledge, the following formula (or really something equivalent) was first derived in Cho and Saul (2009).

Proposition 26 (Correlation propagation for $1-$homogeneous activation functions) *At criticality, we have the following exact formula:*

$$\begin{aligned} 1 - 2\epsilon_{\alpha\beta}^{(\ell+1)} = \frac{2C_W(a_+ - a_-)^2}{\pi}\left[\frac{1}{2}\sqrt{\epsilon_{\alpha\beta}^{(\ell)}(1-\epsilon_{\alpha\beta}^{(\ell)})} + \left(\frac{1}{2} - \epsilon_{\alpha\beta}^{(\ell)}\right)\cos^{-1}\left(\sqrt{\epsilon_{\alpha\beta}^{(\ell)}}\right)\right] \\ + C_W a_+ a_- (1 - 2\epsilon_{\alpha\beta}^{(\ell)}) \end{aligned} \tag{96}$$

In particular, taking $\varepsilon_{\alpha\beta}^{(\ell)}$ small we find

$$\varepsilon_{\alpha\beta}^{(\ell+1)} = \varepsilon_{\alpha\beta}^{(\ell)} - \frac{4}{3\pi}\left(\varepsilon_{\alpha\beta}^{(\ell)}\right)^{3/2} + O\left(\left(\varepsilon_{\alpha\beta}^{(\ell)}\right)^{5/2}\right).$$

Hence, as $\ell \to \infty$,

$$\varepsilon_{\alpha\beta}^{(\ell)} = \frac{2}{3\pi}\ell^{-2}(1 + o(1)).$$

Proof We have from Theorem 2 that

$$K_{\alpha\beta}^{(\ell+1)} = C_b + C_W \left\langle \sigma(z_\alpha)\sigma(z_\beta) \right\rangle_{K^{(\ell)}}, \tag{97}$$

where we recall that the brackets above mean the average with respect to the Gaussian distribution

$$\begin{pmatrix} z_\alpha \\ z_\beta \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K_{\alpha\alpha}^{(\ell)} & K_{\alpha\beta}^{(\ell)} \\ K_{\alpha\beta}^{(\ell)} & K_{\beta\beta}^{(\ell)} \end{pmatrix}\right).$$

Since we are at criticality, we have

$$C_b = 0, \qquad C_W = \frac{2}{a_+^2 + a_-^2}$$

and that moreover

$$K_{\alpha\alpha}^{(\ell)} = K_{\beta\beta}^{(\ell)} = K,$$

where K is the constant from (95). Our first step is to change from the Gaussian variables z_α, z_β to the new Gaussian variables

$$\xi = \frac{z_\alpha + z_\beta}{2\sqrt{K}}, \qquad \eta = \frac{z_\alpha - z_\beta}{2\sqrt{K}}.$$

We have

$$z_\alpha = \sqrt{K}(\xi + \eta), \qquad z_\beta = \sqrt{K}\,(\xi - \eta)\,.$$

Moreover, writing

$$\epsilon := \epsilon_{\alpha\beta}^{(\ell)} = \frac{1}{2}\left(1 - \frac{K_{\alpha\beta}^{(\ell)}}{\left(K_{\alpha\alpha}^{(\ell)} K_{\beta\beta}^{(\ell)}\right)^{1/2}}\right)$$

we find

$$\mathrm{Var}[\xi] = 1 - \varepsilon, \qquad \mathrm{Var}[\eta] = \varepsilon, \qquad \mathrm{Cov}[\xi, \eta] = 0.$$

Hence, the right hand side of the recursion (97) reads

$$C_W K \int_{\mathbb{R}}\int_{\mathbb{R}} \sigma\left((1-\varepsilon)^{1/2}\xi + \varepsilon^{1/2}\eta\right)\sigma\left((1-\varepsilon)^{1/2}\xi - \varepsilon^{1/2}\eta\right)\exp\left[-\frac{1}{2}\left(\xi^2+\eta^2\right)\right]\frac{d\xi d\eta}{2\pi}.$$

Using the definition of σ yields

$$\begin{aligned}&\sigma\left((1-\varepsilon)^{1/2}\xi + \varepsilon^{1/2}\eta\right)\sigma\left((1-\varepsilon)^{1/2}\xi - \varepsilon^{1/2}\eta\right)\\ &\quad = (a_+\mathbf{1}_{\xi+\eta>0} + a_-\mathbf{1}_{\xi+\eta<0})\,(a_+\mathbf{1}_{\xi-\eta>0} + a_-\mathbf{1}_{\xi-\eta<0})\,((1-\epsilon)\xi^2 - \epsilon\eta^2).\end{aligned}$$

Changing variables $(\xi, \eta) \to (-\xi, -\eta)$ inside the integral and averaging yields

$$\begin{aligned}K_{\alpha\beta}^{(\ell+1)} &= C_W K a_+ a_- (1-2\epsilon)\\ &\quad + \frac{C_W K (a_+ - a_-)^2}{2}\int_{\mathbb{R}^2} \mathbf{1}_{(1-\epsilon)\xi^2 - \epsilon\eta^2 > 0}((1-\epsilon)\xi^2 - \epsilon\eta^2)\exp\left[-\frac{1}{2}\left(\xi^2+\eta^2\right)\right]\frac{d\xi d\eta}{2\pi}.\end{aligned}$$

Passing to polar coordinates and explicitly computing the resulting integral is now straightforward and completes the derivation of (96). ■

D.2 Full Distribution of Neuron Pre-activations at a Single Input and the Derivation of (76)

Our purpose in this section is to briefly recall an exact formula for the full distribution of a neuron pre-activation $z_{i;\alpha}^{(L+1)}$. For this, note that since $x_\alpha \mapsto z_\alpha^{(L+1)}$ is piecewise linear and the event that the Jacobian $J_{x_\alpha} z_\alpha^{(L+1)}$ is not well-defined at x_α has probability zero, we may write

$$z_\alpha^{(L+1)} = J_{x_\alpha} z_\alpha^{(L+1)} x_\alpha.$$

Next,

$$J_{x_\alpha} z_\alpha^{(L+1)} = W^{(L+1)} D^{(L)} W^{(L)} \cdots D^{(1)} W^{(1)}, \tag{98}$$

where $W^{(\ell)}$ are simply the weight matrices and

$$D^{(\ell)} := \text{Diag}\left(\sigma'(z_{i;\alpha}^{(\ell)}),\, i = 1, \ldots, n_\ell\right).$$

Arguing exactly as in Proposition 2 of Hanin and Nica (2020b), we have the following equality in distribution:

$$D^{(L)} W^{(L)} \cdots D^{(1)} W^{(1)} \stackrel{d}{=} A \widehat{D}^{(L)} W^{(L)} \cdots \widehat{D}^{(1)} W^{(1)},$$

where A is a diagonal matrix with iid ± 1 entries on the diagonal that is independent of $W^{(\ell)}$ and the diagonal matrices

$$\widehat{D}^{(\ell)} = \text{Diag}\Big(\underbrace{a_+ \xi_i^{(\ell)} + a_-(1 - \xi_i^{(\ell)})}_{=: d_i^{(\ell)}},\, i = 1, \ldots, n_\ell \Big), \qquad \xi_i^{(\ell)} \sim \text{Bernoulli}(1/2) \; iid.$$

Combining this with (98) and recalling that the entries of $W^{(L+1)}$ are iid centered Gaussians with variance C_W / n_L yields

$$z_{i;\alpha}^{(L+1)} \stackrel{d}{=} Z_1 \cdot \left(\frac{C_W}{n_L}\right)^{1/2} \left|\left| \widehat{D}^{(L)} W^{(L)} \cdots \widehat{D}^{(1)} W^{(1)} x_\alpha \right|\right|,$$

where $Z_1 \sim \mathcal{N}(0,1)$ is independent of $\widehat{D}^{(\ell)}, W^{(\ell)}$, $i = 1, \ldots, L$. Further, due to the right orthogonal invariance of the Gaussian matrices $W^{(\ell)}$ and the normalization that the variance of the entries of $W^{(\ell)}$ is $C_W / n_{\ell-11}$, we have that

$$\begin{aligned} &\log\left[\left(\frac{C_W}{n_L}\right)^{1/2} \left|\left| \widehat{D}^{(L)} W^{(L)} \cdots \widehat{D}^{(1)} W^{(1)} x_\alpha \right|\right| \right] \\ &\quad \stackrel{d}{=} \frac{1}{2} \log\left[\frac{C_W}{n_0} \left|\left| x_\alpha \right|\right|^2 \right] + \sum_{\ell=1}^{L} \frac{1}{2} \log\left[\frac{C_W}{n_L} \left|\left| \widehat{D}^{(\ell)} \widehat{W}^{(\ell)} u^{(\ell)} \right|\right|^2 \right] \end{aligned}$$

where $u^{(\ell)} \in \mathbb{R}^{n_{\ell-1}}$ is collection of deterministic unit vectors and $\widehat{W}^{(\ell)}$ are independent random matrices with iid standard Gaussian entries. The summands on the previous line

are independent and are each distributed like the logarithm of a randomly weighted χ^2 random variable:

$$\frac{C_W}{n_\ell}\left|\left|\widehat{D}^{(\ell)}W^{(\ell)}u^{(\ell)}\right|\right|^2 \stackrel{d}{=} \frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\left(d_i^{(\ell)}\right)^2\left(Z_i^{(\ell)}\right)^2,$$

where $Z_i^{(\ell)} \sim \mathcal{N}(0,1)$ are iid and independent of $d_i^{(\ell)}$. Putting this all together, we find that

$$z_{i;\alpha}^{(L+1)} \stackrel{d}{=} \left(\frac{C_W}{n_0}\left|\left|x_\alpha\right|\right|^2\right)^{1/2} \cdot Z_1 \cdot \prod_{\ell=1}^{L}\left(\frac{C_W}{n_\ell}\right)^{1/2}\left|\left|\widehat{D}^{(\ell)}W^{(\ell)}u^{(\ell)}\right|\right|$$

is a product of $L+1$ independent random variables. Moreover, a direct computation shows that

$$\begin{aligned}\mathbb{E}\left[\log\left[\frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\left(d_i^{(\ell)}\right)^2\left(Z_i^{(\ell)}\right)^2\right]\right] &= -\frac{1}{2}\mathrm{Var}\left[\frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\left(d_i^{(\ell)}\right)^2\left(Z_i^{(\ell)}\right)^2\right] + O(n_\ell^{-2}) \\ &= -\frac{1}{2n_\ell}\left(6\frac{a_+^4+a_-^4}{(a_+^2+a_-^2)^2}-1\right) + O(n_\ell^{-2})\end{aligned}$$

and also that

$$\begin{aligned}\mathrm{Var}\left[\log\left[\frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\left(d_i^{(\ell)}\right)^2\left(Z_i^{(\ell)}\right)^2\right]\right] &= \mathrm{Var}\left[\frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\left(d_i^{(\ell)}\right)^2\left(Z_i^{(\ell)}\right)^2\right] + O(n_\ell^{-2}) \\ &= \frac{1}{n_\ell}\left(6\frac{a_+^4+a_-^4}{(a_+^2+a_-^2)^2}-1\right) + O(n_\ell^{-2}).\end{aligned}$$

Combining the preceding two estimates, taking $n, L \to \infty$ with $L/n \to \gamma$ and applying the CLT yields

$$\lim_{\substack{n,L\to\infty \\ L/n\to\gamma\in[0,\infty)}} z_{i;\alpha}^{(L)} \quad \stackrel{d}{\longrightarrow} \quad \left(\frac{C_W}{n_0}\left|\left|x_\alpha\right|\right|^2\right)^{1/2} Z_1 \exp\left[-\mu(\gamma,a_+,a_-) + \sigma(\gamma,a_+,a_-)Z_2\right],$$

where

$$\mu(\gamma,a_+,a_-) = \sigma^2(\gamma,a_+,a_-) := \frac{\gamma}{4}\left(6\frac{a_+^4+a_-^4}{(a_+^2+a_+^2)^2}-1\right), \quad Z_1, Z_2 \sim \mathcal{N}(0,1) \text{ iid.}$$

This is precisely the statement of (76).