The Annals of Applied Probability
2023, Vol. 33, No. 6A, 4798–4819
https://doi.org/10.1214/23-AAP1933

RANDOM NEURAL NETWORKS IN THE INFINITE WIDTH LIMIT AS GAUSSIAN PROCESSES

BY BORIS HANIN[a]

Department of Operations Research and Financial Engineering, Princeton University , [a]*bhanin@princeton.edu*

This article gives a new proof that fully connected neural networks with random weights and biases converge to Gaussian processes in the regime where the input dimension, output dimension, and depth are kept fixed, while the hidden layer widths tend to infinity. Unlike prior work, convergence is shown assuming only moment conditions for the distribution of weights and for quite general nonlinearities.

1. Introduction. In the last decade or so neural networks, originally introduced in the 1940's and 50's [29, 44], have become indispensable tools for machine learning tasks ranging from computer vision [32] to natural language processing [9] and reinforcement learning [46]. Their empirical success has raised many new mathematical questions in approximation theory [13, 54, 55], probability (see §1.2.2 for some references), optimization/learning theory [7, 8, 31, 57] and so on. The present article concerns a fundamental probabilistic question about arguably the simplest networks, the so-called *fully connected* neural networks, defined as follows:

DEFINITION 1.1 (Fully connected network). Fix a positive integer L as well as $L+2$ positive integers $n_0, \dots, n_{L+1}$ and a function $\sigma : \mathbb{R} \to \mathbb{R}$. A fully connected depth L neural network with input dimension n_0, output dimension n_{L+1}, hidden layer widths $n_1, \dots, n_L$, and nonlinearity σ is any function $x_\alpha \in \mathbb{R}^{n_0} \mapsto z_\alpha^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ of the following form

$$z_\alpha^{(\ell)} = \begin{cases} W^{(1)} x_\alpha + b^{(1)} & \ell = 1, \\ W^{(\ell)} \sigma\big(z_\alpha^{(\ell-1)}\big) + b^{(\ell)} & \ell = 2, \dots, L+1, \end{cases}$$

where $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ are matrices, $b^{(\ell)} \in \mathbb{R}^{n_\ell}$ are vectors, and σ applied to a vector is shorthand for σ applied to each component.

The parameters $L, n_0, \dots, n_{L+1}$ are called the *network architecture*, and $z_\alpha^{(\ell)} \in \mathbb{R}^{n_\ell}$ is called the *vector of pre-activations at layer* ℓ corresponding to input x_α. A fully connected network with a fixed architecture and given nonlinearity σ is therefore a finite (but typically high) dimensional family of functions, parameterized by the network weights (entries of the weight matrices $W^{(\ell)}$) and biases (components of bias vectors $b^{(\ell)}$).

This article considers the mapping $x_\alpha \mapsto z_\alpha^{(L+1)}$ when the network's weights and biases are chosen independently at random and the hidden layer widths $n_1, \dots, n_L$ are sent to infinity while the input dimension n_0, output dimension n_{L+1}, and network depth L are fixed. In this *infinite width limit*, akin to the large matrix limit in random matrix theory (see §1.2), neural networks with random weights and biases converge to Gaussian processes (see §1.4 for a review of prior work). Unlike prior work Theorem 1.2, our main result, is that this holds for general nonlinearities σ and distributions of network weights (cf §1.3).

Received July 2021; revised December 2022.
MSC2020 subject classifications. Primary 68T07; secondary 60G15.
Key words and phrases. Neural networks, Gaussian processes, limit theorems.

Moreover, in addition to establishing convergence of wide neural networks to a Gaussian process under weak hypotheses, the present article gives a mathematical take aimed at probabilists of some of the ideas developed in the recent monograph [43]. This book, written in the language and style of theoretical physics by Roberts and Yaida, is based on research done jointly with the author. It represents a far-reaching development of the breakthrough work of Yaida [49], which was the first to systematically explain how to compute *finite width corrections* to infinite width Gaussian process limit of random neural networks for arbitrary depth, width, and nonlinearity. Previously, such finite width (and large depth) corrections were only possible for some special observables in linear and ReLU networks [21, 23–25, 39, 56]. The present article deals only with the asymptotic analysis of random neural networks as the width tends to infinity, leaving to future work a probabilistic elaboration of some aspects of the approach to finite width corrections from [43].

1.1. *Roadmap.* The rest of this article is organized as follows. First, in §1.2 we briefly motivate the study of neural networks with random weights. Then, in §1.3 we formulate our main result, Theorem 1.2. Before giving its proof in §2, we first indicate in §1.4 the general idea of the proof and its relation to prior work.

1.2. *Motivation for studying random neural networks.*

1.2.1. *Practical motivations.* It may seem at first glance that studying neural networks with random weights and biases is of no practical interest. After all, a neural network is only useful after it has been "trained," that is, one has found a setting of its parameters so that the resulting network function (at least approximately) interpolates a given training dataset of input-output pairs $(x, f(x))$ for an otherwise unknown function $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$.

However, the vast majority of neural network training algorithms used in practice are variants of gradient descent starting from *a random initialization* of the weight matrices $W^{(\ell)}$ and bias vectors $b^{(\ell)}$. Studying networks with random weights and biases therefore provides an understanding of the initial conditions for neural network optimization.

Beyond illuminating the properties of networks at the start of training, the analysis of random neural networks can reveal a great deal about networks after training as well. Indeed, on a heuristic level, just as the behavior of the level spacings of the eigenvalues of large random matrices is a surprisingly good match for emission spectra of heavy atoms [48], it is not unreasonable to believe that certain coarse properties of the incredibly complex networks used in practice will be similar to those of networks with random weights and biases. More rigorously, neural networks used in practice often have many more tunable parameters (weights and biases) than the number of datapoints from the training dataset. Thus, at least in certain regimes, neural network training provably proceeds by an approximate linearization around initialization, since no one parameter needs to move much to fit the data. This so-called NTK analysis [14, 16, 30, 31, 35] shows, with several important caveats related to network size and initialization scheme, that in some cases the statistical properties of neural networks at the start of training are the key determinants of their behavior throughout training.

1.2.2. *Motivation from random matrix theory.* In addition to being of practical importance, random neural networks are also fascinating mathematical objects, giving rise to new problems in approximation theory [12, 13, 22, 54, 55], random geometry [26, 27], and random matrix theory (RMT). Perhaps the most direct, though by no means only, connection to RMT questions is to set the network biases $b^{(\ell)}$ to zero and consider the very special case when $\sigma(t) = t$ is the identity (in the machine learning literature these are called deep linear networks). The network function

$$z_\alpha^{(L+1)} = W^{(L+1)} \cdots W^{(1)} x_\alpha \tag{1.1}$$

is then a linear statistic for a product of $L+1$ independent random matrices. Such matrix models have been extensively studied, primarily in two regimes. The first is the multiplicative ergodic theorem regime [10, 17, 18, 45], in which all the layer widths $n_0, \dots, n_{L+1}$ are typically set to a fixed value n and the network depth L tends to infinity. The second regime, where L is fixed and the layer widths n_ℓ (i.e., matrix dimensions) tend to infinity, is the purview of free-probability [38, 47].

In the presence of a nonlinearity σ, random neural networks provide nonlinear generalizations of the usual RMT questions. For instance, the questions taken up in this article are analogs of the joint normality of linear statistics of random matrix products in the free probability regime. Further, random neural networks give additional motivation for studying matrix products appearing in (1.1) when the matrix dimensions n_ℓ and the number of terms L are simultaneously large. This double scaling limit reveals new phenomena [3–6, 20, 24, 25] but is so far poorly understood relative to the ergodic or free regimes.

Finally, beyond studying linear networks, random matrix theory questions naturally appear in neural network theory via nonlinear analogs of the Marchenko–Pastur distribution for empirical covariance matrices of $z_\alpha^{(L+1)}$ when $\alpha \in A$ ranges over a random dataset of inputs [1, 28, 41, 42] as well as through the spectrum of the input-output Jacobian [24, 42] and the NTK [2, 16].

1.3. *Main result.* Our main result shows that under rather general conditions, when the weights $W^{(\ell)}$ and biases $b^{(\ell)}$ of a fully connected network are chosen at random, the resulting field $x_\alpha \mapsto z_\alpha^{(L+1)}$ converges to a centered Gaussian field with iid components when the input dimension n_0 and output dimension n_{L+1} are held fixed but the hidden layer widths $n_1, \dots, n_L$ tend to infinity. To give the precise statement in Theorem 1.2 below, fix a fully connected neural network with depth $L \geq 1$, input dimension n_0, output dimension n_{L+1}, hidden layer widths $n_1, \dots, n_L \geq 1$, and nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$. We assume that σ is absolutely continuous and that its almost-everywhere defined derivative (and hence σ itself) is polynomially bounded:

$$\exists k > 0 \text{ s.t. } \forall x \in \mathbb{R} \quad \left\| \frac{\sigma'(x)}{1+|x|^k} \right\|_{L^\infty(\mathbb{R})} < \infty. \tag{1.2}$$

All nonlinearities used in practice satisfy these rather mild criteria. Further, let us write $W_{ij}^{(\ell)}$ for the entries of the weight matrices $W^{(\ell)}$ and $b_i^{(\ell)}$ for the components of the bias vectors $b^{(\ell)}$. For $\ell \geq 2$ the Definition 1.1 of fully connected networks means that the formula for the components of the pre-activations $z_\alpha^{(\ell)}$ at layer ℓ in terms of those for $z_\alpha^{(\ell-1)}$ reads

$$z_{i;\alpha}^{(\ell)} := b_i^{(\ell)} + \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \sigma\big(z_{j;\alpha}^{(\ell-1)}\big), \quad i = 1, \dots, n_\ell, \tag{1.3}$$

where we've denoted by $z_{i;\alpha}^{(\ell)}$ the ith component of the n_ℓ-dimensional vector of pre-activations $z_\alpha^{(\ell)}$ in layer ℓ corresponding to a network input $x_\alpha \in \mathbb{R}^{n_0}$. We make the following assumption on the network weights:

$$W_{ij}^{(\ell)} := \left(\frac{C_W}{n_{\ell-1}}\right)^{1/2} \widehat{W}_{ij}^{(\ell)}, \qquad \widehat{W}_{ij}^{(\ell)} \sim \mu \quad \text{iid}, \tag{1.4}$$

where μ is a fixed probability distribution on $\mathbb{R}$ such that

$$\mu \text{ has mean 0, variance 1, and finite higher moments.} \tag{1.5}$$

We further assume the network biases are iid Gaussian[1] and independent of the weights:

$$b_i^{(\ell)} \sim \mathcal{N}(0, C_b) \quad \text{iid.} \tag{1.6}$$

In (1.4) and (1.6), $C_W > 0$ and $C_b \geq 0$ are fixed constants. These constants do not play an important role for the analysis in this article but will be crucial for followup work. With the network weights and biases chosen at random the vectors $z_\alpha^{(\ell)}$ are also random. Our main result is that, in the infinite width limit, they have independent Gaussian components.

THEOREM 1.2. *Fix n_0, n_{L+1} and a compact set $T \subseteq \mathbb{R}^{n_0}$. As the hidden layer widths $n_1, \ldots, n_L$ tend to infinity, the sequence of stochastic processes*

$$x_\alpha \in \mathbb{R}^{n_0} \quad \mapsto \quad z_\alpha^{(L+1)} \in \mathbb{R}^{n_{L+1}}$$

converges weakly in $C^0(T, \mathbb{R}^{n_{L+1}})$ to a centered Gaussian process taking values in $\mathbb{R}^{n_{L+1}}$ with iid coordinates. The coordinate-wise covariance function

$$K_{\alpha\beta}^{(L+1)} := \lim_{n_1,\ldots,n_L \to \infty} \operatorname{Cov}(z_{i;\alpha}^{(L+1)}, z_{i;\beta}^{(L+1)})$$

for this limiting process satisfies the layerwise recursion

$$K_{\alpha\beta}^{(\ell+1)} = C_b + C_W \mathbb{E}\big[\sigma(z_\alpha)\sigma(z_\beta)\big], \qquad \begin{pmatrix} z_\alpha \\ z_\beta \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K_{\alpha\alpha}^{(\ell)} & K_{\alpha\beta}^{(\ell)} \\ K_{\alpha\beta}^{(\ell)} & K_{\beta\beta}^{(\ell)} \end{pmatrix}\right) \tag{1.7}$$

for $\ell \geq 2$, with initial condition

$$K_{\alpha\beta}^{(2)} = C_b + C_W \mathbb{E}\big[\sigma(z_{1;\alpha}^{(1)})\sigma(z_{1;\beta}^{(1)})\big], \tag{1.8}$$

where the distribution of $(z_{1;\alpha}^{(1)}, z_{1;\beta}^{(1)})$ is determined via (1.3) by the distribution of weights and biases in the first layer and is not universal.

We prove Theorem 1.2 in §2. First, we explain the main idea and review prior work.

1.4. *Theorem 1.2: Discussion, main idea, and relation to prior work.* At a high level, the proof of Theorem 1.2 (specifically the convergence of finite-dimensional distributions) proceeds as follows:

1. Conditional on the mapping $x_\alpha \mapsto z_\alpha^{(L)}$, the components of the neural network output $x_\alpha \mapsto z_\alpha^{(L+1)}$ are independent sums of n_L independent random fields (see (1.3)), and hence, when n_L is large, are each approximately Gaussian by the CLT.
2. The conditional covariance in the CLT from step 1 is random at finite widths (it depends on $z_\alpha^{(L)}$). However, it has the special form of an average over $j = 1, \ldots, n_L$ of the same function applied to each component $z_{j;\alpha}^{(L)}$ of the vector $z_\alpha^{(L)}$ of pre-activations at the last hidden layer. We call such objects *collective observables* (see §2.1.2 and (1.10)).
3. While $z_{j;\alpha}^{(\ell)}$ are not independent at finite width when $\ell \geq 2$, they are sufficiently weakly correlated that a LLN still applies to any collective observable in the infinite width limit (see §2.1.2).
4. The LLN from step 3 allows us to replace the random conditional covariance matrix from steps 1 and 2 by its expectation, asymptotically as $n_1, \ldots, n_L$ tend to infinity.

[1] As explained in §1.4 the universality results in this article are simply not true if the biases are drawn iid from a fixed non-Gaussian distribution.

We turn to giving a few more details on steps 1-4 and reviewing along the way the relation of the present article to prior work. The study of the infinite width limit for random neural networks dates back at least to Neal [37], who considered networks with one hidden layer:

$$z^{(2)}_{i;\alpha} = b^{(2)}_i + \sum_{j=1}^{n_1} W^{(2)}_{ij}\sigma\big(z^{(1)}_{j;\alpha}\big), \qquad z^{(1)}_{j;\alpha} = b^{(1)}_j + \sum_{k=1}^{n_0} W^{(1)}_{jk} x_{k;\alpha},$$

where $i = 1, \dots, n_2$. In the shallow $L = 1$ setting of Neal if in addition $n_2 = 1$, then neglecting the bias $b^{(2)}_1$ for the moment, the scalar field $z^{(2)}_{1;\alpha}$ is a sum of iid random fields with finite moments, and hence the asymptotic normality of its finite-dimensional distributions follows immediately from the multidimensional CLT. Modulo tightness, this explains why $z^{(2)}_{1;\alpha}$ ought to converge to a Gaussian field. Even this simple case, however, holds several useful lessons:

- If the distribution of the bias $b^{(2)}_1$ is fixed independent of n_1 is and non-Gaussian, then the distribution of $z^{(2)}_{1;\alpha}$ will not be Gaussian, even in the limit when $n_1 \to \infty$.
- If the first layer biases $b^{(1)}_j$ are drawn iid from a fixed distribution μ_b and σ is nonlinear, then higher moments of μ_b will contribute to the variance of each neuron post-activation $\sigma(z^{(1)}_{j;\alpha})$, causing the covariance of the Gaussian field at infinite width to be nonuniversal.
- Unlike in deeper layers, as long as n_0 is fixed, the distribution of each neuron pre-activation $z^{(1)}_{j;\alpha}$ in the first layer will not be Gaussian, unless the weights and biases in layer 1 are themselves Gaussian. This explains why, in the initial condition (1.8) the distribution is non-Gaussian in the first layer.

In light of the first two points, what should one assume about the bias distribution? There are, it seems, two options. The first is to assume that the variance of the biases tends to zero as $n_1 \to \infty$, putting them on par with the weights. The second, which we adopt in this article, is to declare all biases to be Gaussian.

The first trick in proving Theorem 1.2 for general depth and width appears already when $L = 1$ but the output dimension n_2 is at least two.[2] In this case, even for a single network input x_α, at finite values of the network width n_1 different components of the random n_2-dimensional vector $z^{(2)}_\alpha$ are not independent, due to their shared dependence on the vector $z^{(1)}_\alpha$. The key observation, which to the author's knowledge was first presented in [34], is to note that the components of $z^{(2)}_\alpha$ are independent *conditional on the first layer* (i.e., on $z^{(1)}_\alpha$) and are approximately Gaussian when n_1 is large by the CLT. The conditional variance, which captures the main dependence on $z^{(1)}_\alpha$, has the following form:

$$\Sigma^{(2)}_{\alpha\alpha} := \mathrm{Var}\big[z^{(2)}_{i;\alpha} | z^{(1)}_\alpha\big] = C_b + \frac{C_W}{n_1}\sum_{j=1}^{n_1}\sigma\big(z^{(1)}_{j;\alpha}\big)^2. \tag{1.9}$$

This is an example of what we will call a *collective observable*, an average over all neurons in a layer of the same function applied to the pre-activations at each neuron (see §2.1.2 for the precise definition). In the shallow $L = 1$ setting, $\Sigma^{(2)}_{\alpha\alpha}$ is a sum of n_1 iid random variables with finite moments. Hence, by the LLN, it converges almost surely to its mean as $n_1 \to \infty$. This causes the components of $z^{(2)}_\alpha$ to become independent in the infinite width limit, since the source of their shared randomness, $\Sigma^{(2)}_{\alpha\alpha}$, can be replaced asymptotically by its expectation.

The proof for general L follows a similar pattern. Exactly as before, for any $0 \le \ell \le L$, the components of the pre-activations at layer $\ell + 1$ are still conditionally independent, given

[2] Neal [37] states erroneously on page 38 of his thesis that $z^{(2)}_{i;\alpha}$ and $z^{(2)}_{j;\alpha}$ will be independent because the weights going into them are independent. This is not true at finite width but becomes true in the infinite width limit.

the pre-activations at layer ℓ. When the width n_ℓ is large the conditional distribution of each component over any finite collection of inputs is therefore approximately Gaussian by the CLT. Moreover, the conditional covariance across network inputs has the form:

$$\Sigma_{\alpha\beta}^{(\ell+1)} := \mathrm{Cov}\big(z_{i;\alpha}^{(\ell+1)}, z_{i;\beta}^{(\ell+1)} | z_\alpha^{(\ell)}, z_\beta^{(\ell)}\big) = C_b + \frac{C_W}{n_\ell}\sum_{j=1}^{n_\ell}\sigma\big(z_{j;\alpha}^{(\ell)}\big)\sigma\big(z_{j;\beta}^{(\ell)}\big). \tag{1.10}$$

The summands on the right hand side are no longer independent at finite width if $\ell \geq 2$. However, $\Sigma_{\alpha\beta}^{(\ell+1)}$ are still collective observables, and the crucial point is to check that their dependence is sufficiently weak that we may still apply the LLN. Verifying this is the heart of the proof of Theorem 1.2 and is carried out in §2.1.2.

Let us mention that, in addition to the approach outlined above, other methods for showing that wide neural networks are asymptotically Gaussian processes are possible. In the prior article [36], for instance, the idea is to use that the entries of $z_\alpha^{(\ell)}$ are exchangeable and argue using an exchangeable CLT. This leads to some technical complications which, at least in the way the argument is carried out in [36], result in unnatural restrictions on the class of nonlinearities and weight distributions considered there. Let us also mention that in the article [34], the nonindependence at finite width of the components of $z_\alpha^{(\ell)}$ for large ℓ was circumvented by considering only the sequential limit in which $n_\ell \to \infty$ in order of increasing ℓ. The effect is that for every ℓ the conditional covariance $\Sigma_{\alpha\beta}^{(\ell)}$ has already converged to its mean before $n_{\ell+1}$ is taken large. However, this way of taking the infinite width limit seems to the author somewhat unnatural and is in any case not conducive to studying finite width corrections as in [43, 49], which we plan to take up in future work.

We conclude this section by pointing the reader to several other related strands of work. The first are articles such as [11], which quantify the magnitude of the difference

$$\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} z_{i;\alpha}^{(\ell)} z_{i;\beta}^{(\ell)} - \lim_{n_1,\ldots,n_\ell\to\infty} \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} z_{i;\alpha}^{(\ell)} z_{i;\beta}^{(\ell)}\right]$$

between the empirical overlaps $n_\ell^{-1}\langle z_\alpha^{(\ell)}, z_\beta^{(\ell)}\rangle$ of pre-activations and the corresponding infinite width limit uniformly over network inputs x_α, x_β in a compact subset of $\mathbb{R}^{n_0}$. In a similar vein are articles such as [15], which give quantitative estimates at finite width for the distance from $x_\alpha \mapsto z_\alpha^{(\ell)}$ to a nearby Gaussian process.

The second is the series of articles starting with the work of Yang [50–53], which develops the study not only of initialization but also certain aspects of inference with infinitely wide networks using what Yang terms tensor programs. As part of that series, the article [51] establishes that in the infinite width limit many different architectures become Gaussian processes. However, the arguments in those articles are significantly more technical than the ones presented here since they are focused on building the foundation for the tensor program framework. At any rate, to the best of the author's knowledge, no prior article addresses universality of the Gaussian process limit with respect to the weight distribution in deep networks (for shallow networks with $L = 1$ this was considered by Neal in [37]). Finally, that random neural networks converge to Gaussian processes in the infinite width limit under various restrictions but for architectures other than fully connected is taken up in [19, 40, 51].

2. Proof of Theorem 1.2. Let us recall the notation. Namely, we fix a network depth $L \geq 1$, an input dimension $n_0 \geq 1$, an output dimension $n_{L+1} \geq 1$, hidden layer widths $n_1, \ldots, n_L \geq 1$, and a nonlinearity σ satisfying (2.5). We further assume that the networks weights and biases are independent and random as in (1.4) and (1.6). To prove Theorem 1.2

we must show that the random fields $x_\alpha \mapsto z_\alpha^{(L+1)}$ converge weakly in distribution to a Gaussian process in the limit where $n_1, \dots, n_L$ tend to infinity. We start with the convergence of finite-dimensional distributions. Let us therefore fix a collection

$$x_A = \{x_\alpha, \alpha \in A\}$$

of $|A|$ distinct network inputs in $\mathbb{R}^{n_0}$ and introduce for each $\ell = 0, \dots, L+1$, every $i = 1, \dots, n_\ell$, and all $\alpha \in A$ the vectorized notation

$$z_{i;A}^{(\ell)} := \big(z_{i;\alpha}^{(\ell)}, \alpha \in A\big) \in \mathbb{R}^{|A|}, \qquad z_A^{(\ell)} := \big(z_{i;A}^{(\ell)}, i = 1, \dots, n_\ell\big) \in \mathbb{R}^{n_\ell \times |A|}.$$

The following result states that the distribution of the random variable $z_A^{(L+1)}$ with values in $\mathbb{R}^{n_{L+1} \times |A|}$ converges to that of the claimed Gaussian field.

PROPOSITION 2.1 (Convergence of finite-dimensional distributions). *Fix $L \geq 1$ and n_0, n_{L+1}. The distribution of $z_A^{(L+1)}$ converges weakly as $n_1, \dots, n_L \to \infty$ to that of a centered Gaussian in $\mathbb{R}^{n_{L+1} \times |A|}$ with iid rows for which the covariance*

$$K_{\alpha\beta}^{(L+1)} = \lim_{n_1, \dots, n_L \to \infty} \mathrm{Cov}\big(z_{i;\alpha}^{(L+1)}, z_{i;\beta}^{(L+1)}\big), \quad \alpha, \beta \in A$$

between the entries in each row satisfies the recursion (1.7) *with initial condition* (1.8).

Once we have proved Proposition 2.1 in §2.1, it remains to show tightness. For this, we fix a compact subset $T \subseteq \mathbb{R}^{n_0}$. The tightness of $x_\alpha \mapsto z_\alpha^{(L+1)}$ in $C^0(T, \mathbb{R}^{n_{L+1}})$ follows immediately from the Arzelà–Ascoli theorem and the following result, which we prove in §2.2.

PROPOSITION 2.2 (High probability equicontinuity and equiboundedness of $z_\alpha^{(L+1)}$). *For every $L \geq 1$, $\epsilon > 0$ there exists $C = C(\epsilon, \sigma, T, L, C_b, C_W, n_{L+1}) > 0$ so that*

$$\sup_{x_\alpha, x_\beta \in T} \frac{\|z_\alpha^{(L+1)} - z_\beta^{(L+1)}\|_2}{\|x_\alpha - x_\beta\|_2} \leq C \quad \text{and} \quad \sup_{x_\alpha \in T} \|z_\alpha^{(L+1)}\| \leq C \tag{2.1}$$

with probability at least $1 - \epsilon$.

2.1. *Finite-dimensional distributions*: *Proof of Proposition 2.1.* We will prove Proposition 2.1 in two steps. First, we prove a special case in which we keep the weights in layer 1 completely general as in (1.4) but take weights in layers $\ell \geq 2$ to be independent Gaussians:

$$W_{ij}^{(\ell)} \sim \mathcal{N}\big(0, C_W n_{\ell-1}^{-1}\big), \quad \text{iid.}$$

We continue to assume (as in the statement of Theorem 1.2) that all biases are Gaussian:

$$b_i^{(\ell)} \sim \mathcal{N}(0, C_b), \quad \text{iid.}$$

The argument in this case is the technical heart of this paper and is presented in §2.1.1–§2.1.2. Ultimately, it relies on the analysis of collective observables, which we isolate in §2.1.2. A simple Lindeberg swapping argument and induction on layer detailed in §2.1.3 allows us to extend Proposition 2.1 to general weights in layers $\ell \geq 2$ from the Gaussian case.

2.1.1. *Proof of Proposition 2.1 with Gaussian weights in layers $\ell \geq 2$.* Fix

$$\Xi = (\xi_i, i = 1, \ldots, n_{L+1}) \in \mathbb{R}^{n_{L+1} \times |A|}, \qquad \xi_i = (\xi_{i;\alpha}, i = 1, \ldots, n_{L+1}, \alpha \in A) \in \mathbb{R}^{|A|}$$

and consider the characteristic function

$$\chi_A(\Xi) = \mathbb{E}\big[\exp\big[-i\langle z_A^{(L+1)}, \Xi\rangle\big]\big] = \mathbb{E}\left[\exp\left[-i \sum_{\alpha \in A} \sum_{i=1}^{n_{L+1}} z_{i;\alpha}^{(L+1)} \xi_{i;\alpha}\right]\right]$$

of the random variable $z_A^{(L+1)} \in \mathbb{R}^{n_{L+1} \times |A|}$. By Levy's continuity theorem, it is sufficient to show that

$$\lim_{n_1, \ldots, n_L \to \infty} \chi_A(\Xi) = \exp\left[-\frac{1}{2} \sum_{i=1}^{n_{L+1}} \langle K_A^{(L+1)} \xi_i, \xi_i\rangle\right], \tag{2.2}$$

where

$$K_A^{(L+1)} = \big(K_{\alpha\beta}^{(L+1)}\big)_{\alpha,\beta \in A}$$

is the matrix defined by the recursion (1.7) with initial condition (1.8). Writing

$$\mathcal{F}_\ell := \text{filtration defined by } \big\{W^{(\ell')}, b^{(\ell')}, \ell' = 1, \ldots, \ell\big\}, \tag{2.3}$$

we may use the tower property to write

$$\chi_A(\Xi) = \mathbb{E}\big[\mathbb{E}\big[\exp\big[-i\langle z_A^{(L+1)}, \Xi\rangle\big] | \mathcal{F}_L\big]\big]. \tag{2.4}$$

Note that conditional on $\mathcal{F}_L$, the random vectors $z_{i;A}^{(L+1)} \in \mathbb{R}^{|A|}$ in layer $L+1$ for each $i = 1, \ldots, n_{L+1}$ are iid Gaussians, since we have assumed for now that weights in layers $\ell \geq 2$ are Gaussian. Specifically,

$$z_{i;A}^{(L+1)} \overset{d}{=} \big(\Sigma_A^{(L+1)}\big)^{1/2} G_i, \qquad G_i \sim \mathcal{N}(0, \mathrm{I}_{|A|}) \quad \text{iid } 1 \leq i \leq n_{L+1},$$

where for any $\alpha, \beta \in A$ the conditional covariance is

$$\big(\Sigma_A^{(L+1)}\big)_{\alpha\beta} = \mathrm{Cov}\big(z_{i;\alpha}^{(L+1)}, z_{i;\beta}^{(L+1)} | \mathcal{F}_L\big) = C_b + \frac{C_W}{n_L} \sum_{j=1}^{n_L} \sigma\big(z_{j;\alpha}^{(L)}\big)\sigma\big(z_{j;\beta}^{(L)}\big). \tag{2.5}$$

Using (2.4) and the explicit form of the characteristic function of a Gaussian reveals

$$\chi_A(\Xi) = \mathbb{E}\left[\exp\left[-\frac{1}{2} \sum_{i=1}^{n_{L+1}} \langle \Sigma_A^{(L+1)} \xi_i, \xi_i\rangle\right]\right]. \tag{2.6}$$

The crucial observation is that each entry of the conditional covariance matrix $\Sigma_A^{(L+1)}$ is an average over $j = 1, \ldots, n_L$ of the same fixed function applied to the vector $z_{j;A}^{(L)}$. While $z_{j;A}^{(L)}$ are not independent at finite values of $n_1, \ldots, n_{L-1}$ for $L > 1$, they are sufficiently weakly correlated that a weak law of large numbers still holds:

LEMMA 2.3. *Fix n_0, n_{L+1}. There exists a $|A| \times |A|$ PSD matrix*

$$K_A^{(L+1)} = \big(K_{\alpha\beta}^{(L+1)}\big)_{\alpha,\beta \in A}$$

such that for all $\alpha, \beta \in A$

$$\lim_{n_1, \ldots, n_L \to \infty} \mathbb{E}\big[\big(\Sigma_A^{(L+1)}\big)_{\alpha\beta}\big] = K_{\alpha\beta}^{(L+1)} \quad \textit{and} \quad \lim_{n_1, \ldots, n_L \to \infty} \mathrm{Var}\big[\big(\Sigma_A^{(L+1)}\big)_{\alpha\beta}\big] = 0.$$

PROOF. Lemma 2.3 is a special case of Lemma 2.4 (see §2.1.2). □

Lemma 2.3 implies that $\Sigma_A^{(L+1)}$ converges in distribution to $K_A^{(L+1)}$. In view of (2.6) and the definition of weak convergence this immediately implies (2.2). It therefore remains to check that $K_A^{(L+1)}$ satisfies the desired recursion. For this, note that at any values of $n_1, \dots, n_L$ we find

$$\mathrm{Cov}(z_{i;\alpha}^{(L+1)}, z_{i;\beta}^{(L+1)}) = \mathbb{E}[(\Sigma_A^{(L+1)})_{\alpha\beta}] = C_b + C_W \mathbb{E}[\sigma(z_{1;\alpha}^{(L)})\sigma(z_{1;\beta}^{(L)})].$$

When $L = 1$, we therefore see that

$$\mathrm{Cov}(z_{i;\alpha}^{(2)}, z_{i;\beta}^{(2)}) = C_b + C_W \mathbb{E}[\sigma(z_{i;\alpha}^{(1)})\sigma(z_{i;\beta}^{(1)})],$$

where the law of $(z_{i;\alpha}^{(1)}, z_{i;\beta}^{(1)})$ is determined by the distribution μ_W of weights in layer 1 and does not depend on n_1. This confirms the initial condition (1.8). Otherwise, if $L > 1$, the convergence of finite-dimensional distributions that we've already established yields

$$K_{\alpha\beta}^{(L+1)} = \lim_{n_1,\dots,n_L \to \infty} \mathrm{Cov}(z_{i;\alpha}^{(L+1)}, z_{i;\beta}^{(L+1)}) = \lim_{n_1,\dots,n_{L-1} \to \infty} (C_b + C_W \mathbb{E}[\sigma(z_{1;\alpha}^{(\ell)})\sigma(z_{1;\beta}^{(\ell)})]).$$

Since σ is continuous we may invoke the continuous mapping theorem to conclude that

$$K_{\alpha\beta}^{(L+1)} = C_b + C_W \mathbb{E}_{(z_\alpha, z_\beta) \sim G(0, K^{(L)})}[\sigma(z_\alpha)\sigma(z_\beta)],$$

which confirms the recursion (1.7). This completes the proof that the finite-dimensional distributions of $z_\alpha^{(L+1)}$ converge to those of the desired Gaussian process, modulo two issues. First, we must prove Lemma 2.3. This is done in §2.1.2 by proving a more general result, Lemma 2.4. Second, we must remove the assumption that the weights in layers $\ell \geq 2$ are Gaussian. This is done in §2.1.3. □

2.1.2. *Collective observables with Gaussian weights*: *Generalizing Lemma* 2.3. This section contains the key technical argument in our proof of Proposition 2.1. To state the main result, define a *collective observable at layer* ℓ to be any random variable of the form

$$\mathcal{O}_{n_\ell, f; A}^{(\ell)} := \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} f(z_{i;A}^{(\ell)}),$$

where $f : \mathbb{R}^{|A|} \to \mathbb{R}$ is measurable and polynomially bounded:

$$\exists C > 0, k \geq 1 \text{ s.t. } \forall z \in \mathbb{R}^{|A|} \quad |f(z)| \leq C(1 + \|z\|_2^k).$$

We continue to assume, as §2.1.1, that the weights (and biases) in layers $\ell \geq 2$ are Gaussian and will remove this assumption in §2.1.3. Several key properties of collective observables are summarized in the following Lemma:

LEMMA 2.4 (Key properties of collective observables). *Let* $\mathcal{O}_{n_\ell, f; A}^{(\ell)}$ *be a collective observable at layer* ℓ. *There exists a deterministic scalar* $\overline{O}_{f;A}^{(\ell)}$ *such that*

$$\lim_{n_1,\dots,n_{\ell-1} \to \infty} \mathbb{E}[\mathcal{O}_{n_\ell, f; A}^{(\ell)}] = \overline{O}_{f;A}^{(\ell)}. \tag{2.7}$$

Moreover,

$$\lim_{n_1,\dots,n_{\ell} \to \infty} \mathrm{Var}[\mathcal{O}_{n_\ell, f; A}^{(\ell)}] = 0. \tag{2.8}$$

Hence, *we have the following convergence in probability*

$$\lim_{n_1,\dots,n_{\ell} \to \infty} \mathcal{O}_{n_\ell, f; A}^{(\ell)} \xrightarrow{p} \overline{O}_{f;A}^{(\ell)}.$$

PROOF. This proof is by induction on ℓ, starting with $\ell = 1$. In this case, $z^{(1)}_{i;A}$ are independent for different i. Moreover, for each i, α we have

$$z^{(1)}_{i;\alpha} = b^{(1)}_i + \sum_{j=1}^{n_0} W^{(1)}_{ij} x_{j;\alpha} = b^{(1)}_i + \left(\frac{C_W}{n_0}\right)^{1/2} \sum_{j=1}^{n_0} \widehat{W}^{(1)}_{ij} x_{j;\alpha}.$$

Hence, $z^{(1)}_{i;\alpha}$ have finite moments since $b^{(1)}_i$ are iid Gaussian and $\widehat{W}^{(1)}_{ij}$ are mean 0 with finite higher moments. In particular, since f is polynomially bounded, we find for every n_1 that

$$\mathbb{E}[\mathcal{O}^{(1)}_{n_1,f;A}] = \mathbb{E}[f(z^{(1)}_{1;A})],$$

which is finite and independent of n_1, confirming (2.7). Further, $\mathcal{O}^{(1)}_{n_1,f;A}$ is the average of n_1 iid random variables with all moments finite. Hence, (2.8) follows by the weak law of large numbers, completing the proof of the base case.

Let us now assume that we have shown (2.7) and (2.8) for all $\ell = 1, \ldots, L$. We begin by checking that (2.7) holds at layer $L+1$. We have

$$\mathbb{E}[\mathcal{O}^{(L+1)}_{n_{L+1},f;A}] = \mathbb{E}[f(z^{(L+1)}_{1;A})]. \tag{2.9}$$

Since the weights and biases in layer $L+1$ are Gaussian and independent of $\mathcal{F}_L$, we find

$$z^{(L+1)}_{1;A} \overset{d}{=} (\Sigma^{(L+1)}_A)^{1/2} G, \tag{2.10}$$

where $\Sigma^{(L+1)}_A$ is the conditional covariance defined in (2.5) and G is an $|A|$-dimensional standard Gaussian. The key point is that $\Sigma^{(L+1)}_A$ is a collective observable at layer L. Hence, by the inductive hypothesis, there exists a PSD matrix $\overline{\Sigma}^{(L+1)}_A$ such that $\Sigma^{(L+1)}_A$ converges in probability to $\overline{\Sigma}^{(L+1)}_A$ as $n_1, \ldots, n_L \to \infty$. To establish (2.7) it therefore suffices in view of (2.9) to check that

$$\lim_{n_1,\ldots,n_L\to\infty} \mathbb{E}[f((\Sigma^{(L+1)}_A)^{1/2} G)] = \mathbb{E}[f((\overline{\Sigma}^{(L+1)}_A)^{1/2} G)], \tag{2.11}$$

where the right hand side is finite since f is polynomially bounded and all polynomial moments of G are finite. To establish (2.11), let us invoke the Skorohod representation theorem to find a common probability space on which there are versions of $\Sigma^{(L+1)}_A$—which by an abuse of notation we will still denote by $\Sigma^{(L+1)}_A$—that converge to $\overline{\Sigma}^{(L+1)}_A$ almost surely. Next, note that since f is polynomially bounded we may repeatedly apply $ab \le \frac{1}{2}(a^2 + b^2)$ to find

$$|f((\Sigma^{(L+1)}_A)^{1/2} G)| \le p((\Sigma^{(L+1)}_A)^{1/2}) + q(G), \tag{2.12}$$

where p is a polynomial in the entries of $(\Sigma^{(L+1)}_A)^{1/2}$, q a polynomial in the entries of G, and the polynomials p, q don't depend on $n_1, \ldots, n_L$. The continuous mapping theorem shows that

$$\lim_{n_1,\ldots,n_L\to\infty} \mathbb{E}[p((\Sigma^{(L+1)}_A)^{1/2})] = \mathbb{E}[p((\overline{\Sigma}^{(L+1)}_A)^{1/2})].$$

Thus, since all moments of Gaussian are finite, (2.11) follows from the generalized dominated convergence theorem. It remains to argue that (2.8) holds at layer $L+1$. To do this, we write

$$\begin{aligned} &\operatorname{Var}[\mathcal{O}^{(L+1)}_{n_{L+1},f;A}] \\ &\quad = \frac{1}{n_{L+1}} \operatorname{Var}[f(z^{(L+1)}_{1;A})] + \left(1 - \frac{1}{n_{L+1}}\right) \operatorname{Cov}(f(z^{(L+1)}_{1;A}), f(z^{(L+1)}_{2;A})). \end{aligned} \tag{2.13}$$

Observe that

$$\mathrm{Var}\big[f\big(z_{1;A}^{(L+1)}\big)\big] \le \mathbb{E}\big[f\big(z_{1;A}^{(L+1)}\big)^2\big] = \mathbb{E}\left[\frac{1}{n_{L+1}}\sum_{i=1}^{n_{L+1}} f\big(z_{i;A}^{(L+1)}\big)^2\right] < \infty,$$

since we already showed that (2.7) holds at layer $L+1$. Next, recall that, conditional on $\mathcal{F}_L$, neurons in layer $L+1$ are independent. The law of total variance and Cauchy–Schwartz yield

$$\begin{aligned}\big|\mathrm{Cov}\big(f\big(z_{1;A}^{(L+1)}\big), f\big(z_{2;A}^{(L+1)}\big)\big)\big| &= \big|\mathrm{Cov}\big(\mathbb{E}\big[f\big(z_{1;A}^{(L+1)}\big)|\mathcal{F}_L\big], \mathbb{E}\big[f\big(z_{2;A}^{(L+1)}\big)|\mathcal{F}_L\big]\big)\big| \\ &\le \mathrm{Var}\big[\mathbb{E}\big[f\big(z_{1;A}^{(L+1)}\big)|\mathcal{F}_L\big]\big].\end{aligned} \tag{2.14}$$

Using (2.10) and the polynomial estimates (2.12) on f, we conclude that the conditional expectation on the previous line is some polynomially bounded function of the components of $(\Sigma_A^{(L+1)})^{1/2}$. Hence, we may apply dominated convergence as above to find

$$\lim_{n_1,\dots,n_L\to\infty} \mathrm{Var}\big[\mathbb{E}\big[f\big(z_{1;A}^{(L+1)}\big)|\mathcal{F}_L\big]\big] = \mathrm{Var}\big[\mathbb{E}\big[f\big(\overline{\Sigma}_A^{1/2}\big)G\big]\big] = 0, \qquad G \sim \mathcal{N}(0, \mathrm{I}_{|A|})$$

since $\mathbb{E}[f(\overline{\Sigma}_A^{1/2})G]$ is a constant. This proves (2.8) for observables at layer $L+1$ and completes the proof of Lemma 2.4. □

2.1.3. *Proof of Proposition 2.1 for general weights.* In this section, we complete the proof of Proposition 2.1 by removing the assumption from §2.1.1 that weights in layers $\ell \ge 2$ are Gaussian. To do this, let us introduce some notation. Let us write

$$x_\alpha \mapsto z_\alpha^{(\ell)}$$

for the pre-activations at layers $\ell = 0, \dots, L+1$ of a random fully connected network in which, as in the general case of Theorem 1.2, all weights and biases are independent, biases are Gaussian as in (1.6) and weights in all layers are drawn from a general centered distribution with the correct variance and finite higher moments as in (1.4) and (1.5). Next, let us write

$$x_\alpha \mapsto \widetilde{z}_\alpha^{(\ell)},$$

for the vector of pre-activations obtained by replacing, in each layer $\ell = 2, \dots, L+1$, all weights $W_{ij}^{(\ell)}$ by iid centered Gaussians with variance $C_W/n_{\ell-1}$. We already saw that the distribution of $\widetilde{z}_A^{(L+1)}$ converges weakly to that of the desired Gaussian in the infinite width limit. Our goal is thus to show that, as $n_1, \dots, n_L$ tend to infinity, $z_A^{(L+1)}$ and $\widetilde{z}_A^{(L+1)}$ converge weakly to the same distribution. We will proceed by induction on L. When $L = 0$ the claim is trivial since, by construction, the weight and bias distributions in layer 1 are identical in both $z_\alpha^{(1)}$ and $\widetilde{z}_\alpha^{(1)}$ (recall that when we proved Proposition 2.1 in §2.1.1 we had Gaussian weights only starting from layer 2 and general weights in layer 1).

Suppose therefore that we have shown the claim for $\ell = 0, \dots, L$. By the Portmanteau theorem and the density of smooth compactly supported functions in the space of continuous compactly supported functions equipped with the supremum norm, it suffices to show that for any smooth function $g : \mathbb{R}^{n_{L+1}\times|A|} \to \mathbb{R}$ with compact support we have

$$\lim_{n_1,\dots,n_L\to\infty} \big(\mathbb{E}\big[g\big(z_A^{(L+1)}\big)\big] - \mathbb{E}\big[g\big(\widetilde{z}_A^{(L+1)}\big)\big]\big) = 0. \tag{2.15}$$

To check (2.15), let us define an intermediate object:

$$z_\alpha^{(L+1),\bullet} = b^{(L+1)} + W^{(L+1),\bullet}\sigma\big(z_\alpha^{(L)}\big),$$

where the entries $W_{ij}^{(L+1),\bullet}$ of $W^{(L+1),\bullet}$ are iid Gaussian with mean 0 and variance C_W/n_L. That is, we take the vector $\sigma(z_\alpha^{(L)})$ of post-activations from layer L obtained by using general weights in layers $1,\dots,L$ and use Gaussian weights only in layer $L+1$. Our first step in checking (2.15) is to show that this relation holds when $z_A^{(L+1)}$ is replaced by $z_A^{(L+1),\bullet}$.

LEMMA 2.5. *Let $x_A = \{x_\alpha, \alpha \in A\}$ be a finite subset of $\mathbb{R}^{n_0}$ consists of $|A|$ distinct elements. Fix in addition $n_{L+1} \geq 1$ and a smooth compactly supported function $g : \mathbb{R}^{n_{L+1}\times|A|} \to \mathbb{R}$. There exists $C > 0$ and a collective observable $\mathcal{O}_{n_L,f;A}^{(L)}$ at layer L so that*

$$|\mathbb{E}[g(z_A^{(L+1)})] - \mathbb{E}[g(z_A^{(L+1),\bullet})]| \leq C n_{L+1}^3 n_L^{-1/2} \mathbb{E}[\mathcal{O}_{n_L,f;A}^{(L)}].$$

PROOF. This is a standard Lindeberg swapping argument. Namely, for each $\alpha \in A$ and $k = 0,\dots,n_L$ define

$$z_\alpha^{(L+1),k} := b^{(L+1)} + W^{(L+1),k}\sigma(z_\alpha^{(L)}),$$

where the first k entries of each row of $W^{(L+1),k}$ are iid Gaussian with mean 0 and variance C_W/n_L, while the remaining entries are $(C_W/n_L)^{-1/2}$ times iid draws $\widehat{W}_{ij}^{(L+1)}$ from the general distribution μ of network weights, as in (1.4) and (1.5). With this notation, we have

$$z_\alpha^{(L+1)} = z_\alpha^{(L),0}, \qquad \widetilde{z}_\alpha^{(L+1),\bullet} = z_\alpha^{(L),n_L}.$$

Thus,

$$\mathbb{E}[g(z_A^{(L+1)})] - \mathbb{E}[g(\widetilde{z}_A^{(L+1)})] = \sum_{k=1}^{n_L} \mathbb{E}[g(z_A^{(L+1),k-1})] - \mathbb{E}[g(z_A^{(L+1),k})].$$

For any $Z \in \mathbb{R}^{n_{\ell+1}\times|A|}$ and

$$\delta Z = (\delta z_{i;\alpha} i = 1,\dots,n_{\ell+1}, \alpha \in A) \in \mathbb{R}^{n_{\ell+1}\times|A|}$$

consider the third order Taylor expansion of g around Z.

$$g(Z+\delta Z) = g(Z) + \sum_{\substack{\alpha\in A\\ i=1,\dots,n_{L+1}}} g_{i;\alpha}\delta z_{i;\alpha} + \sum_{\substack{\alpha_1,\alpha_2\in A\\ i_1,i_2=1,\dots,n_{L+1}}} g_{i_1,i_2;\alpha_1,\alpha_2}\delta z_{i_1;\alpha_1}\delta z_{i_2;\alpha_2}$$
$$+ O\Bigg(\sum_{\substack{\alpha_1,\alpha_2,\alpha_3\in A\\ i_1,i_2,i_3=1,\dots,n_{L+1}}} |\delta z_{i_1;\alpha_1}\delta z_{i_2;\alpha_2}\delta z_{i_3;\alpha_3}|\Bigg).$$

Let us write

$$z_{i;\alpha}^{(L+1),k-1} = z_{i;\alpha}^{(L+1),k} + n_L^{-1/2} Y_{i,k;\alpha}, \qquad \delta Z_{i,k;\alpha} = C_W^{1/2}(\widetilde{W}_{ik}^{(L+1)} - \widehat{W}_{ik}^{(L+1)})\sigma(z_{k;\alpha}^{(L)}),$$

where $\widetilde{W}_{ik}^{(L+1)} \sim \mathcal{N}(0,1)$. Then, Taylor expanding g to third order around $Z_k = z_{i;\alpha}^{(L+1),k}$ and, using that the first two moments of $(C_W n_L^{-1})^{1/2}\widehat{W}_{ij}^{(L)}$ match those of $\mathcal{N}(0, C_W n_L^{-1})$, we find that

$$\mathbb{E}[g(z_A^{(L+1),k-1})] - \mathbb{E}[g(\widetilde{z}_A^{(L+1),k})] = O(n_L^{-3/2} n_{L+1}^3 \mathbb{E}[p(|\sigma(z_{k;\alpha}^{(L)})|, \alpha \in A))]),$$

where p is a degree 3 polynomial in the $|A|$-dimensional vector of absolute values $|\sigma(z_{k;\alpha}^{(\ell)})|, \alpha \in A$. Summing this over k completes the proof. □

To make use of Lemma 2.5 let us consider any collective observable $\mathcal{O}_{n_L,f;A}^{(L)}$ at layer L. Recall that by (2.9) and (2.13) both the mean and variance of $\mathcal{O}_{n_L,f;A}^{(L)}$ depend only on the

distributions of finitely many components of the vector $z_A^{(L)}$. By the inductive hypothesis we therefore find

$$\lim_{n_1,\dots,n_L\to\infty} \mathbb{E}[\mathcal{O}^{(L)}_{n_L,f;A}] = \lim_{n_1,\dots,n_L\to\infty} \mathbb{E}[\widetilde{\mathcal{O}}^{(L)}_{n_L,f;A}], \tag{2.16}$$

where the right hand side means that we consider the same collective observable but for $\widetilde{z}_A^{(L)}$ instead of $z_A^{(L)}$, which exists by Lemma 2.4. Similarly, again using Lemma 2.4, we have

$$\lim_{n_1,\dots,n_L\to\infty} \mathrm{Var}[\mathcal{O}^{(L)}_{n_L,f;A}] = 0. \tag{2.17}$$

Therefore, we conclude that

$$\mathcal{O}^{(L)}_{n_L,f;A} - \widetilde{\mathcal{O}}^{(L)}_{n_L,f;A} \xrightarrow{d} 0, \quad \text{as } n_1,\dots,n_L\to\infty. \tag{2.18}$$

Note that by (2.16) the mean of any collective observable $\mathbb{E}[\mathcal{O}^{(L)}_{n_L,f;A}]$ is bounded independent of $n_1,\dots,n_L$ since we saw in Lemma 2.4 that the limit exists and is bounded when using Gaussian weights. Since n_{L+1} is fixed and finite, the error term $n_L^{-1/2}n_{L+1}^3\mathbb{E}[\mathcal{O}^{(L)}_{n_L,f;A}]$ in Lemma 2.5 is therefore tends to zero as $n_1,\dots,n_L\to\infty$, and (2.15) is reduced to showing that

$$\lim_{n_1,\dots,n_L\to\infty} \left(\mathbb{E}\left[g\left(z_A^{(L+1),\bullet}\right)\right] - \mathbb{E}\left[g\left(\widetilde{z}_A^{(L+1)}\right)\right]\right) = 0. \tag{2.19}$$

This follows from (2.17) and the inductive hypothesis. Indeed, by construction, conditional on the filtration $\mathcal{F}_L$ defined by weights and biases in layers up to L (see (2.3)), the $|A|$-dimensional vectors $z_{i;A}^{(L+1),\bullet}$ are iid Gaussians:

$$z_{i;A}^{(L+1),\bullet} \overset{d}{=} \left(\Sigma_A^{(L+1)}\right)^{1/2} G_i, \qquad G_i \sim \mathcal{N}(0, \mathrm{I}_{|A|}) \quad \text{iid},$$

where $\Sigma_A^{(L+1)}$ is the conditional covariance matrix from (2.5). The key point, as in the proof with all Gaussian weights, is that each entry of the matrix $\Sigma_A^{(L+1)}$ is a collective observable at layer L. Moreover, since the weights and biases in the final layer are Gaussian for $z_A^{(L+1),\bullet}$ the conditional distribution of $g(z_A^{(L+1),\bullet})$ given $\mathcal{F}_L$ is completely determined by $\Sigma_A^{(L+1)}$. In particular, since g is bounded and continuous, we find that

$$\mathbb{E}\left[g\left(z_A^{(L+1),\bullet}\right)\right] - \mathbb{E}\left[g\left(\widetilde{z}_A^{(L+1)}\right)\right] = \mathbb{E}\left[h\left(\Sigma_A^{(L+1)}\right)\right] - \mathbb{E}\left[h\left(\widetilde{\Sigma}_A^{(L+1)}\right)\right],$$

where $h : \mathbb{R}^{n_{L+1}\times|A|} \to \mathbb{R}$ is a bounded continuous function and $\widetilde{\Sigma}_A^{(L+1)}$ is the conditional covariance matrix at layer $L+1$ for $\widetilde{z}_A^{(L+1)}$. Combining this with the convergence in distribution from (2.18) shows that (2.19) holds and completes the proof of Proposition 2.1 for general weight distributions. □

2.2. *Tightness*: *Proof of Proposition 2.2.* In this section, we provide a proof of Proposition 2.2. In the course of showing tightness, we will need several elementary lemmas, which we record in §2.2.1. We then use them in §2.2.2 to complete the proof of Proposition 2.2.

2.2.1. *Preparatory lemmas.* For the first lemma, let us agree to write $\mathcal{C}(A)$ for the cone over a subset A in a euclidean space and $B_1(\mathbb{R}^n)$ for the unit ball in $\mathbb{R}^n$.

LEMMA 2.6. *Fix integers $n_0, n_1 \geq 1$ and a real number $\lambda \geq 1$. Suppose that T is a compact of $\mathbb{R}^{n_0}$ and $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}$ is λ-Lipschitz with respect to the ℓ_2-norm on both $\mathbb{R}^{n_0}$ and $\mathbb{R}^{n_1}$. Define the Minkowski sums*

$$\widehat{T} = f(T) + \mathcal{C}\big(f(T) - f(T)\big) \cap B_1\big(\mathbb{R}^{n_1}\big).$$

There exists a constant $C > 0$ a compact subset T' of $\mathbb{R}^{3n_0+1}$, and a $C\lambda$-Lipschitz map $g : \mathbb{R}^{3n_0+1} \to \mathbb{R}^{n_1}$ (all depending only $T, \lambda,$), so that $\widehat{T} = g(T')$.

PROOF. By definition,

$$\widehat{T} = g\big(T \times T \times T \times [0,1]\big), \qquad g(x,y,z,r) = f(x) + r\big(f(y) - f(z)\big). \tag{2.20}$$

In particular, for some constant C depending on T_0, λ, we have

$$\begin{aligned}
&\big\|g(x,y,z,r) - g\big(x',y',z',r'\big)\big\|_2 \\
&\quad \le \big\|f(x) - f\big(x'\big)\big\|_2 + \big\|f(y) - f\big(y'\big)\big\|_2 + \big\|f(z) - f\big(z'\big)\big\|_2 \\
&\qquad + |r - r'|\big\|f(y) - f(z)\big\|_2 \\
&\quad \le \lambda\big(\|x - x'\|_2 + \|y - y'\|_2 + \|z - z'\|_2 + \mathrm{Diam}(T_0)|r - r_0|\big) \\
&\quad \le C\lambda\big\|(x - x', y - y', z - z', r - r')\big\|_2.
\end{aligned}$$

Hence, $\widehat{T}$ is the image under a Lipschitz map with a Lipschitz constant depending only on T_0, λ of a compact set in $\mathbb{R}^{3n_0+1}$. □

The second lemma we need is an elementary inequality.

LEMMA 2.7. *Let $a, b, c \ge 0$ be real numbers and $k \ge 1$ be an integer. We have*

$$(a+b+c)^k \le 2^{2k-1}\big(1 + a^{2k}\big) + \frac{1}{4}\big[(2+b)^{4k} + (1+c)^{4k}\big].$$

PROOF. For any $a, b \ge 0$ we have

$$\begin{aligned}
(a+b)^k &= \sum_{j=0}^{k} \binom{k}{j} a^j b^{k-j} \\
&\le \sum_{j=0}^{k} \binom{k}{j} (1+a)^k b^{k-j} = (1+a)^k (1+b)^k \qquad (2.21) \\
&\le \frac{1}{2}\big((1+a)^{2k} + (1+b)^{2k}\big).
\end{aligned}$$

Further, breaking into cases depending on whether $0 \le a \le 1$ or $1 \le a$ we find that

$$(a+b)^k \le 2^{2k-1}\big(1 + a^{2k}\big) + \frac{1}{2}(1+b)^{2k}. \tag{2.22}$$

Combining (2.21) with (2.22) we see as desired that any $a, b, c \ge 0$

$$(a+b+c)^k \le 2^{2k-1}\big(1 + a^{2k}\big) + \frac{1}{4}\big[(2+b)^{4k} + (1+c)^{4k}\big]. \qquad \square$$

The next lemma is also an elementary estimate.

LEMMA 2.8. *Fix an integer $k \ge 1$, and suppose $X_1, \dots, X_k$ are nonnegative random variables. There exists a positive integer $q = q(k)$ such that*

$$\mathbb{E}\left[\prod_{i=1}^{k} X_i\right] \le \left(\prod_{i=1}^{k} \mathbb{E}\big[X_i^q\big]\right)^{1/q}.$$

PROOF. The proof is by induction on k. For the base cases when $k = 1$, we may take $q = 1$ and when $k = 2$ we may take $q = 2$ by Cauchy–Schwartz. Now suppose we have proved the claim for all $k = 1, 2, \dots, K$ for some $K \geq 3$. Note that $1 \leq v\lceil (K+1)/2 \rceil \leq K$. So we may use Cauchy–Schwartz and the inductive hypothesis to obtain

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{K+1} X_i\right] &= \mathbb{E}\left[\prod_{i=1}^{\lceil \frac{K+1}{2} \rceil} X_i \prod_{i=\lceil \frac{K+1}{2} \rceil + 1}^{K} X_i\right] \\
&\leq \mathbb{E}\left[\prod_{i=1}^{\lceil \frac{K+1}{2} \rceil} X_i^2\right]^{1/2} \mathbb{E}\left[\prod_{i=\lceil \frac{K+1}{2} \rceil + 1}^{K} X_i^2\right]^{1/2} \\
&\leq \left(\prod_{i=1}^{K+1} \mathbb{E}[X_i^{2q}]\right)^{1/2q},
\end{aligned}$$

where $q = \max\{q(\lceil \frac{1}{2}(K+1) \rceil), q(K - \lceil \frac{1}{2}(K+1) \rceil - 1)\}$. □

The next Lemma is an elementary result about the moments of marginals of iid random vectors.

LEMMA 2.9. *Fix an even integer $p \geq 2$ a positive integer, and suppose μ is a probability measure on $\mathbb{R}$ with mean 0 and finite higher moments. Assume also that $w = (w_1, \dots, w_n)$ is a vector with iid components, each with distribution μ. Then, there exists a constant C depending only on p and first p moments of μ such that for all $n \geq 1$*

$$\sup_{\|u\|=1} \mathbb{E}[|w \cdot u|^p] \leq C.$$

PROOF. We will use the following result of Łatała [33], Thm. 2, Cor. 2, Rmk. 2. Suppose X_i are independent random variables and p is a positive even integer. Then

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^p\right] \simeq \inf\left\{t > 0 | \sum_{i=1}^{n} \log \mathbb{E}\left[\left|1 + \frac{X_i}{t}\right|^p\right] \leq p\right\}, \tag{2.23}$$

where $\simeq$ means bounded above and below up to universal multiplicative constants. Let us fix a unit vector $u = (u_1, \dots, u_n) \in S^{n-1}$ and apply this to $X_i = u_i w_i$. Since w_i have mean 0 and p is even we find

$$\sum_{i=1}^{n} \log \mathbb{E}\left[\left|1 + \frac{X_i}{t}\right|^p\right] \leq \sum_{i=1}^{n} \log\left(1 + \sum_{k=2}^{p} \binom{p}{k} \frac{|u_i|^k \mathbb{E}[|w_1|^k]}{t^k}\right).$$

Note that for each $k = 2, \dots, p$ we have

$$\mathbb{E}[|w_1|^k] \leq \mathbb{E}[(1 + |w_1|)^p] \text{and} |u_i|^k \leq u_i^2.$$

Hence, using that $\log(1 + x) \leq x$ we find

$$\begin{aligned}
\sum_{i=1}^{n} \log \mathbb{E}\left[\left|1 + \frac{X_i}{t}\right|^p\right] &\leq \sum_{i=1}^{n} \log\left(1 + u_i^2 \mathbb{E}[(1 + |w_1|)^p] \sum_{k=2}^{p} \binom{p}{k} \frac{1}{t^k}\right) \\
&= \sum_{i=1}^{n} \log\left(1 + u_i^2 \mathbb{E}[(1 + |w_1|)^p]\left[\left(1 + \frac{1}{t}\right)^p - 1 - \frac{p}{t}\right]\right) \\
&\leq \mathbb{E}[(1 + |w_1|)^p]\left[\left(1 + \frac{1}{t}\right)^p - 1 - \frac{p}{t}\right].
\end{aligned}$$

Note that for $2 < t$, there is a constant $C > 0$ depending only on p so that

$$\left|\left(1+\frac{1}{t}\right)^p - 1 - \frac{p}{t}\right| \leq \frac{Cp^2}{t^2}.$$

Thus, there exists a constant $C' > 0$ so that

$$t > C'\sqrt{p\mathbb{E}[(1+|w_1|)^p]} \quad \Rightarrow \quad \sum_{i=1}^{n} \log \mathbb{E}\left[\left|1+\frac{X_i}{t}\right|^p\right] \leq p.$$

Combining this with (2.23) completes the proof. □

The final lemma we need is an integrability statement for the supremum of certain non-Gaussian fields over low-dimensional sets.

LEMMA 2.10. *Fix a positive integer n_0, an even integer $k \geq 1$, a compact set $T_0 \subseteq \mathbb{R}^{n_0}$, a constant $\lambda > 0$, and a probability measure μ on $\mathbb{R}$ with mean 0, variance 1, and finite higher moments. For every $\epsilon \in (0,1)$ there exists a constant $C = C(T_0, \epsilon, \lambda, n_0, k, \mu)$ with the following property. Fix any integer $n_1 \geq 1$ and a λ-Lipschitz map $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}$. Define $T_1 := f(T_0)$, and let $w = (w_1, \ldots, w_{n_1})$ be a vector with iid components w_i, each drawn from μ. Then, for any fixed $y_0 \in T_1$*

$$\mathbb{E}\Big[\sup_{y \in T_1} \big(w \cdot (y - y_0)\big)^k\Big] \leq C. \tag{2.24}$$

PROOF. The proof is a standard chaining argument. For each $y \in T_1$ write $\Pi_k(y)$ for the closest point to y in a 2^{-k} net in T_1 and assume without loss of generality that the diameter of T_1 is bounded above by 1 and that $\Pi_0(y) = y_0$ for all $y \in T_1$. We have using the usual chaining trick that

$$\mathbb{E}\Big[\Big(\sup_{y \in T_1} w \cdot (y - y_0)\Big)^k\Big] \leq \sum_{q_1, \ldots, q_k = 0}^{\infty} \mathbb{E}\left[\prod_{i=1}^{k} \sup_{y_i \in T_1} \left|w \cdot \big(\Pi_{q_i}(y_i) - \Pi_{q_i - 1}(y_i)\big)\right|\right]. \tag{2.25}$$

By Lemma 2.8, there exists q depending only on k so that for any $q_1, \ldots, q_k$ we have

$$\begin{aligned} &\mathbb{E}\left[\prod_{i=1}^{k} \sup_{y_i \in T_1} \left|w \cdot \big(\Pi_{q_i}(y_i) - \Pi_{q_i - 1}(y_i)\big)\right|\right] \\ &\quad \leq \prod_{i=1}^{k} \Big(\mathbb{E}\Big[\sup_{y \in T_1} \left|w \cdot \big(\Pi_{q_i}(y) - \Pi_{q_i - 1}(y)\big)\right|^q\Big]\Big)^{1/q}. \end{aligned} \tag{2.26}$$

We seek to bound each expectation on the right hand side in (2.26). To do this, write

$$\mathbb{E}\Big[\sup_{y \in T_1} \left|w \cdot \big(\Pi_{q_i}(y) - \Pi_{q_i - 1}(y)\big)\right|^q\Big] = \int_0^{\infty} \mathbb{P}\Big(\sup_{y \in T_1} \left|w \cdot \big(\Pi_{q_i}(y) - \Pi_{q_i - 1}(y)\big)\right|^q > t\Big)\, dt.$$

Note that the supremum is only over a finite set of cardinality at most

$$|\Pi_{q_i}||\Pi_{q_{i-1}}| \leq 2^{cn_0 q_i}$$

for some $c > 0$ depending only T_0, λ. This is because, by assumption T_1 is the image of T_0 under a λ-Lipschtiz map and Lipschitz maps preserve covering numbers. Thus, by a union bound,

$$\mathbb{P}\Big(\sup_{y \in T_1} \left|w \cdot \big(\Pi_{q_i}(y) - \Pi_{q_i - 1}(y)\big)\right|^q > t\Big) \leq 2^{cn_0 q_i} \sup_{y \in T_1} \mathbb{P}\big(\left|w \cdot \big(\Pi_{q_i}(y) - \Pi_{q_i - 1}(y)\big)\right|^q > t\big).$$

But for any $y \in T_1$ and any $s > 0$, $p \geq 1$ we have

$$\mathbb{P}\big(|w \cdot (\Pi_q(y) - \Pi_{q-1}(y))|^q > s\big) \leq \sup_{\|u\|=1} \mathbb{E}\big[|w \cdot u|^p\big]\bigg(\frac{\|\Pi_k(y) - \Pi_{k-1}(y)\|^p}{s^{p/q}}\bigg)$$
$$= 2^{-pq_i} s^{-p/q} \sup_{\|u\|=1} \mathbb{E}\big[|w \cdot u|^p\big].$$

Putting this all together we find for any $p \geq 2\max\{q+2, cn_0\}$ that

$$\mathbb{E}\Big[\sup_{y \in T_1} |w \cdot (\Pi_{q_i}(y) - \Pi_{q_i-1}(y))|^q\Big] \leq 2^{-cn_0 q_i} \sup_{\|u\|=1} \mathbb{E}\big[|w \cdot u|^p\big].$$

Thus, substituting this into (2.26) yields

$$\mathbb{E}\Big[\Big(\sup_{y \in T_1} w \cdot (y - y_0)\Big)^k\Big] \leq \sup_{\|u\|=1} \mathbb{E}\big[|w \cdot u|^p\big] \sum_{q_1,\ldots,q_k=0}^{\infty} 2^{-cn_0 \sum_{i=1}^k q_i}$$
$$\leq 2^k \sup_{\|u\|=1} \mathbb{E}\big[|w \cdot u|^p\big].$$

Appealing to Lemma 2.9 completes the proof of Lemma 2.10. □

2.2.2. *Proof of Proposition 2.2 using lemmas from §2.2.1.* Let us first establish the equi-Lipschitz condition, which we recall states that for each $\epsilon \in (0,1)$ and each compact set $T \subseteq \mathbb{R}^{n_0}$ there exist $C > 0$ so that with probability at least $1 - \epsilon$ we have

$$\sup_{x_\alpha, x_\beta \in T} \frac{\|z_\alpha^{(L+1)} - z_\beta^{(L+1)}\|_2}{\|x_\alpha - x_\beta\|_2} \leq C. \tag{2.27}$$

For (2.27) to hold, we need a result about the Lipschitz constant of each layer. To ease the notation define a normalized single layer with random weights W and random biases b via the map $\psi : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$:

$$\psi(x; W, b) = \frac{1}{\sqrt{n_2}} \sigma(Wx + b), \tag{2.28}$$

where $b \sim \mathcal{N}(0, C_b \mathrm{I}_{n_2})$ and $W = (w_{ij}) \in \mathbb{R}^{n_2 \times n_1}$ with w_{ij} drawn iid from a distribution with mean 0, variance 1, and finite higher moments. We choose the variance of w_{ij} to be 1 instead of C_W/n_1 since we will later think of x as the normalized vector $(C_W/n_\ell)^{1/2}\sigma(z_\alpha^{(\ell)})$ of post-activations in a given layer.

LEMMA 2.11. *Fix an integer $n_0 \geq 1$, a compact set $T_0 \subseteq \mathbb{R}^{n_0}$, and a constant $\lambda > 0$. For every $\epsilon \in (0,1)$ there exists a constant $C = C(\epsilon, n_0, T_0, \sigma, \lambda)$ with the following property. Fix any integers $n_1, n_2 \geq 1$, and define $\psi : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ as in (2.28). Suppose that $T_1 \subseteq \mathbb{R}^{n_1}$ is the image of T_0 under a λ-Lipschitz map from $\mathbb{R}^{n_0}$ to $\mathbb{R}^{n_1}$. Then,*

$$\mathbb{P}\bigg(\sup_{x_\alpha, x_\beta \in T_1} \frac{\|\psi(x_\alpha) - \psi(x_\beta)\|_2}{\|x_\alpha - x_\beta\|_2} \leq C\bigg) \geq 1 - \epsilon.$$

PROOF. Fix $x_\alpha \neq x_\beta \in T_1$ and define

$$\xi_{\alpha\beta} = \frac{x_\alpha - x_\beta}{\|x_\alpha - x_\beta\|_2}.$$

Write W_i for the ith row of W and b_i for the ith component of b. Since $ab \le \frac{1}{2}(a^2+b^2)$ and σ is absolutely continuous, we have

$$\begin{aligned}\big\|\psi(x_\alpha)-\psi(x_\beta)\big\|_2^2 &= \frac{1}{n_2}\sum_{i=1}^{n_2}\big(\sigma(W_i\cdot x_\alpha+b_i)-\sigma(W_i\cdot x_\alpha+b_i)\big)^2\\ &= \frac{1}{n_2}\sum_{i=1}^{n_2}\Big(W_i\cdot\xi_{\alpha\beta}\int_0^{\|x_\alpha-x_\beta\|_2}\sigma'\big(W_i\cdot(x_\beta+t\xi_{\alpha\beta})+b_i\big)\,dt\Big)^2\\ &\le \|x_\alpha-x_\beta\|_2^2\frac{1}{n_2}\sum_{i=1}^{n_2}\sup_{y\in\widehat{T}}\big(\sigma'(W_i\cdot y+b_i)\big)^2\sup_{\xi\in\widetilde{T}}(W_i\cdot\xi)^2\\ &\le \|x_\alpha-x_\beta\|_2^2\frac{1}{n_2}\sum_{i=1}^{n_2}\Big[\sup_{y\in\widehat{T}}\big(\sigma'(W_i\cdot y+b_i)\big)^4+\sup_{\xi\in\widetilde{T}}(W_i\cdot\xi)^4\Big],\end{aligned}\tag{2.29}$$

where we've set

$$\widetilde{T}=\mathcal{C}(T_1-T_1)\cap S^{n_1-1}\quad\text{and}\quad\widehat{T}:=T_1+\mathcal{C}(T_1-T_1)\cap B_1\big(\mathbb{R}^{n_1}\big),$$

and have denoted by $\mathcal{C}(A)$ the cone over a set A and by $B_1(\mathbb{R}^{n_1})$ the unit ball in $\mathbb{R}^{n_1}$. The estimate (2.29) yields

$$\begin{aligned}&\mathbb{P}\Big(\sup_{x_\alpha,x_\beta\in T}\frac{\|\psi(x_\alpha)-\psi(x_\beta)\|_2^2}{\|x_\alpha-x_\beta\|_2^2}>C\Big)\\ &\quad\le\mathbb{P}\Big(\frac{1}{n_2}\sum_{i=1}^{n_2}\Big[\sup_{y\in\widehat{T}}\big(\sigma'(W_i\cdot y+b_i)\big)^4+\sup_{\xi\in\widetilde{T}}(W_i\cdot\xi)^4\Big]>C\Big).\end{aligned}$$

Since σ' is polynomially bounded by assumption (1.2), we find by Markov's inequality that there exists an even integer $k\ge 2$ so that for any $C>1$

$$\begin{aligned}&\mathbb{P}\Big(\sup_{x_\alpha,x_\beta\in T}\frac{\|\psi(x_\alpha)-\psi(x_\beta)\|^2}{\|x_\alpha-x_\beta\|^2}>C\Big)\\ &\quad\le\frac{1+\mathbb{E}[\sup_{y\in\widehat{T}}|W_1\cdot y+b_1|^k+\sup_{\xi\in\widetilde{T}}|W_1\cdot\xi|^4]}{C-1}.\end{aligned}\tag{2.30}$$

Our goal is now to show that the numerator in (2.30) is bounded above by a constant that depends only on T_0, n_0, λ. For this, let us fix any $y_0\in\widehat{T}$ and apply Lemma 2.7 as follows:

$$\begin{aligned}|W_1\cdot y+b_1|^k&=\big|W_1\cdot(y-y_0)+W_1\cdot y_0+b_1\big|^k\\ &\le 2^{2k-1}\big(1+\big|W_1\cdot(y-y_0)\big|^{2k}\big)+\frac{1}{4}\big[\big(2+|W_1\cdot y_0|^{4k}\big)+\big(1+|b_1|\big)^{4k}\big].\end{aligned}$$

Substituting this and the analogous estimate for $|W\cdot\xi|^4$ into (2.30), we see that since all moments of the entries of the weights and biases exist, there exists a constant $C'>0$ depending on λ, T_0, k so that

$$\begin{aligned}&\mathbb{P}\Big(\sup_{x_\alpha,x_\beta\in T}\frac{\|\psi(x_\alpha)-\psi(x_\beta)\|^2}{\|x_\alpha-x_\beta\|^2}>C\Big)\\ &\quad\le\frac{C'+\mathbb{E}[\sup_{y\in\widehat{T}}|W_1\cdot(y-y_0)|^{2k}+\sup_{\xi\in\widetilde{T}}|W_1\cdot(\xi-\xi_0)|^4]}{C-1},\end{aligned}\tag{2.31}$$

where $\xi_0 \in \widetilde{T}$ is any fixed point. Note that by Lemma 2.6, the sets $\widehat{T}$, $\widetilde{T}$ are both contained in the image of a compact subset $T' \subseteq \mathbb{R}^{3n_0+1}$ under a Lipschitz map, with Lipschitz constant depending only on λ, T. Thus, an application of Lemma 2.6 shows that there exists a constant C'' depending only λ, T, k so that

$$\mathbb{E}\Big[\sup_{y\in\widehat{T}} |W_1 \cdot (y - y_0)|^{2k}\Big] + \mathbb{E}\Big[\sup_{\xi\in\widetilde{T}} |W_1 \cdot (\xi - \xi_0)|^4\Big] \leq C''.$$

Substituting this into (2.31) and taking C sufficiently large completes the proof of Lemma 2.6. □

Lemma 2.11 directly yields the equi-Lipschitz estimate (2.27). Indeed, let us fix $\epsilon \in (0, 1)$ and a compact set $T \subseteq \mathbb{R}^{n_0}$. Let us define

$$h_\alpha^{(\ell)} := \begin{cases} \sqrt{\dfrac{C_W}{n_0}} x_\alpha & \ell = 0, \\ \sqrt{\dfrac{C_W}{n_\ell}} \sigma\big(z_\alpha^{(\ell)}\big) & \ell = 1, \dots, L, \\ \dfrac{1}{\sqrt{n_{L+1}}} z_\alpha^{(L+1)} & \ell = L+1. \end{cases}$$

For $\ell = 1, \dots, L+1$ we have

$$h_\alpha^{(\ell)} = \begin{cases} \sqrt{\dfrac{C_W}{n_\ell}} \sigma\big(\widehat{W}^{(\ell)} h_\alpha^{(\ell-1)} + b^{(\ell)}\big) & \ell = 1, \dots, L, \\ \dfrac{1}{\sqrt{n_{L+1}}} \big(\widehat{W}^{(L+1)} h_\alpha^{(L)} + b^{(L+1)}\big) & \ell = L+1, \end{cases}$$

where the rescaled weight matrices $\widehat{W}^{(\ell)}$, defined in (1.4), has iid mean 0 variance 1 entries with finite higher moments. To each of the transformations $h_\alpha^{(\ell)} \mapsto h_\alpha^{(\ell+1)}$ we may now apply Lemma 2.11. Specifically, applying Lemma 2.11 in the first layer shows that there exists $C^{(1)} > 0$ so that the rescaled first layer map

$$\sup_{x_\alpha, x_\beta \in T} \frac{\|h_\alpha^{(1)} - h_\beta^{(1)}\|}{\|x_\alpha - x_\beta\|} \leq C^{(1)}$$

with probability at least $1 - \epsilon/(L+1)$. Thus, the image

$$T^{(1)} := h^{(1)}(T)$$

of T under the normalized first layer map is the image under a $C^{(1)}$-Lipschitz map of the compact set $T \subseteq \mathbb{R}^{n_0}$. This allows us to apply Lemma 2.11 again, but this time to the second layer, to conclude that, again, there exist $C^{(2)} > 0$ so that with probability at least $1 - 2\epsilon/(L+1)$ the normalized second layer map satisfies

$$\sup_{x_\alpha, x_\beta \in T} \frac{\|h_\alpha^{(2)} - h_\beta^{(2)}\|}{\|x_\alpha - x_\beta\|} \leq \sup_{x_\alpha, x_\beta \in T} \frac{\|h_\alpha^{(2)} - h_\beta^{(2)}\|}{\|h_\alpha^{(1)} - h_\beta^{(1)}\|} \sup_{x_\alpha, x_\beta \in T} \frac{\|h_\alpha^{(1)} - h_\beta^{(1)}\|}{\|x_\alpha - x_\beta\|} \leq C^{(1)} C^{(2)}.$$

Proceeding in this way, with probability at least $1 - \epsilon$ we that

$$\sup_{x_\alpha, x_\beta \in T} \frac{\|z_\alpha^{(L+1)} - z_\beta^{(L+1)}\|}{\|x_\alpha - x_\beta\|} = n_{L+1}^{1/2} \sup_{x_\alpha, x_\beta \in K} \frac{\|h_\alpha^{(L+1)} - h_\beta^{(L+1)}\|}{\|x_\alpha - x_\beta\|} \leq n_{L+1}^{1/2} C^{(1)} \cdots C^{(L+1)}.$$

Since n_{L+1} is fixed and finite, this confirms (2.27). It remains to check the uniform boundedness condition in (2.1). For this note that for any fixed $x_\beta \in K$ by Lemma 2.4, we have

$$\sup_{n_1,\dots,n_L \geq 1} \mathbb{E}\left[\frac{1}{n_{L+1}} \|z_\beta^{(L+1)}\|^2\right] < \infty.$$

Thus, by Markov's inequality, $\|z_\beta^{(L+1)}\|$ is bounded above with high probability. Combined with the equi-Lipschitz condition (2.27), which we just saw holds with high probability on K, we conclude that for each $\epsilon > 0$ there exists $C > 0$ so that

$$\mathbb{P}\Big(\sup_{x_\alpha \in K} \|z_\alpha^{(L+1)}\| \leq C\Big) \geq 1 - \epsilon,$$

as desired.

Funding. The author was supported by an NSF CAREER grant DMS-2143754 and NSF grants DMS-1855684, DMS-2133806. He was also a consultant for an ONR MURI on Foundations of Deep Learning.

REFERENCES

[1] ADLAM, B., LEVINSON, J. and PENNINGTON, J. (2022). A random matrix perspective on mixtures of nonlinearities for deep learning. *AISTATS*.

[2] ADLAM, B. and PENNINGTON, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning* 74–84. PMLR.

[3] AHN, A. (2022). Fluctuations of β-Jacobi product processes. *Probab. Theory Related Fields* **183** 57–123. MR4421171 https://doi.org/10.1007/s00440-022-01109-0

[4] AKEMANN, G. and BURDA, Z. (2012). Universal microscopic correlation functions for products of independent Ginibre matrices. *J. Phys. A* **45** 465201. MR2993423 https://doi.org/10.1088/1751-8113/45/46/465201

[5] AKEMANN, G., BURDA, Z. and KIEBURG, M. (2019). From integrable to chaotic systems: Universal local statistics of Lyapunov exponents. *Europhys. Lett.* **126** 40001.

[6] AKEMANN, G., BURDA, Z., KIEBURG, M. and NAGAO, T. (2014). Universal microscopic correlation functions for products of truncated unitary matrices. *J. Phys. A* **47** 255202. MR3224113 https://doi.org/10.1088/1751-8113/47/25/255202

[7] BARTLETT, P. L., LONG, P. M., LUGOSI, G. and TSIGLER, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA* **117** 30063–30070. MR4263288 https://doi.org/10.1073/pnas.1907378117

[8] BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854. MR3997901 https://doi.org/10.1073/pnas.1903070116

[9] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G. et al. (2020). Language models are few-shot learners. ArXiv Preprint. Available at arXiv:2005.14165.

[10] CRISANTI, A., PALADIN, G. and VULPIANI, A. (2012). *Products of Random Matrices*: *In Statistical Physics* **104**. Springer, Berlin.

[11] DANIELY, A., FROSTIG, R. and SINGER, Y. (2016). Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems* 2253–2261.

[12] DAUBECHIES, I., DEVORE, R., FOUCART, S., HANIN, B. and PETROVA, G. (2022). Nonlinear approximation and (deep) ReLU networks. *Constr. Approx.* **55** 127–172. MR4376561 https://doi.org/10.1007/s00365-021-09548-z

[13] DEVORE, R., HANIN, B. and PETROVA, G. (2021). Neural network approximation. *Acta Numer.* **30** 327–444. MR4298220 https://doi.org/10.1017/S0962492921000052

[14] DU, S. S., ZHAI, X., POCZOS, B. and SINGH, A. (2019). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

[15] ELDAN, R., MIKULINCER, D. and SCHRAMM, T. (2021). Non-asymptotic approximations of neural networks by Gaussian processes. In *Conference on Learning Theory* 1754–1775. PMLR.

[16] FAN, Z. and WANG, Z. (2020). Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Adv. Neural Inf. Process. Syst.* **33** 7710–7721.
[17] FURSTENBERG, H. (1963). Noncommuting random products. *Trans. Amer. Math. Soc.* **108** 377–428. MR0163345 https://doi.org/10.2307/1993589
[18] FURSTENBERG, H. and KESTEN, H. (1960). Products of random matrices. *Ann. Math. Stat.* **31** 457–469. MR0121828 https://doi.org/10.1214/aoms/1177705909
[19] GARRIGA-ALONSO, A. and RASMUSSEN, CE. and AITCHISON, L. (2021). *Deep Convolutional Networks as Shallow Gaussian Processes*. *International Conference on Representation Learning* 2021.
[20] GORIN, V. and SUN, Y. (2022). Gaussian fluctuations for products of random matrices. *Amer. J. Math.* **144** 287–393. MR4401507 https://doi.org/10.1353/ajm.2022.0006
[21] HANIN, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*.
[22] HANIN, B. (2019). Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics* **7** 992.
[23] HANIN, B. and NICA, M. (2019). Finite depth and width corrections to the neural tangent kernel. ICLR 2020 and available at arXiv:1909.05989.
[24] HANIN, B. and NICA, M. (2020). Products of many large random matrices and gradients in deep neural networks. *Comm. Math. Phys.* **376** 287–322. MR4093863 https://doi.org/10.1007/s00220-019-03624-z
[25] HANIN, B. and PAOURIS, G. (2021). Non-asymptotic results for singular values of Gaussian matrix products. *Geom. Funct. Anal.* **31** 268–324. MR4268303 https://doi.org/10.1007/s00039-021-00560-w
[26] HANIN, B. and ROLNICK, D. (2019). Deep ReLU networks have surprisingly few activation patterns. *NeurIPS*.
[27] HANIN, B. and ROLNICK, D. (2019). Complexity of linear regions in deep networks. *ICML*.
[28] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** 949–986. MR4404925 https://doi.org/10.1214/21-aos2133
[29] HEBB, D. O. (1949). *The Organization of Behavior*: *A Neuropsychological Theory*. Wiley, New York.
[30] HUANG, J. and YAU, H.-T. (2020). Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning* 4542–4551. PMLR.
[31] JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems* 8571–8580.
[32] KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.
[33] LATALA, R. (1997). Estimation of moments of sums of independent real random variables. *Ann. Probab.* **25** 1502–1513.
[34] LEE, J., BAHRI, Y., NOVAK, R., SCHOENHOLZ, S. S., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2018). Deep neural networks as Gaussian processes. ICML 2018 and available at arXiv:1711.00165.
[35] LIU, C., ZHU, L. and BELKIN, M. (2020). On the linearity of large non-linear models: When and why the tangent kernel is constant. *Adv. Neural Inf. Process. Syst.* **33** 15954–15964.
[36] MATTHEWS, A. G. D. G., ROWLAND, M., HRON, J., TURNER, R. E. and GHAHRAMANI, Z. (2018). Gaussian process behaviour in wide deep neural networks. ArXiv Preprint. Available at arXiv:1804.11271.
[37] NEAL, R. M. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks* 29–53. Springer, Berlin.
[38] NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. *London Mathematical Society Lecture Note Series* **335**. Cambridge Univ. Press, Cambridge. MR2266879 https://doi.org/10.1017/CBO9780511735127
[39] NOCI, L., BACHMANN, G., ROTH, K., NOWOZIN, S. and HOFMANN, T. (2021). Precise characterization of the prior predictive distribution of deep ReLU networks. *Adv. Neural Inf. Process. Syst.* **34** 20851–20862.
[40] NOVAK, R., XIAO, L., LEE, J., BAHRI, Y., YANG, G., HRON, J., ABOLAFIA, D. A., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2018). Bayesian deep convolutional networks with many channels are Gaussian processes. ArXiv Preprint. Available at arXiv:1810.05148.
[41] PÉCHÉ, S. (2019). A note on the Pennington–Worah distribution. *Electron. Commun. Probab.* **24** 66. MR4029435 https://doi.org/10.1214/19-ecp262
[42] PENNINGTON, J. and WORAH, P. (2017). Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems* 2634–2643.
[43] ROBERTS, D. A., YAIDA, S. and HANIN, B. (2022). *The Principles of Deep Learning Theory*: *An Effective Theory Approach to Understanding Neural Networks*. Cambridge Univ. Press, Cambridge.

[44] ROSENBLATT, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65** 386.
[45] RUELLE, D. (1979). Ergodic theory of differentiable dynamical systems. *Publ. Math. Inst. Hautes étud. Sci.* **50** 27–58.
[46] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V. et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* **529** 484–489.
[47] VOICULESCU, D. (1986). Addition of certain noncommuting random variables. *J. Funct. Anal.* **66** 323–346. MR0839105 https://doi.org/10.1016/0022-1236(86)90062-5
[48] WIGNER, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. of Math.* (2) **67** 325–327. MR0095527 https://doi.org/10.2307/1970008
[49] YAIDA, S. (2020). Non-Gaussian processes and neural networks at finite widths. *MSML*.
[50] YANG, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. ArXiv Preprint. Available at arXiv:1902.04760.
[51] YANG, G. (2019). Tensor programs I: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. ArXiv Preprint. Available at arXiv:1910.12478.
[52] YANG, G. (2020). Tensor programs II: Neural tangent kernel for any architecture. ArXiv Preprint. Available at arXiv:2006.14548.
[53] YANG, G. (2020). Tensor programs III: Neural matrix laws. ArXiv Preprint. Available at arXiv:2009.10685.
[54] YAROTSKY, D. (2016). Error bounds for approximations with deep ReLU networks. ArXiv Preprint. Available at arXiv:1610.01145.
[55] YAROTSKY, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory* 639–649. PMLR.
[56] ZAVATONE-VETH, J. and PEHLEVAN, C. (2021). Exact marginal prior distributions of finite Bayesian neural networks. *Adv. Neural Inf. Process. Syst.* **34**.
[57] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, (*ICLR*).