Modern Stochastics: Theory and Applications 11 (1) (2024) 85–108
https://doi.org/10.15559/23-VMSTA238

A quantitative functional central limit theorem for shallow neural networks

Valentina Cammarota[a], **Domenico Marinucci**[b], **Michele Salvi**[b,*],
Stefano Vigogna[b]

[a]*Department of Statistics, University of Rome La Sapienza, Piazzale Aldo Moro 5, 00185, Italy*

[b]*Department of Mathematics, University of Rome Tor Vergata, Via della Ricerca Scientifica 1, 00133, Italy*

valentina.cammarota@uniroma1.it (V. Cammarota), marinucc@mat.uniroma2.it (D. Marinucci),
salvi@mat.uniroma2.it (M. Salvi), vigogna@mat.uniroma2.it (S. Vigogna)

Received: 5 July 2023, Revised: 23 October 2023, Accepted: 15 November 2023,
Published online: 28 November 2023

Abstract We prove a quantitative functional central limit theorem for one-hidden-layer neural networks with generic activation function. Our rates of convergence depend heavily on the smoothness of the activation function, and they range from logarithmic for nondifferentiable nonlinearities such as the ReLu to $\sqrt{n}$ for highly regular activations. Our main tools are based on functional versions of the Stein–Malliavin method; in particular, we rely on a quantitative functional central limit theorem which has been recently established by Bourguin and Campese [Electron. J. Probab. 25 (2020), 150].

Keywords Quantitative functional central limit theorem, Wiener-chaos expansions, neural networks, Gaussian processes

2010 MSC 60F17, 68T07, 60G60

[*] Corresponding author.

© 2024 The Author(s). Published by VTeX. Open access article under the CC BY license.

1 Introduction and background

In this paper we shall be concerned with one-hidden-layer neural networks with Gaussian random weights, that is, random fields $F : \mathbb{S}^{d-1} \to \mathbb{R}$ of the form

$$F(x) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} V_j \sigma\left(\sum_{\ell=1}^{d} W_{j\ell} x_\ell \right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} V_j \sigma(W_j x), \tag{1}$$

where $V_j \in \mathbb{R}$, $W_j \in \mathbb{R}^{1\times d}$ are, respectively, random variables and vectors, whose entries are independent Gaussian with zero mean and variance $\mathbb{E}[V_j^2] = \mathbb{E}[W_{j\ell}^2] = 1$, $j = 1, \dots, n$, $\ell = 1, \dots, d$. Here $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function whose properties and form we will discuss below, the $\sigma(W_j x)$ represent the artificial neurons, and n is their number, namely the width of the network. The random field F is defined on the unit sphere $\mathbb{S}^{d-1}$, with zero mean and covariance function

$$S(x_1, x_2) := \mathbb{E}\big[F(x_1)F(x_2)\big] = \mathbb{E}\big[\sigma(W_j x_1)\sigma(W_j x_2)\big], \quad x_1, x_2 \in \mathbb{S}^{d-1}. \tag{2}$$

The covariance function S also defines a zero mean Gaussian random field $Z : \mathbb{S}^{d-1} \to \mathbb{R}$, which gives the asymptotic distribution of F for $n \to \infty$ [23].

Our aim is to establish a quantitative functional central limit theorem for the network F as the number of neurons n increases, that is, to study the distance, under a suitable functional probability metric d_2, between F and Z as a function of n. In particular, we shall obtain bounds of the form

$$d_2(F, Z) \le b(n, \alpha),$$

where $\lim_{n\to\infty} b(n, \alpha) \to 0$ and α is a parameter capturing the smoothness of the activation σ. For the sake of brevity and simplicity, in this paper we restrict our attention to univariate neural networks; the extension to the multivariate case can be obtained along similar lines, up to a factor depending on the output dimension.

The distribution of neural networks in the large-width limit is a classical topic in learning theory, the first result going back to the seminal work [23]. The subject has gained considerable attention in the machine learning community as it can shed light on the network training process and draw links with kernel-based learning. Neural networks are usually optimized by (variants of) gradient descent with a random initial condition, hence, at initialization, they can be seen as random fields. In many applications, the Gaussian is in fact the distribution of choice, which leads (in the shallow case) to the model considered in (1). On the other hand, taking large, over-parametrized architectures has become an established practice, achieving impressive empirical performance in spite of classical statistical knowledge that would warn from the risks of overfitting [30, 4]. For these reasons – (Gaussian) random initialization and over-parametrization – central limit theorems of neural networks to infinite width provide useful information on the distribution of a typical network at the beginning of its training. In particular, the Gaussian limit reveals random neural networks as approximations of kernel methods associated with random-features kernels, that is, kernels of the form (2) [27, 2]. Interestingly, some of this information may even carry over beyond initialization. Indeed, it has been observed that, under a proper scaling limit, the evolution of the network through training is well approximated by its linearization around the initial condition, and it is governed by a kernel, called neural tangent kernel, which adds higher-order correlations to the random features (2) [18, 9]. In this lazy regime, the weights do not move too much from their random initialization, and thus the network from its central limit.

As we already recalled, the story of central limit theorems for neural networks starts with [23], which gave the proof for a single hidden layer. This first result was later generalized to deep networks, leading to an extensive picture in [15]. Non-Gaussianity at finite-width perturbations have been investigated studying higher-order cumulants in [28, 29]. Quantitative central limit theorems in suitable probability metrics have been considered only very recently in [3] and [6]. In [3] the authors have proved a finite-dimensional quantitative central limit theorem for neural networks of finite depth whose activation functions satisfy a Lipschitz condition; in [6], the authors have proved second-order Poincaré inequalities (which imply one-dimensional quantitative central limit theorems) for neural networks with C^2 activation functions.

Understanding the Gaussian behavior of a neural network allows, for instance, to investigate the geometry of its landscape, e.g., the cardinality of its minima, the number of nodal components and many other quantities of interest. However, convergence of the finite-dimensional distributions is in general not sufficient to constraint such landscapes. For this reason, functional results, that is, bounds on the speed of convergence in functional spaces, are also of great interest. So far, the literature on quantitative functional central limit theorems is still limited: [13] and [19] have focused on one-hidden-layer networks, where the random coefficients in the inner layer are Gaussian for [13] and uniform on the sphere for [19], whereas the coefficients in the outer layer follow a Rademacher distribution for both. In particular, the authors in [13] manage to establish rates of convergence in Wasserstein distance which are (power of) logarithmic for ReLu and other activation functions, and algebraic for polynomial or very smooth activations, see below for more details. On the other hand, the rates in [19] for ReLu networks are of the form $O(n^{-\frac{1}{2d-1}})$; this is algebraic for fixed values of d, but it can actually converge to zero more slowly than the inverse of a logarithm if d is of the same order as n, as it is the case for many applications.

1.1 Purpose and plan of the paper

We consider in this work functional quantitative central limit theorems under general activations and for coefficients that are Gaussian for both layers, which seems the most relevant case for applications; our approach is largely based upon very recent results by [8] on the Stein–Malliavin techniques for random elements taking values in Hilbert spaces (we refer to [24, 25] for the general foundations of this approach, together with [20, 7, 1, 12] for some more recent references). Our main results are collected in Section 2, whereas their proofs with a few technical lemmas are given in Section 4. A short comparison with the existing literature is provided in Section 3. Appendix A is mainly devoted to background results which we heavily exploit throughout the paper.

Notation. Hereafter, we will write $a_n \sim b_n$ for two positive sequences such that $\lim_{n\to\infty} a_n/b_n = 1$. The expression $A \lesssim B$ means that $A \leq CB$ for some absolute constant $C > 0$. We will denote by $\|\cdot\|$ the L^2 norm corresponding to the uniform probability measure on the unit sphere $\mathbb{S}^{d-1}$.

2 Main results

In order to state our main theorems, we shall need some further assumptions and notations. We shall always be concerned with activation functions which are square integrable with respect to the standard Gaussian measure, i.e., such that

$$\mathbb{E}\big[\sigma^2(\zeta)\big] < \infty, \ \zeta \sim N(0,1);$$

this is a truly minimal conditions, which is guaranteed by $\sigma(z) = O(\exp(z^2/(2+\delta))$ for all $\delta > 0$. For such activation functions, it is well known that the following Hermite expansion holds, in the L^2 sense with respect to the Gaussian measure (see, e.g., [25]):

$$\sigma(x) = \sum_{q=0}^{\infty} J_q(\sigma)\frac{H_q(x)}{\sqrt{q!}}, \quad \text{with } H_q(x) := (-1)^q e^{\frac{x^2}{2}} \frac{d^q}{dx^q} e^{-\frac{x^2}{2}},$$

where $\{H_q\}_{q=0,1,2,\ldots,}$ is the well-known sequence of Hermite polynomials. The coefficients $J_q(\sigma)$, which will play a crucial role in our arguments below, are defined according to the following (normalized) projection:

$$J_q(\sigma) := \frac{1}{\sqrt{q!}}\mathbb{E}\big[\sigma(\zeta)H_q(\zeta)\big].$$

In the following, when no confusion is possible, we may drop the dependence of J on σ for ease of notation. We remark that our notation is to some extent non-standard, insofar we have introduced the factor $\frac{1}{\sqrt{q!}}$ inside the projection coefficient $\mathbb{E}[\sigma(\zeta)H_q(\zeta)]$; equivalently, we are defining the projection coefficients in terms of Hermite polynomials which have been normalized to have unit variance. Indeed, it is well known that

$$\mathbb{E}\left[\left(\frac{H_q(\zeta)}{\sqrt{q!}}\right)^2\right] = \frac{1}{q!}\mathbb{E}\big[\big(H_q(\zeta)\big)^2\big] = 1.$$

In short, our main results state that a quantitative functional central limit theorem for neural networks built on σ holds, and the rate of convergence depends on the rate of decay of $\{J_q(\sigma)\}$, as $q \to \infty$; roughly put, it is logarithmic when this rate is polynomial (e.g., the ReLu case), whereas convergence occurs at algebraic rates for some activation functions which are smoother, with exponential decay of the coefficients. A more detailed discussion of these results and comparisons with the existing literature are given below in Section 3.

Let us discuss an important point about normalization. *In this paper, the measure on the sphere* $\mathbb{S}^{d-1}$ *is normalized to have unit volume*. The bound we obtain are not invariant to this normalization, and indeed they would be much tighter if the measure on the sphere was taken as usual to be $s_d = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}$, the surface volume of $\mathbb{S}^{d-1}$. Indeed, by Stirling's formula

$$s_d = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \sim \frac{2\pi^{d/2}2^{d/2}e^{d/2}}{\sqrt{\pi d}d^{d/2}} = \frac{2}{\sqrt{\pi d}}\left(\sqrt{\frac{2e\pi}{d}}\right)^d;$$

s_d achieves its maximum for $d = 7$ ($s_7 = 33.073$) and decays faster than exponentially as $d \to \infty$. This means that, without the normalization that we chose, our bound on the d_2 metric would be actually smaller by a factor of roughly $d^{-d/2}$ when the dimension grows. On the other hand, if we were to take the standard Lebesgue measure λ then we would obtain, by a standard application of Hermite expansions and the Diagram Formula

$$\mathbb{E}\|F\|^2_{L^2(\lambda)} = \sum_q J_q^2(\sigma) \int_{\mathbb{S}^{d-1}} \lambda(dx) = \sum_q J_q^2(\sigma) s_d,$$

so that the L^2 norm would decay very quickly as d increases, making the interpretation of results less transparent.

Following [8], the convergence in our central limit theorem is measured in the d_2 metric. This is given by

$$d_2(F, Z) = \sup_{\|h\|_{C_b^2(L^2(\mathbb{S}^{d-1}))} \leq 1} \left|\mathbb{E}h(F) - \mathbb{E}h(Z)\right|,$$

where $C_b^2(L^2(\mathbb{S}^{d-1}))$ is the space of real-valued functions on $L^2(\mathbb{S}^{d-1})$ (where L^2 is taken with respect to the uniform measure) with bounded Frechet derivatives up to order 2. It is to be noted that the d_2 metric is bounded by the Wasserstein distance of order 2, i.e.

$$d_2(F, Z) \leq \mathcal{W}_2(F, Z) := \inf_{(\widetilde{F},\widetilde{Z})} \left(\mathbb{E}\|\widetilde{F} - \widetilde{Z}\|^2_{L^2(\mathbb{S}^{d-1})}\right)^{1/2},$$

where the infimum is taken over all the possible couplings of (F, Z).

Our first main statement is as follows.

Theorem 1. *Under the previous assumptions and notations, and letting Z be the Gaussian process with zero mean and same covariance as F, we have that, for all $Q \leq \log_3 \sqrt{n}$,*

$$d_2(F, Z) \leq C\|\sigma\| \frac{1}{\sqrt[4]{n}} \sqrt{\sum_{q=0}^{Q} J_q^2(\sigma) q 3^q} + \frac{3}{2}\sqrt{\sum_{q=Q+1}^{\infty} J_q^2(\sigma)}, \tag{3}$$

where C is an absolute constant (in particular, independent of the input dimension d), and $\|\sigma\|$ is the L^2 norm of σ taken with respect to the Gaussian density on $\mathbb{R}$.

The proof is postponed to Section 4.1. From Theorem 1, optimizing over the choice of Q, it is immediate to obtain much more explicit bounds. In the case of polynomial decay of the Hermite coefficients, the choice $Q = \log n/(3 \log 3)$ yields the following result.

Corollary 2. *In the same setting as in Theorem 1, for $J_q(\sigma) \lesssim q^{-\alpha}$, $\alpha > \frac{1}{2}$, we have*

$$d_2(F, Z) \leq C\|\sigma\| \frac{1}{(\log n)^{\alpha - \frac{1}{2}}}.$$

Example 3 (ReLu)**.** As shown in Lemma 19, for the ReLu activation $\sigma(t) = t\mathbb{I}_{[0,\infty)}(t)$ we have that $J_q(\sigma) \lesssim q^{-\frac{5}{4}}$, whence we obtain the bound $d_2(F, Z) \lesssim (\log n)^{-\frac{3}{2}}$. Once again, we stress that the constant is independent of the input dimension d.

The statement of Theorem 1 is given in order to cover the most general activation functions, allowing for possibly nondifferentiable choices such as the ReLu. Under stronger conditions, the result can be improved; in particular, assuming the activation function has a Malliavin derivative with bounded fourth moment (i.e., it belongs to the class $\mathbb{D}^{1,4}$, see [25, 8]), we obtain the following extension.

Theorem 4. *Under the previous assumptions and notations, and assuming furthermore that $\sigma(Wx) \in \mathbb{D}^{1,4}$, we have that, for all $Q \in \mathbb{N}$,*

$$d_2(F, Z) \leq C \frac{1}{\sqrt{n}} \sum_{q=0}^{Q} J_q^2(\sigma) q 3^q \left(\|\sigma\|^2 + \frac{1}{\sqrt{n}} \sum_{q=0}^{Q} J_q^2(\sigma) 3^q \right) + \frac{3}{2} \sqrt{\sum_{q=Q+1}^{\infty} J_q^2(\sigma)}, \tag{4}$$

where C is an absolute constant (in particular, independend of the input dimension d), and $\|\sigma\|$ is the L^2 norm of σ taken with respect to the Gaussian density on $\mathbb{R}$.

We prove Theorem 4 in Section 4.4. Again, imposing specific decay profiles on the Hermite expansion we can obtain explicit bounds. In particular, when $J_q \lesssim e^{-\beta q}$ with $\beta > \log \sqrt{3}$, the second sum appearing in (4) stays finite for all Q, hence the bound assumes the form

$$d_2(F, Z) \leq C\|\sigma\|^2 \frac{1}{\sqrt{n}} \sum_{q=0}^{Q} J_q^2(\sigma) q 3^q + \frac{3}{2} \sqrt{\sum_{q=Q+1}^{\infty} J_q^2(\sigma)},$$

which is more in line with the bound (3). In such a case, letting Q to go to infinity leads to the next result.

Corollary 5. *In the same setting as in Theorem 4, for $J_q(\sigma) \lesssim e^{-\beta q}$, $\beta > \log \sqrt{3}$, we have*

$$d_2(F, Z) \leq C \frac{1}{\sqrt{n}}.$$

Example 6 (polynomials/erf)**.** The assumptions of Corollary 5 are fulfilled by polynomial activations and by the error function $\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2} ds$, for which $J_q^2(\sigma) \lesssim (2/3)^q$ – cf. [19]. In these cases, the fact that $\sigma(Wx) \in \mathbb{D}^{1,4}$ can be readily shown by means of the triangle inequality and the standard hypercontractivity bound for Wiener chaos components – see [25, Corollary 2.8.14].

Example 7 (tanh/logistic)**.** Of course, other forms of decay could be considered. For instance, for the hyperbolic tangent $\sigma(t) = (e^t - e^{-t})/(e^t + e^{-t})$ the rate of decay of the Hermite coefficients is of order $\exp(-C\sqrt{q})$ (see, e.g., [13]), hence the result of Corollary 5 does not apply; the bounds in Corollary 2 obviously hold, but applying directly Theorem 1 and some algebra we obtain the finer bound

$$d_2(F, Z) \lesssim \exp(-c\sqrt{\log n}), \quad \text{for } J_q(\sigma) \leq \exp(-C\sqrt{q}).$$

The same bound holds also for the sigmoid/logistic activation function $\sigma(t) = (1 + e^{-t})^{-1}$.

Remark 8. Lower bounds on the rates of convergence of neural networks to Gaussian processes are still an open question. In particular, we do not know whether the rates obtained in Corollary 2 and Corollary 5 are optimal. In Section 3 we compare our results to the previous literature.

2.1 Sketch of the proof and discussion

As a first step in our proof, cf. Section 4.1, we decompose our neural network F into two processes: $F_{\leq Q}$, corresponding to its projection onto the first Q Wiener chaoses, and $F_{>Q}$, the remainder, where Q is an integer to be chosen below. This truncation-and-optimization approach is rather standard in the literature on Quantitative Central Limit Theorems, cf. [8, Remark 3.11]. By the triangle inequality, we can bound the distance of F from a suitable Gaussian process Z with the distance of Z from $F_{\leq Q}$ plus the 2-Wasserstein distance of F from $F_{\leq Q}$. This second part can be easily bounded by standard L^2 arguments, see (5).

For the leading term we follow a recent result by Bourguin and Campese [8], which we restate as Theorem 16, adapted to our framework. To the best of our knowledge, this is the first time when a link between the Stein–Malliavin method (see [25, 8] and the references therein) and neural networks has been established.

Thanks to this technique, the problem can be essentially reduced to a thorough analysis of fourth-order cumulants and covariances for the L^2 norms of the Wiener projections. Besides smaller order terms, we dominate the distance between Z and $F_{\leq Q}$ with the sum of two terms, called M and C. Heuristically, M controls the expected distance of the fourth moments of the Wiener projections of Z and $F_{\leq Q}$, while C accounts for the covariances between different projections of $F_{\leq Q}$. In order to control M, in Proposition 9 we exploit the properties of Hermite polynomials and in particular the diagram formula (see [22, Proposition 4.15]). A detailed analysis of the possible configurations of the diagrams (Lemma 10 and Lemma 11) and of the covariances of the Hermite polynomials (Lemma 12) allows us to obtain bounds that, in particular, do not depend on the dimension d of the input, cf. Remark 13 and the discussion in Section 3. Finally, in Proposition 15 we show that C is bounded from above by M itself.

We point out that the strategy we follow relies on the Gaussianity of the distribution of the weights W. While nonquantitative versions of the CLT have been proved assuming only mild finite moment assumptions, see [15], Gaussianity has been required in the literature for the quantitative case so far (cf. [13, 19]). Our main technical tool, [8, Theorem 3.10], goes beyond the Gaussian case and Hermite polynomials, but for more general eigenfunctions no diagram formula is known and explicit computations (e.g., for estimating the cumulants) become impossible.

Another technical point that we shall address is the following. The convergence results by [8] require the limiting process to be nondegenerate; this condition is not always satisfied for arbitrary activation functions if one takes the corresponding Hilbert space to be $L^2(\mathbb{S}^{d-1})$ (counter-examples being finite-order polynomials). However, we note that for activations for which the corresponding networks are dense in the space of continuous functions (such as the ReLu or the sigmoid and basically all nonpolynomials, see for instance the classical universal approximation theorems in [10, 16, 17, 21, 26]), then the nondegeneracy condition is automatically satisfied. On the other hand, when the condition fails, our results continue to hold, but the underlying functional space must be taken to be the reproducing kernel Hilbert space generated by the covariance operator, which is strictly included into $L^2(\mathbb{S}^{d-1})$ when universal approximation fails (e.g., in the polynomial case).

3 A comparison with the existing literature

Two papers that have established quantitative functional central limit theorems for neural networks are [13] and [19]. Their settings and results are not entirely comparable to ours; on the one hand, they use the Wasserstein distance, which is slightly stronger that the d_2 metric we consider here. On the other hand, their model for the random weight is different from ours: for the outer layers, both consider Rademacher variables, while for the inner layer the distribution is Gaussian in [13] and uniform on the sphere in [19]; on the contrary, we assume the Gaussian distribution for both inner and outer layer. As a further (minor) difference, we note that in [13], as well as in our paper, input variables are in $\mathbb{S}^{d-1}$, while [19] considers $\sqrt{d}\ \mathbb{S}^{d-1}$; this is just a notational issue, though, because in [19] the argument of the activation function is normalized by a factor $1/\sqrt{d}$.

Even with these important caveats, it is nevertheless of some interest to compare their bounds with ours, for activation functions for which there is an overlap. We report their results together with ours in Table 1 (the constant C may differ from one box to the other, but in all cases it does not depend neither on d nor on n).

Comparing to [13], our bounds remove a logarithmic factor in the input dimension and a log log factor in the number of neurons for ReLu and tanh networks; for smooth activations, the rate goes from $n^{-1/6}$ to $n^{-1/2}$, and the constants lose the polynomial dependence on the dimension. The rate in [19] in the polynomial case is $n^{-1/2}$ as ours, but with a factor growing in the input dimension d as $d^{d/2}$. In the ReLu setting, [19] displays the algebraic rate $n^{-\frac{3}{4d-2}}$, which *for fixed values of* d decays faster than our logarithmic bound. However, interpretation of these bounds from a “fixed d, growing n” perspective can be incomplete: when considering distances in probability metrics it is of interest to allow both d and n to vary. In particular, for neural networks applications, it is often the case that the input dimension and number of neurons are of comparable order; taking for instance $d = d_n \sim n^{\alpha}$, it is immediate to verify that

Table 1. Comparison of convergence rates established by different functional quantitative central limit theorems for several activation functions. Bear in mind that two different metrics $d_2 \leq \mathcal{W}_2$ are considered, $\mathcal{W}_2$ for [13, 19], and d_2 for this paper. The parameters α and β must satisfy $\alpha > 1/2$ and $\beta > \log\sqrt{3}$

	Eldan et al. [13]	Klukowski [19]	This paper
$J_q \sim q^{-\alpha}$	$(\frac{\log n}{\log\log n \log d})^{-\alpha+\frac{1}{2}}$	–	$(\log n)^{-\alpha+\frac{1}{2}}$
ReLu	$(\frac{\log n}{\log\log n \log d})^{-\frac{3}{4}}$	$n^{-\frac{3}{4d-2}}$	$(\log n)^{-\frac{3}{4}}$
tanh / logistic	$\exp(-c\sqrt{\frac{\log n}{\log d \log\log n}})$	–	$\exp(-c\sqrt{\log n})$
$J_q \sim e^{-\beta q}$	$n^{-c(\log\log n \log d)^{-1}}$	–	$n^{-\frac{1}{2}}$
erf	$n^{-c(\log\log n \log d)^{-1}}$	$C^d(\log n)^{\frac{d}{2}-1} n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}$
polynomial order p	$p^{cp} d^{\frac{5p}{6}-\frac{1}{12}} n^{-\frac{1}{6}}$	$(d+p)^{\frac{d}{2}} n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}$

for all $\alpha > 0$ (no matter how small) one has

$$\lim_{n\to\infty} \frac{(\log n)^{-\frac{3}{4}}}{n^{-\frac{3}{4d-2}}} = \lim_{n\to\infty} \frac{(\log n)^{-\frac{3}{4}}}{\exp(-\frac{3}{4n^{\alpha}-2}\log n)} = 0,$$

so that our bound in the d_2 metric decays faster that the one by [19] in $\mathcal{W}_2$ under these circumstances.

4 Proof of the main results

Our main results, Theorems 1 and 4, are proved in Sections 4.1 and 4.4, respectively. The proofs use auxiliary propositions and lemmas, which are established in Sections 4.2 and 4.3.

4.1 Proof of Theorem 1

The main idea behind our proof is as follows. For some integer Q to be fixed later, write

$$F = F_{\leq Q} + F_{>Q},$$

where

$$F_{\leq Q} := \sum_{q=0}^{Q} F_q, \qquad F_{>Q} := \sum_{q=Q+1}^{\infty} F_q,$$

and

$$F_q(x) := \frac{J_q(\sigma)}{\sqrt{n}} \sum_{j=1}^{n} V_j \frac{H_q(W_j x)}{\sqrt{q!}}, \qquad x \in \mathbb{S}^{d-1}.$$

In words, as anticipated in Section 2.1, we are partitioning our network into a component projected onto the Q lowest Wiener chaoses and the remainder projection on the highest chaoses. Now recall that Z is the zero mean Gaussian process with covariance function

$$\mathbb{E}\big[Z(x_1)Z(x_2)\big] := S(x_1, x_2) = \mathbb{E}\big[F(x_1)F(x_2)\big] = \sum_{q=0}^{\infty} J_q^2(\sigma)\langle x_1, x_2\rangle^q.$$

In the sequel we shall write $\{Z_q\}_{q\in\mathbb{N}}$ for a sequence of independent zero mean Gaussian variables with covariance function $\mathbb{E}[Z_q(x_1)Z_q(x_2)] := J_q^2(\sigma)\langle x_1, x_2\rangle^q$. Our idea is to use Theorem 3.10 in [8] and hence to consider

$$\begin{aligned} d_2(F, Z) &\leq d_2(F_{\leq Q}, Z) + d_2(F, F_{\leq Q}) \\ &\leq \frac{1}{2}\big(\sqrt{M(F_{\leq Q}) + C(F_{\leq Q})} + \|S - S_{\leq Q}\|_{L^2(\Omega,\mathrm{HS})}\big) + \mathcal{W}_2(F, F_{\leq Q}), \end{aligned}$$

where

$$\begin{aligned} M(F_{\leq Q}) &:= \frac{1}{\sqrt{3}}\sum_{p,q}^{Q} c_{p,q}\sqrt{\mathbb{E}\|F_p\|^4\big(\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4\big)}, \\ C(F_{\leq Q}) &:= \sum_{\substack{p,q \\ p\neq q}}^{Q} c_{p,q}\,\mathrm{Cov}\big(\|F_p\|^2, \|F_q\|^2\big), \\ c_{p,q} &:= \begin{cases} 1+\sqrt{3}, & p = q, \\ \frac{p+q}{2p}, & p \neq q, \end{cases} \end{aligned}$$

and we have

$$\mathcal{W}_2(F, F_{\leq Q}) \leq \sqrt{\sum_{q=Q+1}^{\infty} J_q^2(\sigma)}. \tag{5}$$

Moreover,

$$\|S - S_{\leq Q}\|_{L^2(\Omega,\mathrm{HS})}^2 \leq \sum_{q=Q+1}^{\infty} J_q^2(\sigma);$$

indeed, first note that the covariance operator can be written explicitly in coordinates as

$$\begin{aligned} &S_{\leq Q}(x_1, x_2) \\ &\quad = \frac{1}{n}\sum_{p,q}^{Q} J_p(\sigma)J_q(\sigma)\frac{1}{\sqrt{p!q!}}\sum_{j_1,j_2=1}^{n}\mathbb{E}\big[\{V_{j_1}H_p(W_{j_1}x_1)\}\{V_{j_2}H_q(W_{j_2}x_2)\}\big] \\ &\quad = \sum_{q}^{Q} J_q^2(\sigma)\langle x_1, x_2\rangle^q, \end{aligned}$$

and hence

$$S(x, y) - S_{\leq Q}(x, y) = \sum_{q=Q+1}^{\infty} J_q^2(\sigma)\langle x, y\rangle^q.$$

Therefore, taking the standard basis of spherical harmonics $\{Y_{\ell m}\}$, which are eigenfunctions of the covariance operators (see [22]),

$$\begin{aligned}
&\|S - S_{\le Q}\|^2_{L^2(\Omega,\mathrm{HS})}\\
&\quad= \sum_{\ell,\ell',m,m'} \sum_{q=Q+1}^{\infty} \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} Y_{\ell m}(x) Y_{\ell' m'}(y) \sum_{\ell'' m''} C_{\ell''}(q) Y_{\ell'' m''}(x) Y_{\ell'' m''}(y) dxdy\\
&\quad= \sum_{\ell,\ell',m,m'} \sum_{q=Q+1}^{\infty} \sum_{\ell'' m''} C_{\ell''}(q) \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} Y_{\ell m}(x) Y_{\ell' m'}(y) Y_{\ell'' m''}(x) Y_{\ell'' m''}(y) dxdy\\
&\quad= \sum_{\ell,\ell',m,m'} \sum_{q=Q+1}^{\infty} \sum_{\ell'' m''} C_{\ell''}(q) \delta_{\ell}^{\ell''} \delta_{\ell'}^{\ell''} \delta_{m}^{m''} \delta_{m'}^{m''}\\
&\quad= \sum_{\ell} \sum_{q=Q+1}^{\infty} C_{\ell}(q) n_{\ell;d}\\
&\quad= \sum_{q=Q+1}^{\infty} J_q^2(\sigma),
\end{aligned}$$

where $n_{\ell;d}$ is the dimension of the ℓ-th eigenspace in dimension d and $\{C_\ell(q)\}$ is the angular power spectrum of F_q, see again [22] for more discussion and details (the discussion in this reference is restricted to $d = 2$, but the results can be extended to any dimension).

We are left to bound $M(F_{\le Q})$ and $C(F_{\le Q})$. In Section 4.2 we will provide a bound for $M(F_{\le Q})$. Under the conditon $Q \le \log_3 \sqrt{n}$, such bound reduces to

$$M(F_{\le Q}) \lesssim \frac{\|\sigma\|^2}{\sqrt{n}} \sum_{q}^{Q} J_q^2(\sigma) q 3^q.$$

On the other hand, in Section 4.3 we will show that

$$C(F_{\le Q}) \le M(F_{\le Q}).$$

This completes the proof.

4.2 Bounding $M(F_{\le Q})$

The following proposition provides a bound on $M(F_{\le Q})$. The proof relies on several technical lemmas, which are given below.

Proposition 9. *We have*

$$M(F_{\le Q}) \lesssim \frac{1}{\sqrt{n}} \sum_{q=0}^{Q} J_q^2 q 3^q \left(\|\sigma\|^2 + \frac{1}{\sqrt{n}} \sum_{q=0}^{Q} J_q^2 3^q \right).$$

Proof. We can write

$$M(F_{\leq Q}) = \frac{1}{\sqrt{3}} \sum_{p,q}^{Q} c_{p,q} \sqrt{\mathbb{E}\|F_p\|^4 \left(\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4\right)}$$

$$\leq \sum_{p}^{Q} \sqrt{\mathbb{E}\|F_p\|^4} \sum_{q}^{Q} q \sqrt{\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4}.$$

In Lemma 10 we compute

$$\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4 = \frac{1}{n} \frac{J_q^4(\sigma)}{(q!)^2} \sum_{q_1=0}^{q-1} \Upsilon_{q_1,q} \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(q-q_1)} dx_1 dx_2,$$

with

$$\Upsilon_{q_1,q} = \binom{q}{q_1}^4 (q_1!)^2 (2q-2q_1)!.$$

By Lemma 11 we get the bound

$$\max_{0\leq q_1\leq q-1} \Upsilon_{q_1,q} \lesssim \frac{(q!)^2 3^{2q}}{q},$$

whereas Lemma 12 yields

$$\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(q-q_1)} dx_1 dx_2 \leq 1.$$

Therefore,

$$\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4 \lesssim \frac{J_q^4(\sigma) 3^{2q}}{n}.$$

Moreover, in view of Lemma 14, we have

$$\mathbb{E}\|F_p\|^4 \lesssim \frac{J_p^4(\sigma)3^{2p}}{n} + 3J_p^4(\sigma).$$

Collecting all the terms, we finally obtain the claim. □

In the following, we collect the technical lemmas used in the proof of Proposition 9.

Lemma 10. *We have*

$$\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4 = \frac{1}{n} \frac{J_q^4(\sigma)}{(q!)^2} \sum_{q_1=0}^{q-1} \Upsilon_{q_1,q} \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(q-q_1)} dx_1 dx_2$$

with $\Upsilon_{q_1,q} = \binom{q}{q_1}^4 (q_1!)^2 (2q-2q_1)!$.

Proof. We will write $\mathrm{Cum}(\cdot,\cdot,\cdot,\cdot)$ for the joint cumulant of four random variables, that is,

$$\mathrm{Cum}(X,Y,Z,W)=\mathbb{E}[XYZW]-\mathbb{E}[XY]\mathbb{E}[WZ]-\mathbb{E}[XZ]\mathbb{E}[WY]-\mathbb{E}[XW]\mathbb{E}[ZY].$$

We have

$$\begin{aligned}\mathbb{E}\|F_q\|^4 &= \frac{1}{n^2}\frac{J_q^4}{(q!)^2}\sum_{j_1,j_2,j_3,j_4=1}^{n}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\\ &\quad\times\mathbb{E}\big\{V_{j_1}H_q(W_{j_1}x_1)V_{j_2}H_q(W_{j_2}x_1)V_{j_3}H_q(W_{j_3}x_2)V_{j_4}H_q(W_{j_4}x_2)\big\}dx_1dx_2\\ &=\frac{1}{n}\frac{J_q^4}{(q!)^2}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\\ &\quad\times\mathrm{Cum}\big\{V_jH_q(W_jx_1),\,V_jH_q(W_jx_1),\,V_jH_q(W_jx_2),\,V_jH_q(W_jx_2)\big\}dx_1dx_2\\ &\quad+\frac{J_q^4}{(q!)^2}\left\{\int_{\mathbb{S}^{d-1}}\mathbb{E}\big\{V_jH_q(W_jx_1)V_jH_q(W_jx_1)\big\}dx_1\right\}^2\\ &\quad+2\frac{J_q^4}{(q!)^2}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\big\{\mathbb{E}\big\{V_jH_q(W_jx_1)V_jH_q(W_jx_2)\big\}\big\}^2dx_1dx_2.\end{aligned}$$

Now note that, in view of the normalization we adopted for the volume of $\mathbb{S}^{d-1}$,

$$\begin{aligned}&\frac{1}{(q!)^2}\left\{\int_{\mathbb{S}^{d-1}}\mathbb{E}\big\{V_jH_q(W_jx_1)V_jH_q(W_jx_1)\big\}dx_1\right\}^2=1,\\ &\frac{1}{(q!)^2}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\big\{\mathbb{E}\big\{V_jH_q(W_jx_1)V_jH_q(W_jx_2)\big\}\big\}^2dx_1dx_2\\ &\quad=\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\langle x_1,x_2\rangle^{2q}dx_1dx_2.\end{aligned}$$

Moreover,

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathbb{E}\big\{Z_q^2(x_1)Z_q^2(x_2)\big\}dx_1dx_2\\ &\quad=\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathbb{E}\big\{Z_q^2(x_1)\big\}\mathbb{E}\big\{Z_q^2(x_2)\big\}dx_1dx_2\\ &\qquad+2\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathbb{E}\big\{Z_q(x_1)Z_q(x_2)\big\}\mathbb{E}\big\{Z_q(x_1)Z_q(x_2)\big\}dx_1dx_2\\ &\quad=J_q^4+2J_q^4\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\langle x_1,x_2\rangle^{2q}dx_1dx_2.\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{E}\|F_q\|^4-\mathbb{E}\|Z_q\|^4\\ &=\frac{1}{n}\frac{J_q^4}{(q!)^2}\end{aligned}$$

$$\times \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \mathrm{Cum}\big\{V_1 H_q(W_1x_1), V_1 H_q(W_1x_1), V_1 H_q(W_1x_2), V_1 H_q(W_1x_2)\big\} dx_1 dx_2.$$

Using the diagram formula for Hermite polynomials [22, Proposition 4.15] and then isotropy, for $q_1 + q_2 + q_3 + q_4 = 2q$ we have

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \mathrm{Cum}\big\{V_1 H_q(W_1x_1), V_1 H_q(W_1x_1), V_1 H_q(W_1x_2), V_1 H_q(W_1x_2)\big\} dx_1 dx_2 \\
&\quad = \sum_{q_1+q_2+q_3+q_4=2q} \Upsilon_{q_1q_2q_3q_4} \\
&\qquad \times \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_1\rangle^{q_1} \langle x_1, x_2\rangle^{q_2} \langle x_2, x_2\rangle^{q_3} \langle x_2, x_1\rangle^{q_4} dx_1 dx_2 \\
&\quad = \sum_{q_1=0}^{q-1} \Upsilon_{q_1,q} \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(q-q_1)} dx_1 dx_2,
\end{aligned}$$

where $\Upsilon_{q_1q_2q_3q_4}$, $\Upsilon_{q_1,q}$ count the possible configurations of the diagrams. Precisely, $\Upsilon_{q_1,q}$ is the number of connected diagrams with no flat edges between four rows of q nodes each and $q_1 < q$ connections between first and second row. To compute this number explicitly, let us label the nodes of the diagram as

$$\begin{array}{cccccc}
x_1 & x_1 & x_1 & \dots & & x_1 \\
x_1' & x_1' & x_1' & \dots & & x_1' \\
x_2 & x_2 & x_2 & \dots & & x_2 \\
x_2' & x_2' & x_2' & \dots & & x_2'
\end{array}.$$

Because there cannot be flat edges, the number of edges between x_1 and x_1' is the same as the number of edges between x_2 and x_2'. Indeed, assume that the former was larger than the latter; then there would be less edges starting from the pair (x_1, x_1') and reaching the pair (x_2, x_2') than the other way round, which is obviously absurd. There are $\binom{q}{q_1}$ ways to choose the nodes of the first row connected with the second, $\binom{q}{q_1}$ ways to choose the nodes of the second connected with the first, $\binom{q}{q_1}$ ways to choose the nodes of the third connected with the fourth, and $\binom{q}{q_1}$ ways to choose the nodes of the fourth connected with the third, which gives a term of cardinality $\binom{q}{q_1}^4$; the number of ways to match the nodes between first and second row or third and fourth is $(q_1!)^2$. There are now $(2q - 2q_1)$ nodes left in the first two rows, which can be matched in any arbitrary way with the $(2q - 2q_1)$ remaining nodes of the third and the fourth row; the result follows immediately. □

Lemma 11. *The following bound holds true:*

$$\max_{0\le q_1\le q-1} \Upsilon_{q_1,q} \lesssim \frac{(q!)^2 3^{2q}}{q}.$$

Proof. We can write

$$\frac{1}{(q!)^2}\Upsilon_{q_1,q} = \frac{1}{(q!)^2}\binom{q}{q_1}^4 (q_1!)^2 (2q-2q_1)!$$

$$= \frac{(q!)^2(2q-2q_1)!}{(q_1!)^2((q-q_1)!)^2((q-q_1)!)^2} = \binom{q}{q_1}^2 \binom{2q-2q_1}{q-q_1}.$$

Note that both the elements in the last expression are decreasing in q_1, when $q_1 > \frac{q}{2}$ (say). Fix $\alpha \in [\varepsilon, \frac{1}{2}+\varepsilon]$, $\varepsilon > 0$; repeated use of Stirling's approximation gives

$$\begin{aligned}
&\frac{(q!)^2(2q-2q_1)!}{(q_1!)^2((q-q_1)!)^2((q-q_1)!)^2} \\
&\sim \frac{1}{(2\pi)^{3/2}} \frac{q^{2q+1}(2q-2q_1)^{2q-2q_1+\frac{1}{2}} e^{2q_1} e^{2(q-q_1)} e^{2q-2q_1}}{e^{2q} e^{2q-2q_1} q_1^{2q_1+1} (q-q_1)^{4q-4q_1+2}} \\
&\sim \frac{2^{2q-2q_1+\frac{1}{2}}}{(2\pi)^{3/2}} \frac{q^{2q+1}}{q_1^{2q_1+1}(q-q_1)^{2q-2q_1+\frac{3}{2}}}
\end{aligned}$$

Taking $q_1 = \alpha q$ we obtain

$$\begin{aligned}
&\frac{2^{2(1-\alpha)q+\frac{1}{2}}}{(2\pi)^{3/2}} \frac{q^{2q+1}}{(\alpha q)^{2\alpha q+1}((1-\alpha)q)^{2(1-\alpha)q+\frac{3}{2}}} \\
&= \frac{2^{2(1-\alpha)q+\frac{1}{2}}}{(2\pi)^{3/2}} \frac{1}{(\alpha)^{2\alpha q+1}((1-\alpha))^{2(1-\alpha)q+\frac{3}{2}} q^{\frac{3}{2}}} \\
&= \frac{2^{\frac{1}{2}}}{(2\pi)^{3/2} q^{\frac{3}{2}}} \left(\frac{2^{1-\alpha}}{\alpha^{\alpha+\frac{1}{2q}}(1-\alpha)^{1-\alpha+\frac{3}{4q}}} \right)^{2q}.
\end{aligned}$$

It can be immediately checked that the function $f(\alpha) := \frac{2^{1-\alpha}}{\alpha^{\alpha}(1-\alpha)^{1-\alpha}}$ admits a unique maximum at $\alpha = \frac{1}{3}$, for which the quantity gets bounded by $q^{-\frac{3}{2}} 3^{2q}$ up to constants. On the other hand, for $q_1 < \lfloor \varepsilon q \rfloor$ it suffices to notice that

$$\begin{aligned}
\binom{q}{q_1}^2 \binom{2q-2q_1}{q-q_1} &\leq 2^{2q} \binom{q}{q_1}^2 \leq 2^{2q} \binom{q}{\lfloor \varepsilon q \rfloor}^2 \\
&\leq 2^{2q} \frac{q^{2q+1}}{2\pi(\varepsilon q)^{2\varepsilon q+1}((1-\varepsilon)q)^{2(1-\varepsilon)q+1}},
\end{aligned}$$

where we used the fact that $g(\varepsilon) = \varepsilon^{-\varepsilon}(1-\varepsilon)^{-(1-\varepsilon)}$ is strictly increasing in $(0, \frac{1}{2})$; hence we get

$$2^{2q} \frac{q^{2q+1}}{2\pi(\varepsilon q)^{2\varepsilon q+1}((1-\varepsilon)q)^{2(1-\varepsilon)q+1}} = \frac{2^{2q}}{2\pi q} \frac{1}{((\varepsilon)^{\varepsilon+\frac{1}{2q}}((1-\varepsilon))^{(1-\varepsilon)+\frac{1}{2q}})^{2q}}.$$

The result is proved by choosing ε such that

$$\left((\varepsilon)^{\varepsilon+\frac{1}{2q}}\left((1-\varepsilon)\right)^{(1-\varepsilon)+\frac{1}{2q}}\right)^{-1} < \frac{3}{2}. \qquad \square$$

We recall the standard definition of the Beta function $B(\alpha, \beta)$:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp(-t) dt, \quad \alpha, \beta > 0.$$

Lemma 12. *We have*

$$\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(q-q_1)} dx_1 dx_2 = \frac{s_{d-1}}{s_d} B\left(q - q_1 + \frac{1}{2}, \frac{d}{2} - \frac{1}{2}\right) \leq 1.$$

Proof. Fixing a pole and switching to spherical coordinates, we get

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2q-2q_1} dx_1 dx_2\\ &= \frac{s_{d-1}}{s_d}\int_0^{\pi} (\cos\theta)^{2q-2q_1}(\sin\theta)^{d-2} d\theta\\ &= \frac{s_{d-1}}{s_d}\int_0^{\pi/2} \left(\cos^2\theta\right)^{q-q_1-\frac{1}{2}}\left(1-\cos^2\theta\right)^{\frac{d-3}{2}} d\cos\theta\\ &= \frac{s_{d-1}}{s_d}\int_0^{1} t^{q-q_1-\frac{1}{2}}(1-t)^{\frac{d-3}{2}} dt = \frac{s_{d-1}}{s_d} B\left(q - q_1 + \frac{1}{2}, \frac{d-1}{2}\right),\end{aligned}$$

which is smaller than 1 for all d, q. □

Remark 13. The bound we obtain is actually uniform over d. It is likely that it could be further improved for growing numbers of d, because the Beta function decreases quickly as d diverges.

Lemma 14. *We have*

$$\mathbb{E}\|F_p\|^4 \leq \mathbb{E}\|F_p\|^4 - \mathbb{E}\|Z_p\|^4 + 3J_p^4 .$$

Proof. It suffices to observe that, following the calculations of Lemma 10,

$$\mathbb{E}\|Z_q\|^4 = J_q^4 + 2J_q^4 \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2q} dx_1 dx_2 \leq 3J_q^4.$$ □

4.3 Bounding $C(F_{\leq Q})$

The following results reduces the problem of bounding $C(F_{\leq Q})$ to that of bounding $M(F_{\leq Q})$.

Proposition 15. *We have*

$$C(F_{\leq Q}) \leq M(F_{\leq Q}).$$

Proof. We shall show that

$$\begin{aligned}C(F_{\leq Q}) = &\sum_{p,q:p\neq q}^{Q} c_{p,q} \sum_{p_1=p-q}^{p-1} \binom{p}{p_1}^2 (p_1!)^2 \binom{q}{q-p+p_1}^2\\ &\times \left((q-p+p_1)!\right)^2 \left(2(p-p_1)\right)!\\ &\times \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(p-p_1)} dx_1 dx_2\\ \leq &\sum_{p,q}^{Q}\sum_{p_1=1}^{p} c_{p,q} \binom{p}{p_1}^4 (p_1!)^2 (2p-2p_1)! \int_{S^{d-1}\times S^{d-1}} \langle x_1, x_2\rangle^{2(p-p_1)} dx_1 dx_2\end{aligned}$$

$$= M(F_{\leq Q}).$$

Recall that

$$\begin{aligned}&\mathrm{Cov}\big(\|F_p\|^2, \|F_q\|^2\big)\\ &= \frac{J_p^2 J_q^2}{n}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\mathrm{Cum}\big\{VH_p(Wx_1), VH_p(Wx_1), VH_q(Wx_2), VH_q(Wx_2)\big\}dx_1dx_2.\end{aligned}$$

Indeed,

$$\|F_p\|^2 = \frac{J_p^2}{n}\frac{1}{p!}\sum_{j_1,j_2=1}^{n}\int_{S^{d-1}}\big\{V_{j_1}H_p(W_{j_1}x)\big\}\big\{V_{j_2}H_p(W_{j_2}x)\big\}dx,$$

and

$$\mathrm{Cov}\big(\|F_p\|^2, \|F_q\|^2\big) = \mathbb{E}\big(\|F_p\|^2\|F_q\|^2\big) - \mathbb{E}\big(\|F_p\|^2\big)\mathbb{E}\big(\|F_q\|^2\big),$$

where

$$\begin{aligned}&\mathbb{E}\big(\|F_p\|^2\|F_q\|^2\big) = \frac{J_p^2 J_q^2}{n^2}\frac{1}{p!\,q!}\sum_{j_1,j_2=1}^{n}\sum_{j_3,j_4=1}^{n}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\\ &\mathbb{E}\big[V_{j_1}H_p(W_{j_1}x_1)V_{j_2}H_p(W_{j_2}x_1)V_{j_3}H_q(W_{j_3}x_2)V_{j_4}H_q(W_{j_4}x_2)\big]dx_1dx_2\\ &= \frac{J_p^2 J_q^2}{n}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathrm{Cum}\big[V_1H_p(W_1x_1), V_1H_p(W_1x_1), V_1H_q(W_1x_2), V_1H_q(W_1x_2)\big]dx_1dx_2\\ &+ J_p^2J_q^2\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathbb{E}\big[V_1H_p(W_1x_1)V_1H_p(W_1x_1)\big]\mathbb{E}\big[V_1H_q(W_1x_2)V_1H_q(W_1x_2)\big]dx_1dx_2\\ &+ 2J_p^2J_q^2\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathbb{E}\big[V_1H_p(W_1x_1)V_1H_q(W_1x_2)\big]\mathbb{E}\big[V_1H_p(W_1x_1)V_1H_q(W_1x_2)\big]dx_1dx_2.\end{aligned}$$

By the orthogonality of the Hermite polynomials, the third term vanishes and we are left with

$$\begin{aligned}&\frac{J_p^2 J_q^2}{n}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathrm{Cum}\big[V_1H_p(W_1x_1), V_1H_p(W_1x_1), V_1H_q(W_1x_2), V_1H_q(W_1x_2)\big]dx_1dx_2\\ &+ J_p^2J_q^2\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathbb{E}\big[V_1H_p(W_1x_1)V_1H_p(W_1x_1)\big]\mathbb{E}\big[V_1H_q(W_1x_2)V_1H_q(W_1x_2)\big]dx_1dx_2\\ &= \frac{J_p^2 J_q^2}{n}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathrm{Cum}\big[V_1H_p(W_1x_1), V_1H_p(W_1x_1), V_1H_q(W_1x_2), V_1H_q(W_1x_2)\big]dx_1dx_2\\ &+ \left(J_p^2\int_{\mathbb{S}^{d-1}}\mathbb{E}\big[V_1H_p(W_1x_1)V_1H_p(W_1x_1)\big]dx_1\right)\left(J_q^2\int_{\mathbb{S}^{d-1}}\mathbb{E}\big[V_1H_q(W_1x_2)V_1H_q(W_1x_2)\big]dx_2\right)\\ &= \frac{J_p^2 J_q^2}{n}\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}}\mathrm{Cum}\big[V_1H_p(W_1x_1), V_1H_p(W_1x_1), V_1H_q(W_1x_2), V_1H_q(W_1x_2)\big]dx_1dx_2\\ &+ \mathbb{E}\big(\|F_p\|^2\big)\mathbb{E}\big(\|F_q\|^2\big).\end{aligned}$$

Indeed,

$$\mathbb{E}\big(\|F_p\|^2\big) = J_p^2\mathbb{E}\Bigg[\sum_{j_1,j_2=1}^{n}\int_{\mathbb{S}^{d-1}}\big\{V_{j_1}H_p(W_{j_1}x)\big\}\big\{V_{j_2}H_p(W_{j_2}x)\big\}dx\Bigg]$$

$$= J_p^2 \mathbb{E}\Big[n \int_{\mathbb{S}^{d-1}} \big\{V_1 H_p(W_1 x)\big\}\big\{V_1 H_p(W_1 x)\big\} dx \Big].$$

Now note that

$$\frac{J_p^2 J_q^2}{n} \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \mathrm{Cum}\big[V_1 H_p(W_1 x_1), V_1 H_p(W_1 x_1), V_1 H_q(W_1 x_2), V_1 H_q(W_1 x_2)\big] dx_1 dx_2$$
$$= \frac{J_p^2 J_q^2}{n} \sum_{p_1=p-q}^{p-1} \binom{p}{p_1}^2 p_1! \binom{q}{q-p+p_1}^2 (q-p+p_1)!\big(2(p-p_1)\big)!$$
$$\times \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(p-p_1)} dx_1 dx_2.$$

Moreover,

$$\binom{q}{q-p+p_1}^2 \big((q-p+p_1)!\big)^2 = \frac{(q!)^2}{((p-p_1)!)^2} \le \frac{(p!)^2}{((p-p_1)!)^2} = \binom{p}{p_1}^2 (p_1!)^2,$$

and hence

$$\sum_{p_1=p-q}^{p-1} \binom{p}{p_1}^2 (p_1!)^2 \binom{q}{q-p+p_1}^2 \big((q-p+p_1)!\big)^2 \big(2(p-p_1)\big)!$$
$$\times \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(p-p_1)} dx_1 dx_2$$
$$\le \sum_{p_1=1}^{p} \binom{p}{p_1}^4 (p_1!)^2 (2p-2p_1)! \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \langle x_1, x_2\rangle^{2(p-p_1)} dx_1 dx_2,$$

so that our previous bound on the fourth cumulant is sufficient, up to a factor $\frac{J_q^2}{J_p^2}\frac{q!}{p!} \ll 1$. □

4.4 Proof of Theorem 4

The proof of Theorem 4 takes advantage of the tighter bounds which are obtained in [8, Section 4]; we refer to this paper and Section A.1, together with the monograph [25], for more details on the notation and further discussion.

Consider the isonormal Gaussian process with the underlying Hilbert space

$$\mathcal{H} := L^2[0, 2\pi] \otimes L^2[0, 2\pi] \otimes \mathbb{R}^d;$$

we take

$$V_j = I(f_{V_j}) = I\left(\frac{\cos(\cdot)}{\sqrt{\pi}} \otimes \frac{\exp(ij\cdot)}{\sqrt{2\pi}} \otimes z\right) \text{ for some fixed } z \text{ such that } \|z\|_{\mathbb{R}^d} = 1,$$
$$W_j x = I(f_{W_j x}) = I\left(\frac{\sin(\cdot)}{\sqrt{\pi}} \otimes \frac{\exp(ij\cdot)}{\sqrt{2\pi}} \otimes x\right) \text{ for any } x \in \mathbb{S}^{d-1}.$$

It is readily seen that these are two Gaussian, zero mean, unit variance random variables, with covariances

$$\mathbb{E}[V_j V_{j'}] = \delta_j^{j'},\ \mathbb{E}[V_j W_j x] = 0,\ \mathbb{E}[W_j x_1 W_{j'} x_2] = \delta_j^{j'} \langle x_1, x_2 \rangle_{\mathbb{R}^d}.$$

Also, we have that

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{j=1}^{n} V_j \sigma(W_j x) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{p=1}^{\infty} \frac{J_p(\sigma)}{\sqrt{p!}} I(f_{V_j}) H_p(W_j x) \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{p=1}^{\infty} \frac{J_p(\sigma)}{\sqrt{p!}} I(f_{V_j}) I_p\big(f_{W_j x}^{\otimes p}\big),\end{aligned}$$

where we have used the standard identity linking Hermite polynomials and multiple stochastic integrals (i.e., Theorem 2.7.7 in [25]). To evaluate the term $I(f_{V_j}) I_p(f_{W_j x}^{\otimes p})$, we recall the product formula [25, Theorem 2.7.10]

$$I_p\big(f^{\otimes p}\big) I_q\big(g^{\otimes q}\big) = \sum_{r=0}^{p \wedge q} r \binom{p}{r} \binom{q}{r} I_{p+q-2r}(f \widetilde{\otimes}_r g);$$

in our case $p = 1$, $f_{V_j} \widetilde{\otimes}_1 f_{W_j x}^{\otimes p} = 0$, hence we obtain

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} I(f_{V_j}) I_p\big(f_{W_j x}^{\otimes p}\big) = I_{p+1}\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} f_{V_j} \widetilde{\otimes}_r f_{W_j x}^{\otimes p} \right),$$

where $\widetilde{\otimes}$ denotes the symmetrized tensor product. Let us now write

$$f_{p+1;x} := \frac{1}{\sqrt{n}} \frac{J_p(\sigma)}{\sqrt{p!}} \sum_{j=1}^{n} f_{V_j} \widetilde{\otimes}_r f_{W_j x}^{\otimes p};$$

it can then be readily checked that, for $K = L^2(\mathbb{S}^{d-1})$ and $r < (p_1 + 1 \wedge p_2 + 1)$,

$$\| f_{p_1+1;x_1} \otimes f_{p_2+1;x_2} \|^2_{\mathcal{H}^{\otimes(p_1+p_2-2r)}} = \frac{1}{n} \frac{J_p^4(\sigma)}{(p!)^2} \langle x_1, x_2 \rangle^{2r},$$

$$\| f_{p_1+1;x_1} \otimes f_{p_2+1;x_2} \|^2_{\mathcal{H}^{\otimes(p_1+p_2-2r)} \otimes K^{\otimes 2}} = \frac{1}{n} \frac{J_p^4(\sigma)}{(p!)^2} \int_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \langle x_1, x_2 \rangle^{2r} dx_1 dx_2.$$

To complete the proof, it is then sufficient to exploit [8, Theorem 4.3] and to follow similar steps as in the proof of Theorem 1.

A Appendix

A.1 The quantitative functional central limit theorem by Bourguin and Campese (2020)

In this paper, the probabilistic distance for the distance between the random fields we consider is the so-called d_2-metric, which is given by

$$d_2(F, G) = \sup_{\substack{h \in C_b^2(K) \\ \|h\|_{C_b^2(K)} \leq 1}} \big| \mathbb{E}\big[h(F)\big] - \mathbb{E}\big[h(G)\big] \big|;$$

here, $C_b^2(K)$ denotes the space of continuous and bounded applications from the Hilbert space K into $\mathbb{R}$ endowed with two bounded Frechet derivatives h', h''; that is, for each $h \in C_b^2(K)$ there exist a bounded linear operator $h' : K \to \mathbb{R}$ such that $\|h'\|_K \leq 1$

$$\lim_{\|v\|\to 0} \frac{|h(x+v) - h(x) - h'(v)|}{\|v\|} = 0,$$

and similarly for the second derivative.

We will use a simplified version of the results by Bourguin and Campese in [8], which we report below.

Theorem 16 (A special case of Theorem 3.10 in [8]). *Let $F_{\leq Q} \in L^2(\Omega, K)$ be a Hilbert-valued random element $F_{\leq Q} : \Omega \to K$ with zero mean, covariance operator $S_{\leq Q}$ and such that it can be decomposed into a finite number of Wiener chaoses:*

$$F_{\leq Q}(.) = \sum_{p=0}^{Q} F_p(.).$$

Then, for Z a Gaussian process on the same structure with covariance operator S we have that

$$d_2(F_{\leq Q}, Z) \leq \frac{1}{2}\sqrt{M(F_{\leq Q}) + C(F_{\leq Q})} + \|S - S_{\leq Q}\|_{L^2(\Omega,\mathrm{HS})},$$

where

$$M(F_{\leq Q}) = \frac{1}{\sqrt{3}} \sum_{p,q} c_{p,q} \sqrt{\mathbb{E}\|F_p\|^4 \big(\mathbb{E}\|F_q\|^4 - \mathbb{E}\|Z_q\|^4\big)},$$

$$C(F_{\leq Q}) = \sum_{\substack{p,q\\ p\neq q}} c_{p,q} \operatorname{Cov}\big(\|F_p\|^2, \|F_q\|^2\big),$$

Z_q a centred Gaussian process with the same covariance operator as F_q, i.e., ($\mathbb{E}[Z_q(x_1)Z_q(x_2)] = J_q^2(\sigma)\langle x_1, x_2\rangle^q$) and

$$c_{p,q} = \begin{cases} 1+\sqrt{3}, & p = q, \\ \frac{p+q}{2p}, & p \neq q. \end{cases}$$

Remark 17. The general version of Theorem 3.10 in [8] covers a broader class of processes which can be expanded into the eigenfunctions of Markov operators. We do not need this extra generality, and we refer to [8] for more discussion and details.

We will now review another result by [8], which holds under tighter smoothness conditions. We shall omit a number of details, for which we refer to classical references such as [25].

Given a Hilbert space $\mathcal{H}$ we recall the isonormal process is the collection of zero mean Gaussian random variables with covariance function

$$\mathbb{E}\big[X(h_1)X(h_2)\big] = \langle h_1, h_2\rangle_{\mathcal{H}}.$$

In our case these random variables take values in the separable Hilbert space $L^2(\Omega, \mathbb{S}^{d-1})$. For smooth functions $F : \Omega \to L^2(\Omega, \mathbb{S}^{d-1})$ of the form

$$F = f\big(W(h_1), \ldots, W(h_p)\big) \otimes v, \;\; f \in C_b^\infty\big(\mathbb{R}^p\big), \;\; v \in L^2\big(\Omega, \mathbb{S}^{d-1}\big),$$

we recall that the Malliavin derivative is defined as

$$DF = \sum_{i=1}^{p} \partial_i f\big(W(h_1), \ldots, W(h_p)\big) h_i \otimes v$$

whose domain, denoted by $\mathbb{D}^{1,2}$, is the closure of the space of smooth functions with respect to the Sobolev norm $\|F\|^2_{L^2(\Omega,\mathbb{S}^{d-1})} + \|DF\|^2_{L^2(\Omega,\mathcal{H}\otimes\mathbb{S}^{d-1})}$; $\mathbb{D}^{1,4}$ is defined analogously.

In this setting, the Wiener chaos decompositions take the form

$$F = \sum_{p=1}^{\infty} I_p(f_p), \;\; f_p \in \mathcal{H}^{\odot p} \otimes L^2\big(\mathbb{S}^{d-1}\big),$$

where $\mathcal{H}^{\odot p}$ denotes the p-fold symmetrized tensor product of $\mathcal{H}$, see [8, Subsection 4.1.2]. The main result we are going to exploit is their Theorem 4.3, which we can recall as follows.

Theorem 18 (A special case of Theorem 4.3 in [8])**.** *Let Z be a centred random element of $L^2(\mathbb{S}^{d-1})$ with covariance operator S and $F \in \mathbb{D}^{1,4}$ with covariance operator T and chaos decomposition $F = \sum_p I_p(f_p)$, where $f_p \in \mathcal{H}^{\odot p} \otimes L^2(\mathbb{S}^{d-1})$. Then*

$$d_2(F, Z) \leq \frac{1}{2}\big(\widetilde{M}(F) + \widetilde{C}(F) + \|S - T\|_{\mathrm{HS}}\big),$$

where

$$\widetilde{M}(F) = \sum_{p=1}^{\infty} \sqrt{\sum_{r=1}^{p-1} \widetilde{\Upsilon}^2_{p,p}(r) \|f_p \otimes_r f_p\|^2_{\mathcal{H}^{\otimes(2p-2r)} \otimes L^2(\mathbb{S}^{d-1})^{\otimes 2}}},$$

$$\widetilde{C}(F) = \sum_{1 \leq p,q \leq \infty,\; p \neq q}^{\infty} \sqrt{\sum_{r=1}^{p\wedge q} \widetilde{\Upsilon}^2_{p,q}(r) \|f_p \otimes_r f_q\|^2_{\mathcal{H}^{\otimes(p+q-2r)} \otimes L^2(\mathbb{S}^{d-1})^{\otimes 2}}},$$

and

$$\widetilde{\Upsilon}_{p,q}(r) = p^2 (r-1)! \binom{p-1}{r-1} \binom{q-1}{r-1} (p+q-2r)!.$$

A.2 The ReLu activation function

We consider here the most popular activation function, i.e., the standard ReLu defined by $\sigma(t) = t\mathbb{I}_{[0,\infty)}(t)$. The Hermite expansion is known to be given by (see for instance [13, Lemma 17], or [19, Theorem 2], and [14, 11]):

$$J_q(\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}}, & q = 0, \\ \frac{1}{2}, & q = 1, \\ 0, & q > 1, q \text{ odd} \\ \frac{(-1)^{\frac{q}{2}+1}(q-3)!!}{\sqrt{\pi}\sqrt{q!}}, & q \text{ even.} \end{cases}$$

The following lemma is standard (compare [19]), but we include it for completeness.

Lemma 19. *As* $q \to \infty$,

$$J_q^2 \sim \frac{\sqrt{2}}{\sqrt{\pi^3} q^{5/2}}.$$

Proof. The result follows from a straightforward application of Stirling's formula, which gives

$$\begin{aligned} q! &\sim \sqrt{2\pi} q^{q+\frac{1}{2}} \exp(-q), \\ (q-3)!! = \frac{(q-3)!}{2^{\frac{q}{2}-2}(\frac{q}{2}-2)!} &\sim \frac{(q-3)^{q-\frac{5}{2}} \exp(-q+3)}{2^{\frac{q}{2}-2}(\frac{q}{2}-2)^{\frac{q}{2}-\frac{3}{2}} \exp(-\frac{q}{2}+2)} \\ &= \frac{\exp(-\frac{q}{2}+1)}{(q-3)^{1/2}(\frac{q}{2}-2)^{1/2}} \left(1+\frac{1}{q-4}\right)^{\frac{q}{2}-2} (q-3)^{\frac{q}{2}}, \end{aligned}$$

so that

$$\begin{aligned} \frac{((q-3)!!)^2}{\pi(q)!} &\sim \frac{\frac{\exp(-q+2)}{(q-3)(\frac{q}{2}-2)}(1+\frac{1}{q-4})^{q-4}(q-3)^q}{\sqrt{2\pi^3}(q)^{q+\frac{1}{2}}\exp(-q)} \\ &\sim \frac{\exp(3)}{\sqrt{2\pi^3}(q-3)(\frac{q}{2}-2)\sqrt{q}} \left(1-\frac{3}{q}\right)^q \sim \frac{2^{1/2}}{\sqrt{\pi^3}(q)^{5/2}}. \end{aligned}$$ □

Remark 20. The corresponding covariance kernel is given by, for any $x_1, x_2 \in \mathbb{S}^{d-1}$,

$$\begin{aligned} &\mathbb{E}[\sigma(W^T x_1)\sigma(W^T x_2)] \\ &= \frac{1}{2\pi} + \frac{\langle x_1, x_2 \rangle}{4} + \frac{\langle x_1, x_2 \rangle^2}{4\pi} + \frac{1}{2\pi} \sum_{q=2}^{\infty} \frac{((2q-3)!!)^2}{(2q)!} \langle x_1, x_2 \rangle^{2q} \\ &= \frac{1}{\pi}\left(u(\pi - \arccos u) + \sqrt{1-u^2}\right), \end{aligned}$$

for $u = \langle x_1, x_2 \rangle$, see also [5].

Remark 21. The rate for J_q in Lemma 19 is consistent with the one obtained by [19]. In [13], $J_q^2 = O(q^{-3})$ is given instead, yielding in [13, Theorem 3] the rate

$$\left(\frac{\log d \times \log\log n}{\log n}\right).$$

According to Lemma 19, this rate becomes

$$\left(\frac{\log d \times \log\log n}{\log n}\right)^{3/4},$$

which is the one we actually report in Table 1.

Acknowledgement

This paper has originated by a very inspiring short course taught by Boris Hanin at the University of Rome Tor Vergata in January 2023; we are very grateful to him for many deep insights and illuminating conversations.

Funding

The work was partially supported by the MUR Excellence Department Project MatMod@TOV awarded to the Department of Mathematics, University of Rome Tor Vergata, CUP E83C18000100006. We also acknowledge financial support from the MUR 2022 PRIN project *GRAFIA*, project code 202284Z9E4, the INdAM group GNAMPA and the PNRR CN1 High Performance Computing, Spoke 3.

References

[1] Azmoodeh, E., Peccati, G., Yang, X.: Malliavin-Stein method: a survey of some recent developments. Mod. Stoch. Theory Appl. **8**(2), 141–177 (2021). MR4279874. https://doi.org/10.15559/21-vmsta184

[2] Bach, F.: Breaking the curse of dimensionality with convex neural networks. J. Mach. Learn. Res. **18**(19), 1–53 (2017). MR3634886

[3] Basteri, A., Trevisan, D.: Quantitative Gaussian approximation of randomly initialized deep neural networks (2022). arXiv:2203.07379

[4] Belkin, M., Hsu, D., Ma, S., Mandal, S.: Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc. Natl. Acad. Sci. USA **116**(32), 15849–15854 (2019). MR3997901. https://doi.org/10.1073/pnas.1903070116

[5] Bietti, A., Bach, F.: Deep equals shallow for ReLu networks in kernel regimes. In: International Conference on Learning representations (ICLR), 9 (2021)

[6] Bordino, A., Favaro, F., Fortini, S.: Non-asymptotic approximations of Gaussian neural networks via second-order Poincaré inequalities (2023). arXiv:2304.04010

[7] Bourguin, S., Campese, S., Leonenko, N., Taqqu, M.S.: Four moments theorems on Markov chaos. Ann. Probab. **47**(3), 1417–1446 (2019). MR3945750. https://doi.org/10.1214/18-AOP1287

[8] Bourguin, S., Campese, S.: Approximation of Hilbert-valued Gaussians on Dirichlet structures. Electron. J. Probab. **25**, 150 (2020), 30 pp. MR4193891. https://doi.org/10.1214/20-ejp551

[9] Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019) (2019).

[10] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**(4), 303–314 (1989). MR1015670. https://doi.org/10.1007/BF02551274

[11] Daniely, A., Frostig, R., Singer, Y.: Toward deeper understanding of neural networks: the power of initialization and a dual view on expressivity. In: NeurIPS 2016, Volume **29**, 2253–2261 (2016)

[12] Döbler, C., Kasprzak, M., Peccati, G.: The multivariate functional de Jong CLT. Probab. Theory Relat. Fields **184**(1–2), 367–399 (2022). MR4498513. https://doi.org/10.1007/s00440-022-01114-3

[13] Eldan, R., Mikulincer, D., Schramm, T.: Non-asymptotic approximations of neural networks by Gaussian processes (2021). arXiv:2102.08668

[14] Goel, S., Karmalkar, S., Klivans, S.A.: Time/accuracy tradeoffs for learning a ReLu with respect to Gaussian marginals. In: NeurIPS 2019, pp. 8582–8591 (2019)

[15] Hanin, B.: Random neural networks in the infinite width limit as Gaussian processes (2021). arXiv:2107.01562

[16] Hornik, K.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989). https://doi.org/10.1016/0893-6080(89)90020-8

[17] Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Netw. **4**(2), 251–257 (1991). https://doi.org/10.1016/0893-6080(91)90009-T

[18] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems 31 (NeurIPS 2018) (2018).

[19] Klukowski, A.: Rate of convergence of polynomial networks to Gaussian processes (2021). arXiv:2111.03175

[20] Ledoux, M., Nourdin, I., Peccati, G.: Stein's method, logarithmic Sobolev and transport inequalities. Geom. Funct. Anal. **25**(1), 256–306 (2015). MR3320893. https://doi.org/10.1007/s00039-015-0312-0

[21] Leshno, M., Lin, V.Ya., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. **6**(6), 861–867 (1993). https://doi.org/10.1016/S0893-6080(05)80131-5

[22] Marinucci, D., Peccati, G.: Random Fields on the Sphere. Cambridge University Press (2011). MR2840154. https://doi.org/10.1017/CBO9780511751677

[23] Neal, R.M.: Priors for infinite networks. In: Bayesian Learning for Neural Networks, pp. 29–53. Springer, New York, NY (1996). https://doi.org/10.1007/978-1-4612-0745-0_2

[24] Nourdin, I., Peccati, G.: Stein's method on Wiener chaos. Probab. Theory Relat. Fields **145**(1–2), 75–118 (2009). MR2520122. https://doi.org/10.1007/s00440-008-0162-x

[25] Nourdin, I., Peccati, G.: Normal Approximations with Malliavin Calculus. From Stein's Method to Universality. Cambridge Tracts in Math., vol. 192. Cambridge University Press, Cambridge (2012). MR2962301. https://doi.org/10.1017/CBO9781139084659

[26] Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica (1999). MR1819645. https://doi.org/10.1017/S0962492900002919

[27] Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Advances in Neural Information Processing Systems 20 (NeurIPS 2007) (2007)

[28] Roberts, Yaida S, D.A., Hanin, B.: The principles of deep learning theory (2021). arXiv:2106.10165

[29] Yaida, S.: Non-Gaussian processes and neural networks at finite widths (2019). arXiv:1910.00019. MR4198759. https://doi.org/10.1007/s40687-020-00233-4

[30] Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: 5th International Conference on Learning Representations (ICLR) (2017)