Non-asymptotic approximations of neural networks by Gaussian processes

Ronen Eldan   RONENELDAN@GMAIL.COM
Weizmann Institute

Dan Mikulincer   DANMIKU@GMAIL.COM
Weizmann Institute

Tselil Schramm   TSELILSCHRAMM@GMAIL.COM
Stanford University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the extent to which wide neural networks may be approximated by Gaussian processes, when initialized with random weights. It is a well-established fact that as the width of a network goes to infinity, its law converges to that of a Gaussian process. We make this quantitative by establishing explicit convergence rates for the central limit theorem in an infinite-dimensional functional space, metrized with a natural transportation distance. We identify two regimes of interest; when the activation function is polynomial, its degree determines the rate of convergence, while for non-polynomial activations, the rate is governed by the smoothness of the function.

1. Introduction

In the past decade, artificial neural networks have experienced an unprecedented renaissance. However, the current theory has yet to catch-up with the practice and cannot explain their impressive performance. Particularly intriguing is the fact that *over-parameterized* models do not tend to over-fit, even when trained to zero error on the training set. Owing to this seemingly paradoxical fact, researchers have focused on understanding the infinite-width limit of neural networks. This line of research has led to many important discoveries such as the 'lazy-training' regime Chizat et al. (2019); Tzen and Raginsky (2020) which is governed by the limiting 'neural tangent kernel' (see Jacot et al. (2018)), as well as the 'mean-field' limit approach (see Mei et al. (2018, 2019); Karakida et al. (2019) for some examples) to study the training dynamics and loss landscape.

The first to study the limiting distribution of a neural network at (a random) initialization was Neal (Neal, 1996), who proved a *Central Limit Theorem (CLT)* for two-layered wide neural networks. According to Neal's result, when initialized with random weights, as the width of the network goes to infinity, its law converges, in distribution, to a Gaussian process. Subsequent works have generalized this result to deeper networks and other architectures (Williams (1997); Matthews et al. (2018); Novak et al. (2018); Garriga-Alonso et al. (2018); Hazan and Jaakkola (2015); Yang (2019a,b)). We also mention the work of Matsubara et al. (2020), which considered the problem of finding an appropriate random initialization that gives rise to valid approximations for Gaussian processes with a pre-determined covariance structure. This correspondence between Gaussian processes and

neural networks has proved to be highly influential and has inspired many new models (see Yang (2019a) for a thorough review of these models).

Towards supplying a theoretical framework to study *real world* neural networks, one important challenge is to understand the extent to which existing asymptotic results, which essentially only apply to infinite networks, may also be applied to finite ones. While there have been several works in this direction (c.f. Arora et al. (2019); Geiger et al. (2020); Hanin (2019); Andreassen and Dyer (2020); Yaida (2019); Antognini (2019); Aitken and Gur-Ari (2020); Bracale et al. (2021)), to the best of our knowledge, all known results consider finite-dimensional marginals of the random process and the question of finding a finite-width quantitative analog to Neal's CLT, which applies in a functional space, has remained open. The main goal of this paper is to tackle this question.

In essence, we prove a quantitative CLT in the space of functions. On a first glance, this is a completely different setting than the classical CLT, even in high-dimensional regimes. The function space is infinite-dimensional, while all quantitative bounds of CLT deteriorate with the dimension (Bobkov (2018); Courtade et al. (2019); Bonis (2020)). However, by exploiting the special structure of neural networks we are able to reduce the problem to finite-dimensional sets, where we capitalize on recent advances made in understanding the rate of convergence of the high-dimensional CLT. In particular, we give quantitative bounds, depending on the network's width and the dimension of the input, which show that, when initialized randomly, wide but finite networks can be well-approximated by a Gaussian process. The functional nature of our results essentially means that when considering the joint distribution attained on a finite set of inputs to the function, ours bounds do not deteriorate as the number of input points increases.

Roughly speaking, we prove the following results:

- We first consider two-layered networks with polynomial activation functions. By embedding the network into a high-dimensional tensor space we prove a quantitative CLT, with a polynomial rate of convergence in a strong transportation metric.

- We next consider general activations and show that under a (very mild) integrability assumption, one can reduce this case to the polynomial case. This is done at the cost of weakening the transportation distance. The rate of convergence depends on the smoothness of the activation and is typically sub-polynomial.

Let us remark about the methods of proof and the technical contributions. Whereas existing proofs of Gaussian behavior of neural networks apply the central limit theorem on finite dimensional marginals, our proof needs to be different, since we have to apply the CLT to a representation of the function itself (which does not depend on input points). We do this by embedding the network in a finite dimensional Wiener chaos, and invoking a CLT for high dimensions in those spaces. As far as we are aware, this is the first work to follow such a strategy. In order to apply off-the-shelf central limit theorems to our framework, we need to prove results concerning the covariance structures and variance bounds for polynomials. It is plausible that the idea of using Wiener chaos, as well as our implementation of it, could be useful outside the scope of this work.

Organization: The rest of the paper is devoted to describing and proving the results described above. In Section 2 we give the necessary background concerning random initializations of neural networks and we introduce a metric between random processes on the sphere. Our main results are stated in Section 3. In Section 4 we prove results which concern polynomial activations, while in Section 5 we consider general activations.

2. Background

Let $\sigma : \mathbb{R} \to \mathbb{R}$ and fix a dimension $n > 1$. A two-layered network with activation σ is a function $N : \mathbb{R}^n \to \mathbb{R}$, of the form

$$N(x) = \sum_{i=1}^{k} c_i \sigma \left(u_i \cdot x \right),$$

where $u_i \in \mathbb{R}^n, c_i \in \mathbb{R}$, for every $i = 1, \ldots, k$. We will refer to k as the width of the network. In most training procedures, it is typical to initialize the weight as *i.i.d.* random vectors. Specifically, let $\{w_i\}_{i=1}^{k}$ be *i.i.d.* as standard Gaussians in $\mathbb{R}^n$ and let $\{s_i\}_{i=1}^{k}$ be *i.i.d.* with $\mathbb{P}(s_1 = 1) = \mathbb{P}(s_1 = -1) = \frac{1}{2}$. We consider the random network,

$$\mathcal{P}_k \sigma(x) := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i \sigma \left(w_i \cdot x \right).$$

Let $\mathbb{S}^{n-1}$ stand for the unit sphere in $\mathbb{R}^n$. By restricting our attention to $x \in \mathbb{R}^n$, with $\|x\| = 1$, we may consider $\mathcal{P}_k \sigma$ as a random process, indexed by $\mathbb{S}^{n-1}$. In other words, $\mathcal{P}_k \sigma$ is a random vector in $L^2(\mathbb{S}^{n-1})$, equipped with its canonical rotation-invariant probability measure.

A Gaussian process is a random vector $\mathcal{G} \in L^2(\mathbb{S}^{n-1})$, such that for any finite set $\{x_j\}_{j=1}^{m} \subset \mathbb{S}^{n-1}$ the random vector $\{\mathcal{G}(x_j)\}_{j=1}^{m} \in \mathbb{R}^m$, has a multivariate Gaussian law. Since $\mathcal{P}_k \sigma$ is a sum of independent centered vectors, standard reasoning suggests that as $k \to \infty$, $\mathcal{P}_k \sigma$ should approach a Gaussian law in $L^2(\mathbb{S}^{n-1})$, i.e. a Gaussian process. Indeed, this is precisely Neal's CLT, which proves the existence of a Gaussian process $\mathcal{G}$, such that $\mathcal{P}_k \sigma \xrightarrow{k \to \infty} \mathcal{G}$, where the convergence is in distribution.

To make this result quantitative, we must first specify a metric. Our choice is inspired by the classical Wasserstein transportation in Euclidean spaces. The observant reader may notice that our definition, described below, does not correspond to the p-Wasserstein distance on $L^2(\mathbb{S}^{n-1})$, but rather the p-Wasserstein distance on $L^p(\mathbb{S}^{n-1})$. We chose this presentation for ease of exposition and its familiarity.

For $\mathcal{P}, \mathcal{P}'$, random processes on the sphere, and $p \geq 1$ we define the functional p-Wasserstein distance as,

$$\mathcal{WF}_p(\mathcal{P}, \mathcal{P}') := \inf_{(\mathcal{P}, \mathcal{P}')} \left(\int_{\mathbb{S}^{n-1}} \mathbb{E} \left[\left| \mathcal{P}(x) - \mathcal{P}'(x) \right|^p \right] dx \right)^{\frac{1}{p}},$$

3

where the infimum is taken over all couplings of $(\mathcal{P}, \mathcal{P}')$ and where dx is to be understood as the normalized uniform measure on $\mathbb{S}^{n-1}$. For $p = \infty$ we define $\mathcal{WF}_\infty$ as,

$$\mathcal{WF}_\infty(\mathcal{P}, \mathcal{P}') := \inf_{(\mathcal{P}, \mathcal{P}')} \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} \left|\mathcal{P}(x) - \mathcal{P}'(x)\right|\right].$$

The notation $\mathcal{W}_p$ is reserved to the p-Wasserstein distance in finite-dimensional Euclidean spaces. Explicitly, if X and Y are random vectors in $\mathbb{R}^N$, then

$$\mathcal{W}_p(X, Y) := \inf_{(X,Y)} \mathbb{E}\left[\|X - Y\|_2^p\right]^{\frac{1}{p}}.$$

There is a straightforward way to connect between the quadratic functional Wasserstein distance on the sphere and the L^2 distance in Gaussian space.

Lemma 1 *Let $f, g : \mathbb{R} \to \mathbb{R}$, and γ, the standard Gaussian in $\mathbb{R}$. Then,*

$$\mathcal{WF}_2^2(\mathcal{P}_k f, \mathcal{P}_k g) \leq \int_{\mathbb{R}} (f(x) - g(x))^2 d\gamma(x).$$

Proof There is a natural coupling such that

$$\mathcal{WF}_2^2(\mathcal{P}_k f, \mathcal{P}_k g) \leq \frac{1}{k} \int_{\mathbb{S}^{n-1}} \mathbb{E}\left[\left(\sum_{i=1}^k s_i(f(w_i \cdot x) - g(w_i \cdot x))\right)^2\right] dx$$

$$= \frac{1}{k} \int_{\mathbb{S}^{n-1}} \sum_{i=1}^k \mathbb{E}\left[(f(w_i \cdot x) - g(w_i \cdot x))^2\right] dx$$

$$= \int_{\mathbb{R}} (f(x) - g(x))^2 d\gamma(x).$$

The first equality is a result of independence, while the second equality follows from the fact that for any $x \in \mathbb{S}^{n-1}$, $w_i \cdot x \sim \gamma$. ∎

3. Results

We now turn to describe the quantitative CLT convergence rates obtained by our method. Our first result deals with polynomial activations.

Theorem 2 *Let $p(x) = \sum_{m=0}^d a_m x^m$ be a degree d polynomial. Then, there exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that*

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p, \mathcal{G}) \leq C_d \max_m \{|a_m|^2\} \left(\frac{n^{5d-\frac{1}{2}}}{k}\right)^{\frac{1}{3}},$$

where $C_d \leq d^{Cd}$, for some numerical constant $C > 0$.

According to the result, when the degree d is fixed, as long as $k \gg n^{5d - \frac{1}{2}}$, $\mathcal{P}_k p$ is close to a Gaussian process. One way to interpret the metric $\mathcal{WF}_\infty$, in the result, is as follows. For any finite set $\{x_j\}_{j=1}^m \subset \left(\mathbb{S}^{n-1}\right)^m$, the random vector $\{\mathcal{P}_k p(x_j)\}_{j=1}^m \subset \mathbb{R}^m$ converges to a Gaussian random vector, uniformly in m. Let us also mention that while the result is stated for Gaussian weights, the Gaussian plays no special role here (as will become evident from the proof), and the weights could be initialized by any symmetric random vector with sub-Gaussian tails.

One drawback of using polynomial activations is that the resulting network will always be a polynomial of bounded degree, which limits its expressive power. For this reason, in practice, neural networks are usually implemented using non-polynomial activations. By using a polynomial approximation scheme in Gaussian space, we are able to extend our result to this setting as well. We defer the necessary definitions and formulation of the result to Section 5, but mention here two specialized cases of common activations.

We first consider the Rectified Linear Unit (ReLU) function, denoted as $\psi(x) := \max(0, x)$. For this activation, we prove:

Theorem 3 *There exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that,*

$$\mathcal{WF}_2^2(\mathcal{P}_k \psi, \mathcal{G}) \leq C \left(\frac{\log(n) \log(\log(k))}{\log(k)} \right)^2,$$

where $C > 0$ is a numerical constant.

The reader might get the impression that this is a weaker result than Theorem 2. Indeed, the rate of convergence here is much slower. In order to get $\mathcal{WF}_2(\mathcal{P}_k \psi, \mathcal{G}) \leq \varepsilon$, the theorem requires that $k \gtrsim n^{\frac{1}{\varepsilon} \log \log n}$. Also, $\mathcal{WF}_2$ is a weaker metric than $\mathcal{WF}_\infty$. Let us point out that it may not be reasonable to expect similar behavior for polynomial and non-polynomial activations. The celebrated universal approximation theorem of Cybenko (Cybenko (1989), see also Leshno et al. (1993); Barron (1993)) states that any function in $L^2(\mathbb{S}^{n-1})$ can be approximated, to any precision, by a sufficiently wide neural network with a non-polynomial activation. Thus, as $k \to \infty$, the limiting support of $\mathcal{P}_k \psi$ will encompass all of $L^2(\mathbb{S}^{n-1})$. This is in sharp contrast to a polynomial activation function, for which the support of $\mathcal{P}_k p$ is always contained in some finite-dimensional subspace of $L^2(\mathbb{S}^{n-1})$, uniformly in k.

Another explanation for the slow rate of convergence, is the fact that ψ is non-differentiable. For smooth functions, the rate can be improved, but will still be typically sub-polynomial. As an example, we consider the hyperbolic tangent activation, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Theorem 4 *There exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that,*

$$\mathcal{W}_2^2(\mathcal{P}_k \tanh, \mathcal{G}) \leq C \exp\left(-\frac{1}{C} \sqrt{\frac{\log(k)}{\log(n) \log(\log(k))}} \right),$$

where $c > 0$ is an absolute constant.

Finally, let us remark about possible improvements to our obtained rates. We do not know whether the constant C_d in Theorem 5 is necessarily exponential, and we have made no effort to optimize it. We do conjecture that the dependence on the ratio $\frac{n^{5d - \frac{1}{2}}}{k}$ is not tight. To support this claim we prove an improved rate when the activation is monomial.

Theorem 5 *Let $p(x) = x^d$ for some $d \in \mathbb{N}$. Then, there exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that*

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p, \mathcal{G}) \leq C_d \frac{n^{2.5d - 1.5}}{k},$$

where $C_d \leq d^{Cd}$, for some numerical constant $C > 0$.

Remark 6 *It is plausible the dependence on d and k could be further improved. Let us note that when $d = 2$, the best rate one could hope for is proportional to $\frac{n^3}{k}$. This is a consequence of the bounds proven in Jiang and Li (2015); Bubeck et al. (2016), which show that if $n^3 \gg k$, then when considered as a random bi-linear form (or a Wishart matrix) $\mathcal{P}_k p$ is far from any Gaussian law. In fact, our proof of Theorem 5 can actually be improved when $d = 2$ (or, in general, for even d), and we are able to obtain the sharp rate $\frac{n^3}{k}$. It is an interesting question to understand the correct rates when $d > 2$.*

Remark 7 *Trivially, $\mathcal{WF}_2 \leq \mathcal{WF}_\infty$. Thus, Theorems 2 and 5 could be weakened and stated with respect to $\mathcal{WF}_2$. However, it is not clear to us whether one can actually obtain better bounds for polynomial activations if the stronger metric is replaced by the weaker one.*

4. Polynomial processes

For this section, fix a polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree d, $p(x) = \sum\limits_{m=0}^{d} a_m x^m$. The goal of this section is to show that when k is large enough, $\mathcal{P}_k p$ can be well approximated by a Gaussian process in the $\mathcal{WF}_\infty$ metric. Towards this, we will use the polynomial p to embed $\mathbb{R}^n$ into some high-dimensional tensor space.

4.1. The embedding

For $m \in \mathbb{N}$, we make the identification $(\mathbb{R}^n)^{\otimes m} = \mathbb{R}^{n^m}$ and focus on the subspace of symmetric tensors, which we denote $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$. If $\{e_i\}_{i=1}^n$ is the standard orthonormal basis of $\mathbb{R}^n$, then an orthonormal basis for $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$, is given by the set

$$\{e_I | I \in \mathrm{MI}_n(m)\}.$$

where $\mathrm{MI}_n(m)$ is the set of multi-indices,

$$\mathrm{MI}_n(m) = \{(I_1, \ldots I_n) \in \mathbb{N}^n | I_1 + \cdots + I_n = m\},$$

With this perspective, we have $e_I = \otimes_{i=1}^n \left(e_i^{\otimes I_i}\right)$, and we denote the inner product on $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$ by $\langle \cdot, \cdot \rangle_m$. We also use the following multi-index notation: if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote $x^I = \prod\limits_{i=1}^n x_i^{I_i}$.

Define the feature space $H := \oplus_{m=0}^d \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$. If $\pi_m : H \to \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$ is the natural projection, then an inner product on H may be defined by

$$\langle v, u \rangle_H := \sum_{m=0}^d \langle \pi_m v, \pi_m u \rangle_m.$$

6

We further define the embedding $P : \mathbb{R}^n \to H$, $P(x) = \sum_{m=0}^{d} \sqrt{|a_m|} x^{\otimes m}$, which induces a bi-linear form on H as,

$$Q(u, v) := \sum_{m=0}^{d} \text{sign}(a_m) \langle \pi_m u, \pi_m v \rangle_m.$$

Observe that Q is not necessarily positive definite, but still satisfies the following Cauchy-Schwartz type inequality,

$$Q(u, v) \leq \|u\|_H \|v\|_H. \tag{1}$$

Furthermore, it is clear that for any $x, y \in \mathbb{R}^n$,

$$Q(P(x), P(y)) = \sum_{m=0}^{d} a_m (x \cdot y)^m = p(x \cdot y),$$

and we have the identity,

$$\mathcal{P}_k p(x) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i p(w_i \cdot x) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i Q(P(x), P(w_i)) = Q\left(P(x), \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)\right). \tag{2}$$

Consider the random vector $X_k := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)$ taking values in H. By the central limit theorem, we should expect X_k to approach a Gaussian law. The next result shows that approximate Gaussianity of X_k implies that the process $\mathcal{P}_k p$ is approximately Gaussian as well.

Lemma 8 *Let G be a Gaussian random vector in H and define the random process on $\mathcal{G}$ in $\mathbb{S}^{n-1}$ by $\mathcal{G}(x) := Q(P(x), G)$. Then, $\mathcal{G}$ is a Gaussian process and,*

$$\mathcal{WF}_{\infty}^2(\mathcal{P}_k p, \mathcal{G}) \leq \left(\sum_{m=0}^{d} |a_m|\right) \mathcal{W}_2^2(X_k, G).$$

Proof Let (X_k, G) be the optimal coupling so that $\mathcal{W}_2^2(X_k, G) = \mathbb{E}\left[\|X_k - G\|_H^2\right]$. We then have

$$\mathcal{WF}_{\infty}(\mathcal{P}_k p, \mathcal{G}) \leq \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} |\mathcal{P}_k p(x) - \mathcal{G}(x)|\right] = \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} |Q(P(x), X_k - G)|\right]$$

$$\leq \sup_{x \in \mathbb{S}^{n-1}} \|P(x)\|_H \sqrt{\mathbb{E}\left[\|X_k - G\|_H^2\right]} = \sup_{x \in \mathbb{S}^{n-1}} \|P(x)\|_H \cdot \mathcal{W}_2(X_k, G),$$

where we have used (1) in the second inequality. Now, for any $x \in \mathbb{S}^{n-1}$,

$$\|P(x)\|_H = \sqrt{\sum_{m=0}^{d} |a_m| \langle x^{\otimes m}, x^{\otimes m} \rangle_m} = \sqrt{\sum_{m=0}^{d} |a_m|}.$$

7

■

So, we wish to show that the random vector $X_k := \frac{1}{\sqrt{k}} \sum_{i=I}^{k} s_i P(w_i)$ is approximately Gaussian inside H. For this, we will apply the following Wasserstein CLT bound, recently proven by Bonis, in Bonis (2020).

Theorem 9 *(Bonis, 2020, Theorem 1) Let Y_i be i.i.d isotropic random vectors in $\mathbb{R}^N$ and let G be the standard Gaussian. Then, if $S_k = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} Y_i$,*

$$\mathcal{W}_2^2(S_k, G) \leq \frac{\sqrt{N}}{k} \left\| \mathbb{E} \left[Y Y^T \|Y\|_2^2 \right] \right\|_{HS}.$$

Since the theorem applies to isotropic random vectors, for which the covariance matrix is the identity, we first need to understand $\Sigma := \text{Cov}(P(w))$. Let us emphasize the fact that Σ is a bi-linear operator on H. Thus it can be regarded as a $\dim(H) \times \dim(H)$ positive semi-definite matrix.

4.2. The covariance matrix

We first show that one may disregard small eigenvalues of Σ. Let (λ_j, v_j) stand for the eigenvalue/vector pairs of Σ. Fix $\delta > 0$ define $V_\delta = \text{span}(v_j | \lambda_j \leq \delta)$ and let $\Pi_\delta, \Pi_\delta^\perp$ be the orthogonal projection unto $V_\delta, V_\delta^\perp$, respectively.

Lemma 10 *Let $G \sim \mathcal{N}(0, \Sigma)$ be a Gaussian in H, then*

$$\mathcal{W}_2^2(X_k, G) \leq \mathcal{W}_2^2(\Pi_\delta^\perp X_k, \Pi_\delta^\perp G) + 8n^d \delta.$$

Proof For any coupling (X_k, G) we have

$$\mathcal{W}_2^2(X_k, G) \leq \mathbb{E}\left[\|X_k - G\|^2\right] = \mathbb{E}\left[\|\Pi_\delta X_k - \Pi_\delta G\|^2\right] + \mathbb{E}\left[\left\|\Pi_\delta^\perp X_k - \Pi_\delta^\perp G\right\|^2\right]$$

$$\leq 2\mathbb{E}\left[\|\Pi_\delta G\|^2\right] + 2\mathbb{E}\left[\|\Pi_\delta X_k\|^2\right] + \mathbb{E}\left[\left\|\Pi_\delta^\perp X_k - \Pi_\delta^\perp G\right\|^2\right]$$

$$\leq 4\dim(H)\delta + \mathbb{E}\left[\left\|\Pi_\delta^\perp X_k - \Pi_\delta^\perp G\right\|^2\right].$$

The proof concludes by taking the coupling for which $\Pi_\delta^\perp X_k, \Pi_\delta^\perp G$ is optimal, and by noting $\dim(H) \leq 2n^d$. ■

In the next Lemma (whose proof may be found in Appendix A), we bound from above the eigenvalues of Σ.

Lemma 11 *Let $\Sigma = \text{Cov}(P(w))$, where $P(w)$ is defined as in (2). Then*

$$\|\Sigma\|_{op} \leq (4d)! \max_m \{|a_m|\} n^{\frac{d-1}{2}}.$$

4.3. A functional CLT for polynomial processes

Proof [Proof of Theorem 2] Let δ be some small number to be determined later and set $\tilde{X}_k = \Sigma^{-1/2} X_k$ and $\tilde{G}$, the standard Gaussian in H. By Lemma 10, and since Σ and Π_δ commute,

$$
\begin{aligned}
\mathcal{W}_2^2(X_k, G) &\leq \mathcal{W}_2^2(\Pi_\delta^\perp X_k, \Pi_\delta^\perp G) + 8n^d \delta \\
&= \mathcal{W}_2^2(\Sigma^{1/2} \Pi_\delta^\perp \tilde{X}_k, \Sigma^{1/2} \Pi_\delta^\perp \tilde{G}) + 8n^d \delta \leq \|\Sigma\|_{op} \mathcal{W}_2^2(\Pi_\delta^\perp \tilde{X}_k, \Pi_\delta^\perp \tilde{G}) + 8n^d \delta.
\end{aligned}
$$

We focus on the term $\mathcal{W}_2^2(\Pi_\delta^\perp \tilde{X}_k, \Pi_\delta^\perp \tilde{G})$ for which Theorem 9 may be invoked,

$$
\begin{aligned}
\mathcal{W}_2^2(\Pi_\delta^\perp \tilde{X}_k, \Pi_\delta^\perp \tilde{G}) &\leq \frac{\sqrt{\dim(H)}}{k} \mathbb{E}\left[\left\|\Pi_\delta^\perp \Sigma^{-1/2} P(w)\right\|_H^4\right] \\
&\leq \frac{\sqrt{\dim(H)}}{k} \mathbb{E}\left[\|P(w)\|_H^4\right] \left\|\Pi_\delta^\perp \Sigma^{-1}\right\|_{op}^2 \\
&\leq \frac{\sqrt{\dim(H)}}{\delta^2 k} \mathbb{E}\left[\|P(w)\|_H^4\right].
\end{aligned}
$$

In the first inequality, we have used Jensen's inequality on the bound from Theorem 9. Let us estimate $\mathbb{E}\left[\|P(w)\|_H^4\right]$. By definition,

$$
\begin{aligned}
\mathbb{E}\left[\|P(w)\|_H^4\right] = \mathbb{E}\left[\left(\sum_{m=0}^d |a_m| \|w^{\otimes m}\|_m^{2m}\right)^2\right] &\leq \left(\sum_{m=0}^d a_m^2\right)\left(\sum_{m=0}^d \mathbb{E}\left[\|w\|_2^{4m}\right]\right) \\
&\leq \left(\sum_{m=0}^d a_m^2\right)\left(\sum_{m=0}^d (2m)! \left(4\mathbb{E}\left[\|w\|_2^2\right]\right)^{2m}\right) \leq \left(\sum_{m=0}^d a_m^2\right)\left(\sum_{m=0}^d (2m)! (4n)^{2m}\right) \\
&\leq \left(\sum_{m=0}^d a_m^2\right) 16^d (2d)! n^{2d} \leq \max_m\{a_m^2\}(100d)! n^{2d}
\end{aligned}
$$

The first inequality is Cauchy-Schwartz and in the second inequality we have used the fact that $\|w\|_2$ has sub-exponential tails.

Since $\dim(H) \leq 2n^d$, it follows that,

$$
\mathcal{W}_2^2(X_k, G) \leq \|\Sigma\|_{op} \frac{(100d)! n^{\frac{5d}{2}}}{\delta^2 k} \max_m\{a_m^2\} + 8n^d \delta.
$$

We plug the estimate for $\|\Sigma\|_{op}$ from Lemma 11 to deduce:

$$
\mathcal{W}_2^2(X_k, G) \leq \frac{(110d)! n^{3d-0.5}}{\delta^2 k} \max_m\{|a_m|^3\} + 8n^d \delta.
$$

We now take $\delta = \left(\frac{(110d)! n^{2d-0.5} \max_m\{|a_m|^3\}}{k}\right)^{\frac{1}{3}}$ to obtain

$$
\mathcal{W}_2^2(X_k, G) \leq 16 \max_m\{|a_m|\} ((110d)!) \frac{n^{\frac{5d-0.5}{3}}}{k^{\frac{1}{3}}}
$$

To finish the proof, define the Gaussian process $\mathcal{G}$ by $\mathcal{G}(x) = Q(P(x), G)$, and invoke Claim 8. ∎

4.4. An improved rate for tensor powers

Throughout this section we assume that $p(x) = x^d$ for some $d \in \mathbb{N}$. Under this assumption, we improve Theorem 2. This improvement is enabled by two factors:

- A specialized CLT for tensor powers, as proven in (Mikulincer, 2021).

- An improved control on the eigenvalues of Σ, which allows to bypass Lemma 10.

Let us first state the result about approximating tensor powers by Gaussians. Note that for a polynomial p as above, we have the embedding map $P(x) = x^{\otimes d}$. Since the image of P is always a symmetric d-tensor, we allow ourselves to restrict the embedding map P and overload notations, so that $P : \mathbb{R}^n \to \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$. In this case, for $w \sim \mathcal{N}(0, \mathrm{I}_d)$, we have $\Sigma := \mathrm{Cov}(P(w))$, and $X_k := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)$.

Theorem 12 *(Mikulincer, 2021, Theorems 2 and 5) Let the above notations prevail. Then, there exists a Gaussian random vector G, in $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$, such that,*

$$\mathcal{W}_2^2(X_k, G) \leq C_d \|\Sigma\|_{op} \|\Sigma^{-1}\|_{op}^2 \frac{n^{2d-1}}{k},$$

where $C_d = d^{Cd}$, for some universal constant $C > 0$.

Remark that the result in (Mikulincer, 2021) actually deals with the random vector $\sqrt{\Sigma^{-1}} X_k$. Since we care about the un-normalized vector X_k we incur a dependence on $\|\Sigma\|_{op}$. Next we show that the eigenvalues of Σ, may be bounded from below (see Appendix 3 for the proof).

Lemma 13 *Let $\lambda_{\min}(\Sigma)$ stand for the minimal eigenvalue of Σ. Then*

$$\lambda_{\min}(\Sigma) \geq \frac{1}{d!}.$$

We are now in a position to prove Theorem 5.

Proof [Proof of Theorem 5] By combining Lemma 11 and Lemma 13, there exists some numerical constant $C' > 0$, such that

$$\|\Sigma\|_{op} \|\Sigma^{-1}\|_{op}^2 \leq d^{C'd} n^{\frac{d-1}{2}}.$$

Thus, Theorem 12 shows that there exists a Gaussian vector G in $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$, such that

$$\mathcal{W}_2^2(X_k, G) \leq d^{Cd} \frac{n^{2.5d-1.5}}{k},$$

for some other constant $C > 0$. Define the Gaussian process $\mathcal{G}(x) = \langle P(x), G \rangle$, then Lemma 1 shows,

$$\mathcal{W}\mathcal{F}_\infty^2(\mathcal{P}_k p, \mathcal{G}) \leq d^{Cd} \frac{n^{2.5d-1.5}}{k},$$

10

which concludes the proof. When $d = 2$, it is not hard to see that $\|\Sigma\|_{op}$ can be bounded by an absolute constant (see Lemma 16 in the appendix). In this case,

$$\|\Sigma\|_{op}\|\Sigma^{-1}\|_{op}^2 \leq C,$$

which is the reason behind Remark 6. ∎

5. General activations

In this section we consider a general (non-polynomial) activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Our goal is to derive a quantitative CLT for the random process $\mathcal{P}_k\sigma$. Our strategy will be to approximate σ by some polynomial, for which Theorem 2 applies. We set γ to be the law of the standard Gaussian in $\mathbb{R}$. Lemma 1 suggests that, in order to control the remainder in the approximation, it would be beneficial to find a polynomial p, such that p and σ are close in $L^2(\gamma)$.

In $L^2(\gamma)$ there is a distinguished set of polynomials, the so-called Hermite polynomials. Henceforth we denote h_m to be the mth normalized Hermite polynomial,

$$h_m(x) = \frac{(-1)^m}{\sqrt{m!}}\left(\frac{d^m}{dx^m}e^{-\frac{x^2}{2}}\right)e^{\frac{x^2}{2}}.$$

The reader is referred to Janson (1997) for the necessary definitions and proofs pertaining to Hermite polynomials. We will mainly care about the fact that $\{h_m\}_{m=0}^\infty$ forms a complete orthonormal system in $L^2(\gamma)$. Thus, assuming that $\sigma \in L^2(\gamma)$, it may be written as,

$$\sigma = \sum_{m=0}^\infty \hat{\sigma}_m h_m \text{ , where } \hat{\sigma}_m := \int_{\mathbb{R}} \sigma(x)h_m(x)d\gamma(x).$$

Let us also define the remainder function of σ as,

$$R_\sigma(d) = \sum_{m=d+1}^\infty \hat{\sigma}_r^2.$$

If we define the degree d polynomial

$$p_d := \sum_{m=1}^d \hat{\sigma}_m h_m, \tag{3}$$

we then have,

$$\|\sigma - p_d\|_{L^2(\gamma)}^2 \leq R_\sigma(d). \tag{4}$$

With these notations, the main result of this section is:

Theorem 14 *Suppose that $\sigma \in L^2(\gamma)$. Then, there exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that,*

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C'\frac{\max_m |\hat{\sigma}_m|^2}{k^{\frac{1}{6}}} + R_\sigma\left(\frac{\log(k)}{C'\log(n)\log(\log(k))}\right),$$

where $C' > 0$ is a numerical constant.

Before proving the theorem, we first focus on the coefficients of the polynomial p_d, defined in (3), with respect to the standard monomial basis. For this, we write h_m, explicitly (see (Janson, 1997, Chapter 3)) as,

$$h_m(x) = \sqrt{m!} \sum_{j=0}^{\frac{m}{2}} \frac{(-1)^j}{j!(m-2j)!2^j} x^{m-2j}.$$

Write now $p_d = \sum_{m=0}^{d} a_m x^m$ and let us estimate a_m.

Lemma 15 *It holds that*

$$|a_m| \leq \max_i |\hat{\sigma}_i| \frac{2}{\sqrt{m!}} 2^d.$$

Proof We have:

$$
\begin{aligned}
|a_m| &\leq \sum_{i=m}^{d} |\hat{\sigma}_i| \frac{\sqrt{i!}}{m!((i-m)/2)!2^{(i-m)/2}} \\
&\leq \max_i |\hat{\sigma}_i| \sum_{i=m}^{d} \frac{\sqrt{i!}}{m!((i-m)/2)!2^{(i-m)/2}} \\
&\leq \max_i |\hat{\sigma}_i| \sum_{i=m}^{d} \frac{\sqrt{i!}}{m!\sqrt{(i-m)!}} = \max_i |\hat{\sigma}_i| \sum_{i=m}^{d} \frac{1}{\sqrt{m!}} \sqrt{\binom{i}{m}} \\
&\leq \max_i |\hat{\sigma}_i| \frac{1}{\sqrt{m!}} \sum_{i=m}^{d} \binom{i}{m} = \max_i |\hat{\sigma}_i| \frac{1}{\sqrt{m!}} \binom{d+1}{m+1} \leq \max_i |\hat{\sigma}_i| \frac{2}{\sqrt{m!}} 2^d,
\end{aligned}
$$

where the last equality is Pascal's identity. ∎

We may now prove Theorem 14.
Proof [Proof of Theorem 14] Fix d and let $\mathcal{G}$ be the Gaussian process promised by Theorem 2, for p_d. By the triangle inequality,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq 2\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{P}_k p_d) + 2\mathcal{WF}_2^2(\mathcal{P}_k p_d, \mathcal{G}) \leq 2\mathcal{WF}_\infty^2(\mathcal{P}_k\sigma, \mathcal{P}_k p_d) + 2\mathcal{WF}_2^2(\mathcal{P}_k p_d, \mathcal{G}).$$

We now invoke Lemma 1 with (4) to obtain,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{P}_k p_d) \leq \|p_d - \sigma\|_{L^2(\gamma)}^2 \leq R_\sigma(d).$$

For the other term, Theorem 2 along with Lemma 15 imply,

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p_d, \mathcal{G}) \leq \max_i |\hat{\sigma}_i|^2 \cdot d^{Cd} \frac{n^{2d}}{k^{\frac{1}{3}}},$$

for some numerical constant $C > 0$. So,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq 2\max_i |\hat{\sigma}_i|^2 d^{Cd} \frac{n^{2d}}{k^{\frac{1}{3}}} + 2R_\sigma(d).$$

Finally, choose $d = \lceil \frac{\log(k)}{100C\log(n)\log(\log(k))} \rceil$. It can be verified that for any $\delta > 0, \alpha > 0$,

$$d^{Cd} \cdot n^{2d} \leq \log(k)^{\frac{\log(k)}{100\log(\log(k))}} \cdot e^{\frac{\log(k)}{10}} = O(k^{\frac{1}{6}}).$$

This implies the existence of an absolute constant $C' > 0$, for which,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C' \left(\frac{\max_i |\hat{\sigma}_i|^2}{k^{\frac{1}{6}}} + R_\sigma(d) \right).$$

The proof is complete. ∎

References

Kyle Aitken and Guy Gur-Ari. On the asymptotics of wide networks with polynomial activations. *arXiv preprint arXiv:2006.06687*, 2020.

Anders Andreassen and Ethan Dyer. Asymptotics of wide convolutional neural networks. *arXiv preprint arXiv:2008.08675*, 2020.

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

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8141–8150, 2019.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Sergey G. Bobkov. Berry-Esseen bounds and Edgeworth expansions in the central limit theorem for transport distances. *Probab. Theory Related Fields*, 170(1-2):229–262, 2018. ISSN 0178-8051. doi: 10.1007/s00440-017-0756-2. URL https://doi.org/10.1007/s00440-017-0756-2.

Thomas Bonis. Stein's method for normal approximation in Wasserstein distances with application to the multivariate central limit theorem. *Probab. Theory Related Fields*, 178 (3-4):827–860, 2020. ISSN 0178-8051. doi: 10.1007/s00440-020-00989-4. URL https://doi.org/10.1007/s00440-020-00989-4.

Daniele Bracale, Stefano Favaro, Sandra Fortini, and Stefano Peluchetti. Large-width functional asymptotics for deep gaussian neural networks. *arXiv preprint arXiv:2102.10307*, 2021.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures Algorithms*, 49(3):503–532, 2016. ISSN 1042-9832. doi: 10.1002/rsa.20633. URL https://doi.org/10.1002/rsa.20633.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

Thomas A. Courtade, Max Fathi, and Ashwin Pananjady. Existence of Stein kernels under a spectral gap, and discrepancy bounds. *Ann. Inst. Henri Poincaré Probab. Stat.*, 55(2): 777–790, 2019. ISSN 0246-0203. doi: 10.1214/18-aihp898. URL https://doi.org/10.1214/18-aihp898.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*, 2018.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stephane d'Ascoli, Giulio Biroli, Clement Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal Of Statistical Mechanics-Theory And Experiment*, 2020(ARTICLE):023401, 2020.

Surbhi Goel, Sushrut Karmalkar, and Adam Klivans. Time/accuracy tradeoffs for learning a relu with respect to Gussian marginals. In *Advances in Neural Information Processing Systems*, pages 8584–8593, 2019.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10):992, 2019.

Tamir Hazan and Tommi Jaakkola. Steps toward deep kernel methods from infinite neural networks. *arXiv preprint arXiv:1508.05133*, 2015.

Einar Hille. Contributions to the theory of Hermitian series II. the representation problem. *Transactions of the American Mathematical Society*, 47(1):80–94, 1940.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Svante Janson. *Gaussian Hilbert spaces*, volume 129. Cambridge university press, 1997.

Tiefeng Jiang and Danning Li. Approximation of rectangular beta-Laguerre ensembles and large deviations. *J. Theoret. Probab.*, 28(3):804–847, 2015. ISSN 0894-9840. doi: 10.1007/s10959-013-0519-7. URL https://doi.org/10.1007/s10959-013-0519-7.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Takuo Matsubara, Chris J Oates, and François-Xavier Briol. The ridgelet prior: A covariance function approach to prior specification for bayesian neural networks. *arXiv preprint arXiv:2010.08488*, 2020.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

S Mei, A Montanari, and PM Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

Dan Mikulincer. A CLT in Stein's distance for generalized Wishart matrices and higher-order tensors. *International Mathematics Research Notices*, 01 2021. ISSN 1073-7928. doi: 10.1093/imrn/rnaa336. URL https://doi.org/10.1093/imrn/rnaa336. rnaa336.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer, 1996.

Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2012. ISBN 978-1-107-01777-1. doi: 10.1017/CBO9781139084659. URL https://doi.org/10.1017/CBO9781139084659. From Stein's method to universality.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2018.

Abhishek Panigrahi, Abhishek Shetty, and Navin Goyal. Effect of activation functions on the training of overparametrized neural nets. In *International Conference on Learning Representations*, 2019.

Belinda Tzen and Maxim Raginsky. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled Mckean-Vlasov dynamics. *arXiv preprint arXiv:2002.01987*, 2020.

Christopher KI Williams. Computing with infinite networks. In *Advances in neural information processing systems*, pages 295–301, 1997.

Sho Yaida. Non-gaussian processes and neural networks at finite widths. *arXiv preprint arXiv:1910.00019*, 2019.

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

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

Appendix A. Variance estimates for polynomials in Gaussian variables

A.1. An upper bound for the covariance matrix

Proof [Proof of Lemma 11] Let $\sum_I v_I e_I = v \in H$ be a unit vector, we wish to bound $\langle v, \Sigma v \rangle = \mathrm{Var}\left(\langle P(w), v \rangle_H\right)$ from above. Let us denote the degree d polynomial

$$\sum_{i=0}^{m} \sum_{I \in \mathrm{MI}_n(m)} \sqrt{|a_m|} v_I x^I = q(x) = \langle P(x), v \rangle_H.$$

We will prove the claim by induction on d. The case $d = 1$, is rather straightforward to check. For the general case, we will use the Gaussian Poincaré inequality (see (Nourdin and Peccati, 2012, Proposition 1.3.7), for example) to reduce the degree. According to the inequality,

$$\begin{aligned}
\mathrm{Var}\left(\langle P(w), v \rangle_H\right) = \mathrm{Var}\left(q(w)\right) &\leq \mathbb{E}\left[\|\nabla q(w)\|_2^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\left|\frac{d}{dx_i} q(w)\right|^2\right] \\
&= \sum_{i=1}^{n} \mathrm{Var}\left(\frac{d}{dx_i} q(w)\right) + \mathbb{E}\left[\frac{d}{dx_i} q(w)\right]^2. \quad (5)
\end{aligned}$$

Fix $i = 1, \ldots, n$, if $I \in \mathrm{MI}_n(m)$ we denote by $\partial_i I \in \mathrm{MI}_n(m-1)$, to be a multi-index set such that

$$\partial_i I_j = \begin{cases} I_j & \text{if } i \neq j \\ \max(0, I_i - 1) & \text{if } i = j. \end{cases}$$

With this notation, we have,

$$\begin{aligned}
\frac{d}{dx_i} q(w) &= \frac{d}{dx_i} \left(\sum_{m=0}^{d} \sum_{I \in \mathrm{MI}_n(m)} \sqrt{|a_m|} w^I v_I \right) \\
&= \sum_{m=0}^{d} \sum_{I \in \mathrm{MI}_n(m)} \sqrt{|a_m|} I_i w^{\partial_i I} v_I.
\end{aligned}$$

Since $\frac{d}{dx_i} q$ is a polynomial of degree $d - 1$, we thus get by induction,

$$\mathrm{Var}\left(\frac{d}{dx_i} q(w)\right) \leq (4d-4)! \max_m \{|a_m|\} n^{\frac{d-2}{2}} \sum_{m=0}^{d} \sum_{I \in \mathrm{MI}_n(m)} I_i^2 v_I^2.$$

Observe that $I_i \leq d$ and that for every $I \in \mathrm{MI}_n(m)$, there are at most d different indices $i \in [n]$, for which $I_i \neq 0$. Therefore,

$$\sum_{i=1}^{n} \mathrm{Var}\left(\frac{d}{dx_i} q(w)\right) \leq d^2 (4d-4)! \max_m \{|a_m|\} n^{\frac{d-2}{2}} \left(d \sum_I v_I^2 \right) \leq (4d-1)! \max_m \{|a_m|\} n^{\frac{d-2}{2}}. \quad (6)$$

Furthermore, if for some $j \in [n]$, $\partial_i I_j$ is odd, then $\mathbb{E}\left[w^{\partial_i I}\right] = 0$. Otherwise,

$$|\mathbb{E}\left[w^{\partial_i I}\right]| \leq |\mathbb{E}\left[w_1^{d-1}\right]| \leq \sqrt{d!}.$$

It is easy to verify that the size of the following set,

$$A_i = \{I \in \cup_{m=0}^d \mathrm{MI}_n(m) | I_i \mathbb{E}[w^{\partial_i I}] \neq 0\},$$

is at most $(2n)^{\frac{d-1}{2}}$. Thus, since there are at most $(2n)^{\frac{d-1}{2}}$ elements which do not vanish, Cauchy-Schwartz's inequality shows,

$$\mathbb{E}\left[\frac{d}{dx_i}q(w)\right]^2 \leq d^2 \max_m\{|a_m|\}\mathbb{E}\left[\sum_{m=0}^d \sum_{I \in \mathrm{MI}_n(m)} w^{\partial_i I} v_I\right]^2$$

$$\leq (4d-1)! \max_m\{|a_m|\}n^{\frac{d-1}{2}} \sum_{I \in A_i} v_I^2.$$

Note that if $I \in A_i$, then necessarily I_i is odd. In this case, it follows that for $j \neq i$, $\mathbb{E}\left[w^{\partial_j I}\right] = 0$. Hence, $A_i \cap A_j = \emptyset$, and

$$\sum_{i=1}^n \mathbb{E}\left[\frac{d}{dx_i}q(w)\right]^2 \leq (4d-1)! \max_m\{|a_m|\}n^{\frac{d-1}{2}} \sum_{i=1}^n \sum_{I \in A_i} v_I^2$$

$$\leq (4d-1)! \max_m\{|a_m|\}n^{\frac{d-1}{2}} \sum_I v_I^2 = (4d-1)! \max_m\{|a_m|\}n^{\frac{d-1}{2}}. \qquad (7)$$

We now plug (6) and (7) into (5) to obtain

$$\mathrm{Var}\left(\langle P(w), v\rangle_H\right) \leq (4d)! \max_m\{|a_m|\}n^{\frac{d-1}{2}}.$$

$\blacksquare$

Remark that, up to the multiplicative dependence on d, this bound is generally sharp. As an example, when $d = 2\ell - 1$ is odd, one can consider the degree d polynomial,

$$q(x) = \frac{1}{n^{\ell/2}} \sum_{i_1,\dots i_\ell=1}^n x_{i_1} x_{i_2}^2 \dots x_{i_\ell}^2.$$

For this polynomial it may be verified that $\mathrm{Var}(q(w)) = \Omega(n^{\ell-1}) = \Omega\left(n^{\frac{d-1}{2}}\right)$.

A.2. A lower bound for the covariance matrix of a tensor power

Proof Let $v \in \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$ be a unit vector. We can thus write $v = \sum_{|I|=d} v_I e_I$, with $\sum_{I \in \mathrm{MI}_n(d)} v_I^2 = 1$. Define the degree d homogeneous polynomial $q : \mathbb{R}^n \to \mathbb{R}$, by $q(x) = \sum_{I \in \mathrm{MI}_n(d)} v_I x^I$. In this case we have $\langle v, P(w)\rangle = q(w)$, and it will be enough to show,

$$\mathrm{Var}\left(\langle v, P(w)\rangle\right) = \mathrm{Var}\left(q(w)\right) \geq \frac{1}{d!}.$$

18

We will use the variance expansion for functions of Gaussian vectors, which can be found at (Nourdin and Peccati, 2012, Proposition 1.5.1). According to this expansion, for any smooth enough function $f : \mathbb{R}^n \to \mathbb{R}$,

$$\mathrm{Var}(f(w)) = \sum_{m=1}^{\infty} \frac{\|\mathbb{E}\left[\nabla^m f(w)\right]\|_m^2}{m!}. \tag{8}$$

Here $\nabla^m f$ is the mth total derivative of f, which we regard as an element in $(\mathbb{R}^n)^{\otimes m}$. In particular, we have,

$$\mathrm{Var}\left(q(w)\right) \geq \frac{\left\|\mathbb{E}\left[\nabla^d q(w)\right]\right\|_d^2}{d!}.$$

Now, if $I \neq J$ are two multi-subsets of $[n]$, with $I, J \in \mathrm{MI}_n(d)$, we have

$$\frac{d}{dx^I} x^J = 0 \text{ and } \frac{d}{dx^I} x^I = I!.$$

So, since $\nabla^d f = \{\frac{d}{dx^I} q\}_{I \in \mathrm{MI}_n(d)}$,

$$\|\mathbb{E}\left[\nabla^d f(w)\right]\|_d^2 = \sum_{I \in \mathrm{MI}_n(d)} I! v_I^2 \geq 1,$$

and

$$\mathrm{Var}\left(q(w)\right) \geq \frac{1}{d!},$$

as required. ∎

A.3. Dimension-free covariance estimates for quadratic tensor powers

When considering the polynomial $p(x) = x^2$, we can strictly improve upon Lemma 11 and obtain dimension-free bounds. As noted in the proof of Theorem 5, this explains Remark 6.

Lemma 16 *Suppose that $d = 2$. Then,*

$$\|\Sigma\|_{op} \leq 1.$$

Proof As in the proof of Lemma 13, let $v = \sum_{i,j=1}^{n} v_{i,j} e_i \otimes e_j$, with $\sum v_{i,j}^2 = 1$. Define $q(x) = \sum_{i,j=1}^{n} v_{i,j} x_i x_j$. It will suffice to bound $\mathrm{Var}(q(w))$ from above. Since q is a quadratic polynomial, the variance decomposition (8) gives,

$$\mathrm{Var}(q(w)) = \|\mathbb{E}\left[\nabla q(w)\right]\|^2 + \frac{1}{2}\|\mathbb{E}\left[\nabla^2 q(w)\right]\|^2.$$

for $i \in [n]$, we have $\frac{d}{dx_i} q(w) = \sum_{j=1}^{n} (1 + \delta_{i,j}) v_{i,j} w_j$. So, $\mathbb{E}\left[\nabla q(w)\right] = 0$. On the other hand,

$$\|\mathbb{E}\left[\nabla q(w)\right]\|^2 = \sum_{i,j=1}^{n} \mathbb{E}\left[\frac{d^2}{dx_i dx_j} q(w)\right]^2 \leq 2 \sum_{i=1}^{n} v_{i,j}^2 = 2,$$

and the claim is proven. ∎

Appendix B. Normal approximations for non-polynomial activations

B.1. ReLU activation; Proof of Theorem 3

In this section we specialize Theorem 14 to the ReLU activation $\psi(x) := \max(0, x)$. The calculation of $\hat{\psi}_m$ may be found in Daniely et al. (2016); Goel et al. (2019). We repeat it here for completeness.

Lemma 17 *Let $m \in \mathbb{N}$. Then,*

$$|\hat{\psi}_m| = \begin{cases} \frac{1}{\sqrt{2}} & m = 1 \\ 0 & m > 1 \ and \ odd \ . \\ \frac{(m-3)!!}{\sqrt{\pi}\sqrt{m!}} & otherwise \end{cases} \tag{9}$$

In particular, $|\hat{\psi}_m| \leq \frac{1}{m^{\frac{3}{2}}}$, and

$$R_\psi(d) \leq \frac{1}{d^2}.$$

Proof Note that once (9) is established the rest of the proof is trivial. Thus, let us focus on calculating $\hat{\psi}_m$. We will use the following formula for the derivative of Hermite polynomials,

$$h'_m(x) = \sqrt{m} h_{m-1}(x). \tag{10}$$

Using this, we have, with an application of integration by parts,

$$\begin{aligned} \hat{\psi}_m = \int_{\mathbb{R}} h_m(x)\psi(x)d\gamma(x) &= \int_{x>0} h_m(x)x d\gamma(x) = \frac{h_m(0)}{\sqrt{2\pi}} - \int_{x>0} h'_m(x)d\gamma(x) \\ &= \frac{h_m(0)}{\sqrt{2\pi}} - \sqrt{m} \int_{x>0} h_{m-1}(x)d\gamma(x) \\ &= \frac{h_m(0)}{\sqrt{2\pi}} + (-1)^m \sqrt{\frac{m}{2\pi(m-1)!}} \int_{x>0} \frac{d^{m-1}}{dx^{m-1}} e^{-\frac{x^2}{2}}(x)dx \\ &= \frac{h_m(0)}{\sqrt{2\pi}} + (-1)^m \sqrt{\frac{m}{2\pi(m-1)!}} \frac{d^{m-2}}{dx^{m-2}} e^{-\frac{x^2}{2}}(0) \\ &= \frac{h_m(0) + \sqrt{\frac{m}{(m-1)}} h_{m-2}(0)}{\sqrt{2\pi}}. \end{aligned}$$

For $h_m(0)$, the following explicit formula holds:

$$h_m(0) = \begin{cases} 0 & \text{for } m \text{ odd} \\ (-1)^{m/2} \frac{(m-1)!!}{\sqrt{m!}} & \text{for } m \text{ even} \end{cases}.$$

In this case, for m even,

$$\sqrt{\frac{m}{(m-1)}} h_{m-2}(0) = (-1)^{m/2-1} \sqrt{\frac{m}{m-1}} \frac{(m-3)!!}{\sqrt{(m-2)!}} = (-1)^{m/2-1} \frac{m(m-3)!!}{\sqrt{(m-1)!}},$$

and (9) follows. ∎

Theorem 3 follows immediately, by plugging the above Lemma into Theorem 14.
Proof [Proof of Theorem 3] From Lemma 17 we see that $\max_i |\hat{\psi}_i| \leq 1$, and so coupled with Theorem 14, we get

$$\mathcal{W}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C\left(\frac{1}{k^{\frac{1}{6}}} + \left(\frac{\log(n)\log(\log(k))}{\log(k)}\right)^2\right).$$

It is now enough to observe,

$$\frac{1}{k^{\frac{1}{6}}} = O\left(\left(\frac{\log(n)\log(\log(k))}{\log(k)}\right)^2\right).$$

∎

B.2. Hyperbolic tangent activation; Proof of Theorem 4

Let us now consider the function $\tanh(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}}$ as an activation. Since it is smooth, we should expect it to have better polynomial approximations than the ReLU. This will lead to a faster convergence rate along the CLT. An explicit expression for $\widehat{\tanh}_m$ may be difficult to find. However, one may combine the smoothness of tanh with a classical result of Hille (Hille (1940)) in order to bound the coefficients from above.

This calculation was done in Panigrahi et al. (2019), where it was shown that for the derivative $|\widehat{\tanh}'_m| \leq e^{-C\sqrt{m}}$, where $C > 0$, does not depend on m. We now extend this result to tanh.

Lemma 18 *Let $m \geq 0$. It holds that*

$$|\widehat{\tanh}_m| \leq e^{-C\sqrt{m}},$$

for some absolute constant $C > 0$.

Proof Since $|\widehat{\tanh}'_m| \leq e^{-C\sqrt{m}}$, Hille's result ((Hille, 1940, Theorem 1)) shows that we have the point-wise equality,

$$\tanh'(x) = \sum_{m=0}^{\infty} \widehat{\tanh}'_m h_m(x).$$

We now use (10), and integrate the series, term by term, so that

$$\tanh(x) = \sum_{m=1}^{\infty} \frac{\widehat{\tanh}'_{m-1}}{\sqrt{m}} h_m(x).$$

So, $\widehat{\tanh}_m = \frac{\widehat{\tanh}'_{m-1}}{\sqrt{m}}$, which proves the claim. ∎

From the lemma, we get that there is some absolute constant $C > 0$, such that $R_{\mathrm{tanh}}(d) \leq e^{-C\sqrt{d}}$. This allows us to prove Theorem 4.

Proof [Proof of Theorem 4] From Lemma 18 along with Theorem 14, we get

$$\mathcal{W}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C \left(\frac{1}{k^{\frac{1}{6}}} + \exp\left(-\frac{1}{C}\sqrt{\frac{\log(k)}{\log(n)\log(\log(k))}} \right) \right).$$

As before, the claim follows since,

$$\frac{1}{k^{\frac{1}{6}}} = O\left(\exp\left(-\frac{1}{C}\sqrt{\frac{\log(k)}{\log(n)\log(\log(k))}} \right) \right).$$

■