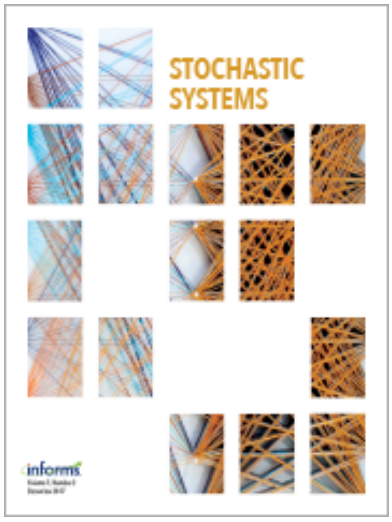

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Normal Approximation of Random Gaussian Neural Networks

Nicola Apollonio, Daniela De Canditiis, Giovanni Franzina, Paola Stolfi, Giovanni Luca Torrisi

To cite this article:
Nicola Apollonio, Daniela De Canditiis, Giovanni Franzina, Paola Stolfi, Giovanni Luca Torrisi (2025) Normal Approximation of Random Gaussian Neural Networks. Stochastic Systems 15(1):88-110. https://doi.org/10.1287/stsy.2023.0033

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 15, No. 1, March 2025, pp. 88–110
ISSN 1946-5238 (online)

Normal Approximation of Random Gaussian Neural Networks

Nicola Apollonio,[a] Daniela De Canditiis,[a] Giovanni Franzina,[a] Paola Stolfi,[a,*] Giovanni Luca Torrisi[a]

[a] Istituto per le Applicazioni del Calcolo "M. Picone", Consiglio Nazionale delle Ricerche, 00185 Rome, Italy
*Corresponding author
Contact: nicola.apollonio@cnr.it, https://orcid.org/0000-0001-6089-1333 (NA); d.decanditiis@iac.cnr.it, https://orcid.org/0000-0002-3022-3411 (DDC); giovanni.franzina@cnr.it, https://orcid.org/0000-0002-8274-3224 (GF); p.stolfi@iac.cnr.it, https://orcid.org/0000-0003-3688-5464 (PS); g.torrisi@iac.rm.cnr.it, https://orcid.org/0000-0001-6953-5407 (GLT)

Received: July 12, 2023
Revised: January 17, 2024
Accepted: September 18, 2024
Published Online in Articles in Advance: October 23, 2024

https://doi.org/10.1287/stsy.2023.0033

Abstract. In this paper, we provide explicit upper bounds on some distances between the (law of the) output of a random Gaussian neural network and (the law of) a random Gaussian vector. Our main results concern deep random Gaussian neural networks with a rather general activation function. The upper bounds show how the widths of the layers, the activation function, and other architecture parameters affect the Gaussian approximation of the output. Our techniques, relying on Stein's method and integration by parts formulas for the Gaussian law, yield estimates on distances that are indeed integral probability metrics and include the convex distance. This latter metric is defined by testing against indicator functions of measurable convex sets and so allows for accurate estimates of the probability that the output is localized in some region of the space, which is an aspect of a significant interest both from a practitioner's and a theorist's perspective. We illustrated our results by some numerical examples.

Funding: This research was supported by the European Union's Horizon 2020 research project WARIFA under grant agreement no. 101017385, by the PRIN project 2022 "Variational Analysis of Complex Systems in Materials Science, Physics and Biology" (CUP B53D23009290006), and by the INdAM project "Modelli ed Algoritmi per dati ad elevata dimensionalità" (CUP E53C23001670001).

Keywords: Gaussian approximation • neural networks • Stein's method

1. Introduction

This work is part of the literature studying random neural networks (NNs for short), that is, NNs whose biases and weights are random variables. In the context of modern deep learning, the interest in these types of networks is twofold; on the one hand, they naturally constitute a prior in a Bayesian approach, and on the other hand they may represent the initialization of gradient flows in empirical risk minimization. See Roberts et al. (2022) for a general reference on the subject.

Within the boundaries of this topic, many contributions in the literature have been handling the asymptotic Gaussianity of random NNs, as the number of neurons in the hidden layers tends to infinity. A seminal paper is Neal (1996), where the output of a shallow (i.e., having a single hidden layer) random NN, viewed as a stochastic process on the sphere, is shown to converge to a Gaussian process, as the number of neurons in the hidden layer grows large. From that point onward, many sophisticated results have been published for deep (i.e., having more than one hidden layer) random NNs at the large width limit. Early contributions in this direction were from Lee et al. (2018), Matthews et al. (2018), Yaida (2019), and Yang (2019). These achievements were extended in Hanin (2023), where it was proved that the output of a deep random NN, with Gaussian biases and weights, viewed as a random element on the space of continuous functions on a compact set, converges to a Gaussian process as the number of neurons in all the hidden layers tends to infinity. Large and moderate deviations of the output of a deep random NN, with Gaussian biases and weights, were studied in Macci et al. (2024).

Recently, the problem of the quantitative Gaussian approximation of the output of a random NN has received a lot of attention. For instance, Eldan et al. (2021), exploiting Wasserstein distances, provided quantitative versions of the results in Neal (1996) when the activation function was polynomial, ReLU, and hyperbolic tangent. We emphasize that the shallow random NN model considered in Neal (1996) and Eldan et al. (2021) has weights

on the outer layer given by Rademacher random variables and weights on the inner layer distributed according to the Gaussian law (see Remark 3.4). Slightly different models were investigated in Klukowski (2022) and Cammarota et al. (2024). Indeed, as the number of neurons on the hidden layer grew large, Klukowski (2022) provided a quantitative functional central limit theorem for a shallow random NN model with input variables still on the sphere and weights on the outer layer still given by Rademacher random variables but weights on the inner layer uniformly distributed on the sphere, whereas as the number of neurons on the hidden layer grew large, Cammarota et al. (2024) gave quantitative functional central limit theorems for a shallow random NN model with weights on outer and inner layers distributed according to the Gaussian law and again input variables on the sphere. A quantitative functional central limit theorem for deep random NNs, with input variables on the sphere and Lipschitz continuous activation functions, was proved in Balasubramanian et al. (2024); this paper shares with our work the idea to apply Stein's method for the Gaussian approximation in the context of deep NNs, albeit in a different mathematical setting.

A significant achievement is provided in Basteri and Trevisan (2024), where, for the first time in the literature, a quantitative proof of the Gaussian behavior of the output of a deep random Gaussian NN (i.e., a random NN whose biases and weights are Gaussian distributed) with a Lipschitz continuous activation function was given; in Basteri and Trevisan (2024), the distance from Gaussianity was measured by means of the 2-Wasserstein metric, which comes from the Monge-Kantorovich problem with quadratic cost. As far as shallow random Gaussian NNs with univariate output is concerned, we mention Bordino et al. (2024), which provided quantitative bounds on the Kolmogorov, the total variation and the 1-Wasserstein distances between the output and a Gaussian random variable, when the activation function was sufficiently smooth and had a sub-polynomial growth. A special mention is deserved for the independently written paper Favaro et al. (2023), where Stein's method was used to obtain tight probabilistic bounds for various distances between the output (and its derivatives) of a deep random Gaussian NN and a Gaussian random vector. We refer the reader to Remarks 4.2, 6.2, and 6.4 for comparisons between our results and the corresponding achievements in Favaro et al. (2023).

The main contribution of our paper concerns the Gaussian approximation of the output of deep random Gaussian NNs in the convex and 1-Wasserstein distances under mild assumptions on the activation function (which, differently from Basteri and Trevisan (2024), can be non-Lipschitz). A specialization of these results clearly provides approximations for shallow random Gaussian NN with single input and real valued (i.e., univariate) output. However, in this specific case we furnish direct proofs, which (for various technical reasons) give the same rates under more general assumptions on the activation function.

For shallow random Gaussian NNs with univariate output, combining the Stein method for the Gaussian approximation with the integration by parts formula for the Gaussian law, we provide explicit bounds for the Kolmogorov, the total variation and the 1-Wasserstein distances between the output and a Gaussian random variable, under a minimal assumption on the activation function (see Theorem 4.1). Remarkably, we obtain the same rate of convergence as in Bordino et al. (2024) as the number of neurons in the hidden layer grows large, our constants being presumably better than the ones in Bordino et al. (2024) (see Table 1). For deep random Gaussian NNs, the novelty of our results is that we measure the error in the Gaussian approximation of the output in terms of the convex distance (see Theorem 6.1) and of the 1-Wasserstein distance (see Theorem 6.3) for a class of activation functions that strictly includes the family of Lipschitz continuous functions if either the biases are non null or the activation function vanishes at 0 (see Proposition 5.2). Remarkably, for both the convex and the 1-Wasserstein distances the rate of convergence that we obtain is of the same order as the one in Basteri and Trevisan (2024) because the number of neurons in all the hidden layers tend to infinity. The proofs of Theorems 6.1 and 6.3 are based on the Stein method for the multivariate Gaussian approximation and the integration by parts formula for the multivariate Gaussian law. The presence of more than one hidden layer complicates the derivations of the bounds, which rely on a key estimate for the L^2-distance between the so-called collective observables and their limiting values (see Theorem 5.1). We emphasize that, when considering the convex distance, an expedient tool is provided by a smoothing lemma that we borrow from Schulte and Yukich (2019).

Table 1. Values of the Constants Given in Theorems 3.1 and 4.1

	$d_{TV}(z^{(2)},z)$	$d_K(z^{(2)},z)$	$d_{W_1}(z^{(2)},z)$
Theorem 3.1	5.05	2.52	2.01
Theorem 4.1	1.68	0.84	0.67

Note. Here, $\sigma(x) = x^3$, $\gamma = 3$, $r_2 = 1$, $r_1 = 6$, $L = 1$, $C_b = C_W = 1$, and $x = 1$.

It is well known that localizing the output of a random NN, that is, having a control over the probability that the output lies in a region of the space (belonging to a large class of measurable sets), is of valuable interest for practitioners. From a theorist's point of view, the output distribution of a NN is often analytically untractable, and computing the probability that the output belongs to some measurable set results in performing a "heroic" mathematical integration (see, e.g., Roberts et al. 2022, p. 49). Our Theorems 4.1(ii) and 6.1 offer some insights into the localization problem in a simple and efficient way for both the univariate output of a shallow random Gaussian NN and the output of a deep random Gaussian NN, respectively. We refer the reader to Section 7 for some numerical illustrations of this issue.

The paper is organized as follows. In Section 2, we introduce our toolkit such as the integral probability metrics considered in the paper and some preliminaries on the Stein method. In Section 3, first we introduce all of the NNs considered in this work and then give a brief overview of the main results in Basteri and Trevisan (2024) and Bordino et al. (2024), comparing them with our achievements. In Section 4, we give upper bounds for the Kolmogorov, the total variation and the 1-Wasserstein distances between the (univariate) output of a shallow random Gaussian NN and a Gaussian random variable. In Section 5, we prove the aforementioned key estimate on the L^2-distance between the collective observables and their limiting values. In Section 6, we furnish explicit upper bounds on the convex and the 1-Wasserstein distances between the output of a deep random Gaussian NN and a Gaussian random vector. Finally, in Section 7, we present some numerical illustrations concerning the above-mentioned issue of the output localization.

2. Preliminaries

In the present section, we introduce some notation, and we recall some results that will be of use throughout the paper.

2.1. Distances Between Probability Measures

In this paper, we consider various distances between probability measures on $\mathbb{R}^d$, $d \in \mathbb{N} := \{1, 2, \dots\}$: the total variation distance, the convex distance, the Komogorov distance, and the p-Wasserstein distances. Hereon, the symbol $\|\cdot\|_d$ denotes the Euclidean norm on $\mathbb{R}^d$.

Definition 2.1. The total variation distance between the laws of two $\mathbb{R}^d$-valued random vectors $\mathbf{X}$ and $\mathbf{Y}$, written $d_{TV}(\mathbf{X}, \mathbf{Y})$, is given by

$$d_{TV}(\mathbf{X}, \mathbf{Y}) := \sup_{B \in \mathcal{B}(\mathbb{R}^d)} |\mathbb{P}(\mathbf{X} \in B) - \mathbb{P}(\mathbf{Y} \in B)|,$$

where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-field on $\mathbb{R}^d$.

Definition 2.2. The convex distance between the laws of two $\mathbb{R}^d$-valued random vectors $\mathbf{X}$ and $\mathbf{Y}$, written $d_c(\mathbf{X}, \mathbf{Y})$, is given by

$$d_c(\mathbf{X}, \mathbf{Y}) := \sup_{C \in \mathcal{C}_d} |\mathbb{P}(\mathbf{X} \in C) - \mathbb{P}(\mathbf{Y} \in C)|,$$

where $\mathcal{C}_d$ denotes the collection of all Borel convex sets in $\mathbb{R}^d$.

Definition 2.3. The Kolmogorov distance between the laws of two $\mathbb{R}^d$-valued random vectors $\mathbf{X} = (X_1, \dots, X_d)$ and $\mathbf{Y} = (Y_1, \dots, Y_d)$, written $d_K(\mathbf{X}, \mathbf{Y})$, is given by

$$d_K(\mathbf{X}, \mathbf{Y}) := \sup_{\mathbf{y}=(y_1, \dots, y_d) \in \mathbb{R}^d} |\mathbb{P}(X_1 \le y_1, \dots, X_d \le y_d) - \mathbb{P}(Y_1 \le y_1, \dots, Y_d \le y_d)|.$$

Definition 2.4. For $p \in [1, +\infty)$, the p-Wasserstein distance between the laws of two $\mathbb{R}^d$-valued random vectors $\mathbf{X}$ and $\mathbf{Y}$, written $d_{W_p}(\mathbf{X}, \mathbf{Y})$, is given by

$$d_{W_p}(\mathbf{X}, \mathbf{Y}) := \inf_{(\mathbf{U}, \mathbf{V}) \in C_{(\mathbf{X}, \mathbf{Y})}} \mathbb{E}[\|\mathbf{U} - \mathbf{V}\|_d^p]^{1/p},$$

where $C_{(\mathbf{X}, \mathbf{Y})}$ is the family of all the couplings of $\mathbf{X}$ and $\mathbf{Y}$, that is, the family of all random vectors $(\mathbf{U}, \mathbf{V})$ such that $\mathbf{U}$ is distributed as $\mathbf{X}$ and $\mathbf{V}$ is distributed as $\mathbf{Y}$.

Clearly, by Jensen's inequality, $d_{W_1} \le d_{W_2}$, and it follows directly by the definitions that $d_K \le d_c \le d_{TV}$. Furthermore, for all $s = TV, c, K, W_p$, if $d_s(\mathbf{Y}_n, \mathbf{Y}) \to 0$, as $n \to +\infty$, where $\mathbf{Y}_n$, $n \in \mathbb{N}$, and $\mathbf{Y}$ are random vectors with values in $\mathbb{R}^d$, then $\mathbf{Y}_n$ converges in law to $\mathbf{Y}$, as $n \to +\infty$ (see, e.g., Villani 2009 and Nourdin and Peccati 2012).

In view of the Kantorovich-Rubinstein duality (see theorem 5.10 and equation (5.11) in Villani 2009), the 1-Wasserstein distance between the laws of two $\mathbb{R}^d$-valued random vectors $\mathbf{X}$ and $\mathbf{Y}$ such that $\max\{\mathbb{E}\|\mathbf{X}\|_d, \mathbb{E}\|\mathbf{Y}\|_d\} < \infty$ satisfies the relation

$$d_{W_1}(\mathbf{X},\mathbf{Y}) = \sup_{g\in\mathcal{L}_d(1)} |\mathbb{E}[g(\mathbf{X})] - \mathbb{E}[g(\mathbf{Y})]|, \tag{1}$$

where $\mathcal{L}_d(1)$ is the collection of Lipschitz continuous functions with Lipschitz constant less than or equal to 1.

Because d_c is defined by testing against indicator functions of Borel convex sets rather than arbitrary Borel sets, the convex distance can be expected to be estimated more easily than the total variation distance; moreover, the convex distance looks more flexible than the Kolmogorov distance; for example, it enjoys a number of invariance properties not satisfied by d_K (see Benktus 2003).

As for the relation between the convex distance and the optimal transport metric d_{W_1}, it turns out that the convex distance to a fixed centered Gaussian law is bounded from above by a multiple of the square root of the 1-Wasserstein distance. More precisely, one has the following Proposition 2.5, which is proved in Nourdin et al. (2022).

Here and henceforth, we denote by $\mathbf{N}_\Sigma = ((N_\Sigma)_1,\dots,(N_\Sigma)_d)$, $d\in\mathbb{N}$, a centered Gaussian vector with invertible covariance matrix $\Sigma = (\Sigma_{ij})_{1\le i,j\le d}$.

Proposition 2.5. *For any d-dimensional random vector* $\mathbf{Y}$*, we have*

$$d_c(\mathbf{Y},\mathbf{N}_\Sigma) \le 2\sqrt{2}\Gamma(\Sigma)^{1/2} d_{W_1}(\mathbf{Y},\mathbf{N}_\Sigma)^{1/2},$$

where $\Gamma(\Sigma)$ *is the constant defined by*

$$\Gamma(\Sigma) := \sup_{Q,\,\epsilon>0} \frac{\mathbb{P}(\mathbf{N}_\Sigma\in Q^\epsilon) - \mathbb{P}(\mathbf{N}_\Sigma\in Q)}{\epsilon}, \tag{2}$$

where Q *ranges over all the Borel measurable convex subsets of* $\mathbb{R}^d$*, and* Q^ϵ *denotes the set of all elements of* $\mathbb{R}^d$ *whose Euclidean distance from* Q *does not exceed* ϵ*.*

We note that $\Gamma(\Sigma)$ is an isoperimetric constant that satisfies the following relation (see Nazarov 2004):

$$e^{-\frac{5}{4}} d^{1/4} \le \Gamma(\Sigma) \le (2\pi)^{-\frac{1}{4}} d^{1/4}. \tag{3}$$

2.2. The One-Dimensional Stein Equation

Throughout this paper, we denote by $\mathcal{N}(\mu,\eta)$ the one-dimensional Gaussian law with mean $\mu\in\mathbb{R}$ and variance $\eta>0$ and let $Z\sim\mathcal{N}(0,1)$.

The celebrated Stein equation for the one-dimensional Normal approximation Stein (1972) is given by

$$g(w) - \mathbb{E}g(Z) = f_g'(w) - wf_g(w), \tag{4}$$

where $g:\mathbb{R}\to\mathbb{R}$ is a measurable function such that $\mathbb{E}|g(Z)|<\infty$, and $f_g:\mathbb{R}\to\mathbb{R}$ is unknown. The following lemma holds; see, for example, proposition 3.2.2, theorem 3.3.1, theorem 3.4.2, and proposition 3.5.1 in Nourdin and Peccati (2012). See Chen et al. (2011) for an introduction on the Stein method.

Hereon, for a Lipschitz continuous function $g:\mathbb{R}^d\to\mathbb{R}$, we denote by $\mathrm{Lip}(g)$ the Lipschitz constant of g, and for a function $f:\mathbb{R}^d\to\mathbb{R}$ we denote by $\|f\|_\infty$ the supremum norm of f.

Lemma 2.6. *The following claims hold:*

i. *For any* $y\in\mathbb{R}$*, the Stein Equation* (4) *with* $g(w) := \mathbf{1}_{(-\infty,y]}(w)$ *has a unique solution* f_g *and* $\|f_g'\|_\infty \le 1$.

ii. *Let* $g:\mathbb{R}\to[0,1]$ *be a measurable function. Then, there exists a unique solution* f_g *of the Stein Equation* (4) *and* $\|f_g'\|_\infty \le 2$.

iii. *Let* $g:\mathbb{R}\to\mathbb{R}$ *be a Lipschitz continuous function. Then, there exists a unique solution* f_g *of the Stein Equation* (4) *and* $\|f_g'\|_\infty \le \mathrm{Lip}(g)\sqrt{2/\pi}$.

2.3. The Multidimensional Stein Equation

Throughout this paper, given a sufficiently smooth function $f:\mathbb{R}^d\to\mathbb{R}$, we define

$$\partial^n_{i_1 i_1\dots i_n} f(x_1,\dots,x_d) := \frac{\partial^n f}{\partial x_{i_1}\dots\partial x_{i_n}}(x_1,\dots,x_d).$$

Let $\mathcal{M}_{d\times d}(\mathbb{R})$, $d\in\mathbb{N}$, be the set of $d\times d$ real matrices. For a function $f\in C^2(\mathbb{R}^d)$, we denote by $\mathrm{Hess}\, f(\mathbf{y})\in\mathcal{M}_{d\times d}(\mathbb{R})$ the Hessian matrix of f at $\mathbf{y}\in\mathbb{R}^d$ and by $\|\cdot\|_{op}$ the operator norm on $\mathcal{M}_{d\times d}(\mathbb{R})$, that is, for any $\Gamma\in\mathcal{M}_{d\times d}(\mathbb{R})$, $\|\Gamma\|_{\mathrm{op}} := \sup_{\mathbf{y}:\,\|\mathbf{y}\|_d=1}\|\Gamma\mathbf{y}\|_d$. We consider the Hilbert-Schmidt inner product and the Hilbert-Schmidt norm on $\mathcal{M}_{d\times d}(\mathbb{R})$, which are defined, respectively, by

$$\langle\Gamma,\Psi\rangle_{H.S.} := \mathrm{Tr}(\Gamma\Psi^\top) = \sum_{i,j=1}^{d}\Gamma_{ij}\Psi_{ij} \text{ and } \|\Gamma\|_{H.S.} = \sqrt{\langle\Gamma,\Gamma\rangle_{H.S.}}$$

for every pair of matrices $\Gamma=(\Gamma_{ij})_{1\le i,j\le j}$ and $\Psi=(\Psi_{ij})_{1\le i,j\le d}$, where the symbols $\mathrm{Tr}(\Gamma)$ and $\Gamma^\top$ denote, respectively, the trace and the transpose of the matrix Γ.

The Stein equation for multivariate Normal approximation is defined as

$$g(\mathbf{y})-\mathbb{E}[g(\mathbf{N}_\Sigma)] = \langle\mathbf{y},\nabla f_g(\mathbf{y})\rangle_d - \langle\Sigma,\mathrm{Hess}\, f_g(\mathbf{y})\rangle_{H.S.},\quad \mathbf{y}\in\mathbb{R}^d \tag{5}$$

where $g:\mathbb{R}^d\to\mathbb{R}$ is given, f_g is unknown, Σ is the (invertible) covariance matrix of the centered Gaussian vector $\mathbf{N}_\Sigma$, and the symbol $\langle\cdot,\cdot\rangle_d$ denotes the inner product in $\mathbb{R}^d$.

The following lemmas provide solutions to Stein's Equation (5) under different assumptions on g.

Lemma 2.7. *Let $g\in\mathcal{L}_d(1)$. Then, the function*

$$f_g(\mathbf{y}) := \int_0^\infty \mathbb{E}[g(\mathbf{N}_\Sigma)-g(\mathrm{e}^{-t}\mathbf{y}+\sqrt{1-\mathrm{e}^{-2t}}\mathbf{N}_\Sigma)]\,\mathrm{d}t,\quad \mathbf{y}\in\mathbb{R}^d$$

is such that $f_g\in C^2(\mathbb{R}^d)$, f_g solves (5), f_g satisfies

$$\|\partial_i f_g\|_\infty \le 1,\quad \text{for any } i=1,\dots,d, \tag{6}$$

and

$$\sup_{\mathbf{y}\in\mathbb{R}^d}\|\mathrm{Hess}\, f_g(\mathbf{y})\|_{H.S.} \le \sqrt{d}\|\Sigma^{-1}\|_{op}\|\Sigma\|_{op}^{1/2}.$$

Lemma 2.8. *For $g:\mathbb{R}^d\to\mathbb{R}$ measurable and bounded, define the smoothed function*

$$g_t(\mathbf{y}) := \mathbb{E}[g(\sqrt{t}\mathbf{N}_\Sigma+\sqrt{1-t}\mathbf{y})],\quad \mathbf{y}\in\mathbb{R}^d, \tag{7}$$

where $t\in(0,1)$ is a smoothing parameter. Then:

i. *For any $t\in(0,1)$, the function*

$$f_{t,g}(\mathbf{y}) := \frac{1}{2}\int_t^1\frac{1}{1-s}\mathbb{E}[g(\sqrt{s}\mathbf{N}_\Sigma+\sqrt{1-s}\mathbf{y})-g(\mathbf{N}_\Sigma)]\,\mathrm{d}s,\quad \mathbf{y}\in\mathbb{R}^d$$

is such that $f_{t,g}\in C^2(\mathbb{R}^d)$, $f_{t,g}$ solves (5) with g_t in place of g, and

$$\|\partial_i f_{t,g}\|_\infty \le \|g\|_\infty\sqrt{\frac{1-t}{t}}\sum_{j=1}^{d}(\Sigma^{-1/2})_{ji}\sqrt{\Sigma_{jj}},\quad \text{for any } i=1,\dots,d. \tag{8}$$

ii. *For any d-dimensional random vector $\mathbf{Y}$, it holds*

$$\sup_{g\in\mathcal{I}_d}\mathbb{E}\|\mathrm{Hess}(f_{t,g}(\mathbf{Y}))\|_{H.S.}^2 \le \|\Sigma^{-1}\|_{op}^2(d^2(\log t)^2 d_c(\mathbf{Y},\mathbf{N}_\Sigma)+530d^{17/6}),\quad \text{for any } t\in(0,1).$$

See proposition 4.3.2 in Nourdin and Peccati (2012) for Lemma 2.7; in particular, in Nourdin and Peccati (2012) it is noticed that

$$\partial_i f_g(\mathbf{y}) = -\int_0^\infty \mathrm{e}^{-t}\mathbb{E}[\partial_i g(\mathrm{e}^{-t}\mathbf{y}+\sqrt{1-\mathrm{e}^{-2t}}\mathbf{N}_\Sigma)]\mathrm{d}t,\quad i=1,\dots,d$$

which, combined with the fact that $g\in\mathcal{L}_d(1)$, gives the bound (6); see Schulte and Yukich (2019) p. 12, and proposition 2.3 for lemma 2.8 and lemma 3.6(ii) in Torrisi (2023) for the bound (8).

2.4. The Smoothing Lemma and the Integration by Parts Formula for Gaussian Random Vectors

We state a remarkable smoothing lemma for the convex distance proved in Schulte and Yukich (2019); see lemma 2.2 therein. It plays a crucial role in the proof of the Normal approximation of the output of a deep random Gaussian NN in the metric d_c; see Theorem 6.1.

Lemma 2.9. *Let* $\mathbf{Y}$ *be a d-dimensional random vector. Then, for any* $t \in (0,1)$,

$$d_c(\mathbf{Y}, \mathbf{N}_\Sigma) \le \frac{4}{3} \sup_{C \in \mathcal{C}_d} |\mathbb{P}(\sqrt{t}\mathbf{N}_\Sigma + \sqrt{1-t}\mathbf{Y} \in C) - \mathbb{P}(\sqrt{t}\mathbf{N}_\Sigma + \sqrt{1-t}\mathbf{N}_\Sigma \in C)| + \frac{20d}{\sqrt{2}} \frac{\sqrt{t}}{1-t},$$

where $\mathcal{C}_d$ *is given in Definition 2.2.*

We recall the Gaussian integration by parts formula (we refer the reader to exercise 3.1.4 in Nourdin and Peccati (2012) for the Part (i) of Lemma 2.10 and to exercise 3.1.5 in Nourdin and Peccati (2012) for the Part (ii) of Lemma 2.10).

Lemma 2.10. *The following claims hold:*

i. $N \sim \mathcal{N}(\mu, \eta)$, $\eta > 0$, *if and only if, for any differentiable function* $g : \mathbb{R} \to \mathbb{R}$ *such that* $\mathbb{E}|g'(N)| < \infty$, *we have* $\mathbb{E}(N-\mu)g(N) = \eta \mathbb{E}g'(N)$.

ii. *Let* $g \in C^1(\mathbb{R}^d)$ *with bounded first partial derivatives. Then,*

$$\mathbb{E}[(N_\Sigma)_i g(\mathbf{N}_\Sigma)] = \sum_{j=1}^{d} \Sigma_{ij} \mathbb{E}[\partial_j g(\mathbf{N}_\Sigma)], \quad \text{for any } i = 1, \dots, d.$$

This relation holds true even if Σ *is not positive definite.*

3. Random Neural Networks

We let $L \in \mathbb{N}$, we take $L+2$ positive integers $n_0, \dots, n_{L+1} \in \mathbb{N}$, and we fix a function $\sigma : \mathbb{R} \to \mathbb{R}$. A fully connected NN of depth L with input dimension n_0, output dimension n_{L+1}, hidden layer widths $n_1, \dots, n_L$, and nonlinearity σ is a mapping

$$\mathbf{x} := (x_1, \dots, x_{n_0}) \in \mathbb{R}^{n_0} \longmapsto \mathbf{z}^{(L+1)}(\mathbf{x}) = (z_1^{(L+1)}(\mathbf{x}), \dots, z_{n_{L+1}}^{(L+1)}(\mathbf{x}))) \in \mathbb{R}^{n_{L+1}}$$

that is defined by a recursive relation of the form

$$z_i^{(1)}(\mathbf{x}) = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j, \quad i = 1, \dots, n_1$$
$$z_i^{(\ell)}(\mathbf{x}) = b_i^{(\ell)} + \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \sigma(z_j^{(\ell-1)}(\mathbf{x})), \quad i = 1, \dots, n_\ell,\ \ell = 2, \dots, L+1$$

where the parameters $b_i^{(\ell)} \in \mathbb{R}$ and $W_{ij}^{(\ell)} \in \mathbb{R}$ are called network biases and weights, respectively. The quantities L and $n_0, \dots, n_{L+1}$ constitute the so-called network architecture. The function σ is usually called activation function. NNs of this kind will be denoted by

$$\mathrm{NN}(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W}),$$

where $\mathbf{n}_L := (n_1, \dots, n_L)$, $\mathbf{b} := (b_i^{(\ell)})$ and $\mathbf{W} := (W_{ij}^{(\ell)})$.

We say that the neural network $\mathrm{NN}(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ is a (fully connected and) deep random Gaussian neural network denoted by

$$\mathrm{GNN}(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W}),$$

if $\sigma : \mathbb{R} \to \mathbb{R}$ is measurable and $b_i^{(\ell)}, W_{ij}^{(\ell)}$, $i = 1, \dots, n_\ell$, $j = 1, \dots, n_{\ell-1}$, $\ell = 1, \dots, L+1$, are independent random variables with

$$b_i^{(\ell)} \sim \mathcal{N}(0, C_b) \quad \text{and} \quad W_{ij}^{(\ell)} \sim \mathcal{N}(0, C_W/n_{\ell-1}), \quad \ell = 1, \dots, L+1$$

for constants $C_b \ge 0$ and $C_W > 0$.

NNs of depth $L = 1$ are called shallow NNs. We will denote shallow NNs (respectively, shallow random Gaussian NNs) by $\mathrm{NN}(1, n_0, n_2, n_1, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ (respectively, by $\mathrm{GNN}(1, n_0, n_2, n_1, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$).

Throughout this paper, we will also consider NNs with univariate output, that is, NNs with $n_{L+1} = 1$.

Consider a deep random Gaussian neural network GNN$(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$. It turns out that the random variables $z_i^{(1)} = z_i^{(1)}(\mathbf{x})$, $i = 1, \dots, n_1$, are independent and identically distributed with

$$z_i^{(1)} \sim \mathcal{N}\left(0, C_b + \frac{C_W}{n_0}\sum_{j=1}^{n_0} x_j^2\right).$$

For $\ell = 1, \dots, L$, let $\mathcal{F}_\ell$ be the σ-field generated by the random variables

$$\{b_i^{(h)}, W_{ij}^{(h)}, \quad i = 1, \dots, n_h, j = 1, \dots, n_{h-1}, h = 1, \dots, \ell\}.$$

By construction, for any fixed $\ell \in \{2, \dots, L+1\}$, given $\mathcal{F}_{\ell-1}$, the random variables $z_i^{(\ell)} = z_i^{(\ell)}(\mathbf{x})$, $i = 1, \dots, n_\ell$, are independent and Gaussian (as linear combination of independent Gaussian random variables). A straightforward computation yields

$$\mathbb{E}[z_i^{(\ell)} | \mathcal{F}_{\ell-1}] = 0, \quad i = 1, \dots, n_\ell$$

and

$$\mathbb{E}[|z_i^{(\ell)}|^2 | \mathcal{F}_{\ell-1}] = C_b + \frac{C_W}{n_{\ell-1}}\sum_{j=1}^{n_{\ell-1}} |\sigma(z_j^{(\ell-1)})|^2, \quad i = 1, \dots, n_\ell. \tag{9}$$

Setting $\mathbf{n}_\ell = (n_1, \dots, n_\ell)$, $\ell = 1, \dots, L$, we define the quantities

$$\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} := \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \sigma(z_j^{(\ell)})^2, \quad \ell = 1, \dots, L$$

and

$$\mathcal{O}^{(\ell)} := \mathbb{E}\sigma(\mathfrak{s}_{\ell-1} Z)^2, \qquad \mathfrak{s}_{\ell-1}^2 := C_b + C_W \mathcal{O}^{(\ell-1)}, \quad \ell = 1, \dots, L \tag{10}$$

where

$$\mathcal{O}^{(0)} := \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2 \quad \text{and} \quad Z \sim \mathcal{N}(0,1). \tag{11}$$

The random variable $\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)}$ is often referred to as collective observable at layer ℓ; see, for example, Roberts et al. (2022) and Hanin (2023).

3.1. Some Related Literature

Consider the output $\mathbf{z}^{(L+1)} := (z_1^{(L+1)}, \dots, z_{n_{L+1}}^{(L+1)})$ of a deep random Gaussian neural network GNN$(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma; \mathbf{x}; \mathbf{b}, \mathbf{W})$, and let

$$\mathbf{z} := (z_1, \dots, z_{n_{L+1}})$$

be a centered n_{L+1}-dimensional Gaussian random vector with covariance matrix

$$\Sigma_{n_{L+1}} := \mathfrak{s}_L^2 \mathrm{Id}_{n_{L+1}}, \quad \mathfrak{s}_L^2 := C_b + C_W \mathcal{O}^{(L)}, \tag{12}$$

where $\mathrm{Id}_{n_{L+1}}$ is the identity matrix of $\mathcal{M}_{n_{L+1} \times n_{L+1}}(\mathbb{R})$. Here and henceforth, we suppose that the quantities $C_b, \|\mathbf{x}\|, \sigma(0)$, where $\mathbf{x}$ is the input of the network, are not all simultaneously equal to zero. This causes no loss of generality, because otherwise $\mathbf{z}^{(L+1)} = \mathbf{z} = 0$ almost surely.

It follows from theorem 1.2 in Hanin (2023) (which indeed, more generally, establishes a functional weak convergence) that, if σ is continuous and polynomially bounded, then,

$$\mathbf{z}^{(L+1)} \to \mathbf{z} \quad \text{in law, as } \min\{n_1, \dots, n_L\} \to +\infty. \tag{13}$$

The following result for shallow random Gaussian NNs was proved in Bordino et al. (2024); see theorem 3.2 therein.

Theorem 3.1. *Let* GNN$(1, n_0, 1, n_1, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ *be a shallow random Gaussian NN with univariate output. If*

$$\sigma \in C^2(\mathbb{R}) \text{ and } \max\{|\sigma(x)|, |\sigma'(x)|, |\sigma''(x)|\} \le r_1 + r_2|x|^\gamma, \ x \in \mathbb{R} \tag{14}$$

for some $r_1, r_2, \gamma \geq 0$, *then*

$$d_s(z^{(2)}, z) \leq c_s \|r_1 + r_2 |\mathfrak{s}_0 Z|^\gamma\|_{L^4}^2 \sqrt{\mathfrak{s}_0^2 + \mathfrak{s}_0^4 (2 + \sqrt{3(1 + 2\mathfrak{s}_0^2 + 3\mathfrak{s}_0^4)})} \times \frac{1}{\sqrt{n_1}},$$

where $s = TV, K, W_1$ *and*

$$c_{TV} := \frac{4}{\mathfrak{s}_1^2}, \quad c_K := \frac{2}{\mathfrak{s}_1^2}, \quad c_{W_1} := \frac{1}{\mathfrak{s}_1} \sqrt{8/\pi}.$$

In Section 4, we will give bounds on the quantities $d_s(z^{(2)}, z)$, $s = TV, K, W_1$, of order $1/\sqrt{n_1}$, as $n_1 \to \infty$, under a minimal assumption on the activation function (note that Condition (14) excludes the important case of the ReLU function, that is, $\sigma(x) := x\mathbf{1}\{x \geq 0\}$); see Theorem 4.1. In Section 6, we will give two general bounds on $d_c(\mathbf{z}^{(L+1)}, \mathbf{z})$ and $d_{W_1}(\mathbf{z}^{(L+1)}, \mathbf{z})$ for deep random Gaussian NNs; see Theorems 6.1 and 6.3. When specialized to shallow random Gaussian neural networks GNN$(1, n_0, n_2, n_1, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$, they provide computable bounds, respectively, on $d_c(\mathbf{z}^{(2)}, \mathbf{z})$ and $d_{W_1}(\mathbf{z}^{(2)}, \mathbf{z})$ of order $1/\sqrt{n_1}$, as $n_1 \to \infty$, see theorem 3.3 in Bordino et al. (2024) for a related result.

The first result in the literature that quantifies the convergence in distribution (13) with $L \geq 2$ is given in Basteri and Trevisan (2024), where the following theorem has been proven.

Theorem 3.2. *Let* GNN$(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ *be a deep random Gaussian NN, and suppose that the activation function* σ *is Lipschitz continuous. Then,*

$$d_{W_2}(\mathbf{z}^{(L+1)}, \mathbf{z}) \leq \sqrt{n_{L+1}} \sum_{i=1}^{L} \frac{C^{(i+1)} [\mathrm{Lip}(\sigma) \sqrt{C_W}]^{L-i}}{\sqrt{n_i}}, \tag{15}$$

where, for any $i = 1, \ldots, L$, $C^{(i+1)}$ *are explicitly known positive constants, depending upon* σ, $\mathbf{x}$, C_b *and* C_W.

The next corollary is an immediate consequence of Theorem 3.2, the fact that $d_{W_1} \leq d_{W_2}$, Proposition 2.5, and (3).

Corollary 3.3. *Let the assumptions and notation of Theorem 3.2 prevail. Then,*

$$d_{W_1}(\mathbf{z}^{(L+1)}, \mathbf{z}) \leq \sqrt{n_{L+1}} \sum_{i=1}^{L} \frac{C^{(i+1)} [\mathrm{Lip}(\sigma) \sqrt{C_W}]^{L-i}}{\sqrt{n_i}} \tag{16}$$

and

$$d_c(\mathbf{z}^{(L+1)}, \mathbf{z}) \leq 2^{\frac{11}{8}} \pi^{-\frac{1}{8}} n_{L+1}^{\frac{3}{8}} \left(\sum_{i=1}^{L} \frac{C^{(i+1)} [\mathrm{Lip}(\sigma) \sqrt{C_W}]^{L-i}}{\sqrt{n_i}} \right)^{1/2}. \tag{17}$$

Theorems 6.1 and 6.3 will provide, under alternate assumptions on σ (see by Proposition 5.2(ii)), explicit bounds, respectively, on $d_c(\mathbf{z}^{(L+1)}, \mathbf{z})$ and $d_{W_1}(\mathbf{z}^{(L+1)}, \mathbf{z})$ that are of the same order of the bound in (16), as $n_1, \ldots, n_L \to \infty$. In particular, the bound on the convex distance of Theorem 6.1 considerably improves the one in (17). We emphasize that to compare the results in Basteri and Trevisan (2024) with our achievements, we specialized those results to the case of a single input $\mathbf{x} \in \mathbb{R}^{n_0}$. In fact, the bound (15) was proven in Basteri and Trevisan (2024) for multiple inputs, and consequently the bound (17) can be stated for multiple inputs too.

Remark 3.4. Although not strictly related to our results, the pioneering papers Eldan et al. (2021) and Neal (1996) deserve a special mention. In Neal (1996) the author considered a random shallow $NN(1, n_0, 1, n_1, \sigma, \mathbf{x}, \mathbf{0}, \mathbf{W})$ with univariate output, $b_i^{(1)} := 0$, for all $i = 1, \ldots, n_1$, $b_1^{(2)} := 0$, $W_{ij}^{(1)}$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_0$, independent with law $\mathcal{N}(0, 1)$ and $W_{1j}^{(2)}$, $j = 1, \ldots, n_1$, independent, identically distributed with law $\mathbb{P}(W_{1j}^{(2)} = \pm 1/\sqrt{n_1}) = 1/2$ and independent of the random variables $W_{ij}^{(1)}$. It was proven in Neal (1996) that there exists a Gaussian process G on $\mathbb{S}^{n_0 - 1}$ (the unit sphere in $\mathbb{R}^{n_0}$) such that the process $\{z^{(2)}(\mathbf{x})\}_{\mathbf{x} \in \mathbb{S}^{n_0 - 1}}$ converges in distribution to G, as $n_1 \to \infty$. Quantitative versions of this result (for various Wasserstein metrics and some specific choices of σ) are provided in Eldan et al. (2021).

4. Normal Approximation of Shallow Random Gaussian NNs with Univariate Output

The following theorem holds.

Theorem 4.1. *Let* $\mathrm{GNN}(1, n_0, 1, n_1, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ *be a shallow random Gaussian NN with univariate output, and assume that the activation function* σ *is such that*

$$0 < \mathbb{V}\mathrm{ar}(\sigma(\mathfrak{s}_0 Z)^2) < \infty. \tag{18}$$

Then:

i.

$$d_K(z^{(2)}, z) \le \frac{C_W \sqrt{\mathbb{V}\mathrm{ar}(\sigma(\mathfrak{s}_0 Z)^2)}}{C_b + C_W \mathbb{E}\sigma(\mathfrak{s}_0 Z)^2} \frac{1}{\sqrt{n_1}}.$$

ii.

$$d_{TV}(z^{(2)}, z) \le 2 \frac{C_W \sqrt{\mathbb{V}\mathrm{ar}(\sigma(\mathfrak{s}_0 Z)^2)}}{C_b + C_W \mathbb{E}\sigma(\mathfrak{s}_0 Z)^2} \frac{1}{\sqrt{n_1}}.$$

iii.

$$d_{W_1}(z^{(2)}, z) \le \sqrt{2/\pi} \frac{C_W \sqrt{\mathbb{V}\mathrm{ar}(\sigma(\mathfrak{s}_0 Z)^2)}}{\sqrt{C_b + C_W \mathbb{E}\sigma(\mathfrak{s}_0 Z)^2}} \frac{1}{\sqrt{n_1}}.$$

Note that the bound on the Kolmogorov distance is better than the one that can be obtained using the relation $d_K \le d_{TV}$.

Remark 4.2. Remarkably, theorem 3.3 in Favaro et al. (2023) shows that if $\mathrm{GNN}(1, n_0, 1, n_1, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ is a shallow random Gaussian NN with univariate output and the activation function σ is polynomially bounded to order $r \ge 1$ (see definition 2.1 in Favaro et al. (2023)), then there exist two constants $C, C_0 > 0$ such that

$$\frac{C_0}{n_1} \le \max\{d_{W_1}(z^{(2)}, z), d_{TV}(z^{(2)}, z)\} \le \frac{C}{n_1}.$$

Although this inequality shows the optimality of the rate $1/n_1$, because the constants are not provided in closed form, it cannot be directly used for the purpose of output localization (see Section 7). Moreover, the assumption (18) on σ does not require any regularity of the activation function.

Proof. We prove the three bounds (i), (ii), and (iii) separately by the Stein method.

Proof of Part (i). Consider the Stein Equation (4) with

$$g(w) := \mathbf{1}_{(-\infty, y]}(\mathfrak{s}_1 w).$$

Let f_g be the unique solution of the Stein equation (see Lemma 2.6(i)). Then, for any $y \in \mathbb{R}$,

$$\mathbf{1}\{z^{(2)} \le y\} - \mathbb{P}(z \le y) = f_g'(z^{(2)}/\mathfrak{s}_1) - (z^{(2)}/\mathfrak{s}_1) f_g(z^{(2)}/\mathfrak{s}_1).$$

Taking the expectation, we have

$$\mathbb{P}(z^{(2)} \le y) - \mathbb{P}(z \le y) = \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1) - (z^{(2)}/\mathfrak{s}_1) f_g(z^{(2)}/\mathfrak{s}_1)].$$

By Lemma 2.10(i), we have

$$\begin{aligned} \mathbb{E}[(z^{(2)}/\mathfrak{s}_1) f_g(z^{(2)}/\mathfrak{s}_1) \mid \mathcal{F}_1] &= \mathfrak{s}_1^{-2}(C_b + C_W \mathcal{O}_{n_1}^{(1)}) \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1) \mid \mathcal{F}_1] \\ &= \mathbb{E}[\mathfrak{s}_1^{-2}(C_b + C_W \mathcal{O}_{n_1}^{(1)}) f_g'(z^{(2)}/\mathfrak{s}_1) \mid \mathcal{F}_1], \end{aligned} \tag{19}$$

where the latter equality follows by the $\mathcal{F}_1$-measurability of $\mathcal{O}_{n_1}^{(1)}$. Then,

$$\mathbb{P}(z^{(2)} \le y) - \mathbb{P}(z \le y) = \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1)(1 - \mathfrak{s}_1^{-2}(C_b + C_W \mathcal{O}_{n_1}^{(1)}))], \quad y \in \mathbb{R}.$$

Setting

$$\varphi(n_1) := \mathfrak{s}_1^{-2} C_W \sqrt{\mathbb{V}\mathrm{ar}(\mathcal{O}_{n_1}^{(1)})} = \mathfrak{s}_1^{-2} C_W \sqrt{\mathbb{V}\mathrm{ar}(\sigma(z_1^{(1)})^2)} / \sqrt{n_1}, \tag{20}$$

we have

$$\frac{\mathbb{P}(z^{(2)} \le y) - \mathbb{P}(z \le y)}{\varphi(n_1)} = \mathbb{E}[f_g'(z^{(2)}/\nu)V_{n_1}], \tag{21}$$

where

$$V_{n_1} := \frac{1 - \mathfrak{s}_1^{-2}(C_b + C_W\mathcal{O}_{n_1}^{(1)}))}{\varphi(n_1)} = -\frac{\mathcal{O}_{n_1}^{(1)} - \mathcal{O}^{(1)}}{\sqrt{\mathbb{V}\mathrm{ar}(\mathcal{O}_{n_1}^{(1)})}} = -\frac{\sum_{j=1}^{n_1}\sigma(z_j^{(1)})^2 - n_1\mathbb{E}\sigma(z_1^{(1)})^2}{\sqrt{n_1}\sqrt{\mathbb{V}\mathrm{ar}(\sigma(z_1^{(1)})^2)}}.$$

Taking the modulus in (21) and then using that $\|f_g'\|_\infty \le 1$ uniformly in $y \in \mathbb{R}$ (see Lemma 2.6(i)) and that $\mathbb{E}V_{n_1}^2 = 1$, we have

$$d_K(z^{(2)}, z) \le \varphi(n_1)\mathbb{E}[|V_{n_1}|] \le \varphi(n_1),$$

which, combined with (20), proves the statement.

Proof of Part (ii). Consider the Stein Equation (4) with $g(w) := \mathbf{1}_B(\mathfrak{s}_1 w)$, where $B \subseteq \mathbb{R}^d$ is a Borel set. Let f_g be the unique solution of the Stein equation (see Lemma 2.6(ii)). Then,

$$\mathbf{1}_B(z^{(2)}) - \mathbb{E}\mathbf{1}_B(z) = f_g'(z^{(2)}/\mathfrak{s}_1) - (z^{(2)}/\mathfrak{s}_1)f_g(z^{(2)}/\mathfrak{s}_1).$$

Taking the expectation and arguing as in (19), we have

$$\mathbb{P}(z^{(2)} \in B) - \mathbb{P}(z \in B) = \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1)(1 - \mathfrak{s}_1^{-2}(C_b + C_W\mathcal{O}_{n_1}^{(1)}))].$$

Along similar computations as for (21), we have

$$\frac{\mathbb{P}(z^{(2)} \in B) - \mathbb{P}(z \in B)}{\varphi(n_1)} = \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1)V_{n_1}].$$

Taking the modulus on this relation and then using that $\|f_g'\|_\infty \le 2$ (see Lemma 2.6(ii)) and that $\mathbb{E}V_{n_1}^2 = 1$, we have

$$d_{TV}(z^{(2)}, z) \le 2\varphi(n_1)\mathbb{E}[|V_{n_1}|] \le 2\varphi(n_1),$$

which, combined with (20), proves the statement.

Proof of Part (iii). Consider the Stein Equation (4) with $g(y) := h(\mathfrak{s}_1 y)$, where $h : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with $\mathrm{Lip}(h) \le 1$. Let f_g be the unique solution of the Stein equation (see Lemma 2.6(iii)). Then,

$$h(z^{(2)}) - \mathbb{E}h(z) = f_g'(z^{(2)}/\mathfrak{s}_1) - (z^{(2)}/\mathfrak{s}_1)f_g(z^{(2)}/\mathfrak{s}_1).$$

Taking the expectation and arguing as in (19), we have

$$\mathbb{E}h(z^{(2)}) - \mathbb{E}h(z) = \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1)(1 - \mathfrak{s}_1^{-2}(C_b + C_W\mathcal{O}_{n_1}^{(1)}))].$$

Along similar computations as for (21), we have

$$\frac{\mathbb{E}h(z^{(2)}) - \mathbb{E}h(z)}{\varphi(n_1)} = \mathbb{E}[f_g'(z^{(2)}/\mathfrak{s}_1)V_{n_1}].$$

Taking the modulus on this relation and then using that $\|f_g'\|_\infty \le \mathfrak{s}_1\sqrt{2/\pi}$ (see Lemma 2.6(iii)) and that $\mathbb{E}V_{n_1}^2 = 1$, we have

$$d_{W_1}(z^{(2)}, z) \le \mathfrak{s}_1\sqrt{2/\pi}\varphi(n_1)\mathbb{E}[|V_{n_1}|] \le \mathfrak{s}_1\sqrt{2/\pi}\varphi(n_1),$$

which, combined with (20), proves the statement. □

Note that both Theorem 3.1 and Theorem 4.1 provide bounds on $d_s(z^{(2)}, z)$, $s = TV, K, W_1$, with a common rate $1/\sqrt{n_1}$, but different constants. In Table 1, we compare those constants in a special case. We observe that the constants given by Theorem 4.1 are more effective than those in Bordino et al. (2024). We also note that Condition (18) is satisfied by the ReLU activation function, whereas the assumptions of Theorem 3.1 do not hold for the ReLU.

5. A Key Estimate for the Collective Observables

The next theorem provides an estimate for the L^2-norm of the random variable $\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell} - \mathcal{O}^{(\ell)}$, $\ell = 1, \dots, L$. In Section 6, such an estimate will play a crucial role in the proofs of the results on the Normal approximation of the output of a deep random Gaussian NN both in the convex and in the 1-Wasserstein distances (see Theorems 6.1 and 6.3).

Hereon, we denote by $\|Y\|_{L^2} := (\mathbb{E}[|Y|^2])^{1/2}$ the L^2-norm of a real-valued random variable Y.

Theorem 5.1. *Let* GNN$(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ *be a deep random Gaussian NN. Suppose that the activation function* σ *is such that:*

i. *For any* $a_1, a_2 \geq 0$, *there exists a polynomial*

$$P(x) := \sum_{k=0}^{m} d_k x^k,$$

with nonnegative coefficients $d_k = d_k(\sigma(\cdot), C_b, C_W) \geq 0$ *dependent only on* $\sigma(\cdot), C_b, C_W$ *and degree* $m \geq 0$ *independent of* $\sigma(\cdot), a_1, a_2, C_b, C_W$, *such that*

$$|\sigma(x\sqrt{C_b + C_W a_2})^2 - \sigma(x\sqrt{C_b + C_W a_1})^2| \leq P(|x|)|a_2 - a_1|, \quad \text{for all } x \in \mathbb{R}. \tag{22}$$

ii. *For any* $\kappa \in \mathbb{R}$, $\mathbb{E}\sigma(\kappa Z)^4 < \infty$.

Then, for any $\ell = 1, \dots, L$, *we have*

$$\|\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell} - \mathcal{O}^{(\ell)}\|_{L^2} \leq \sum_{k=1}^{\ell} (4\sqrt{2}\|P(|Z|)\|_2)^{\ell-k} \frac{c_k}{\sqrt{n_k}},$$

where

$$c_k = c_k(n_0, \sigma, \mathbf{x}, C_b, C_W) := \sqrt{2\mathbb{E}\sigma(\mathfrak{s}_{k-1}Z)^4} < \infty, \quad k = 1, \dots, L. \tag{23}$$

The proof of the theorem is given later on in this section. We proceed stating a proposition and a remark, which clarify the generality of our assumptions on the activation function σ.

Proposition 5.2. *The following statements hold:*

i. *If* σ *is the perceptron function, that is,* $\sigma(x) := \mathbf{1}\{x \geq 0\}$, $x \in \mathbb{R}$, *then it satisfies Conditions* (i) *and* (ii) *of Theorem* 5.1.

ii. *If* σ *is Lipschitz continuous and* $C_b > 0$, *then* σ *satisfies Conditions* (i) *and* (ii) *of Theorem* 5.1. *In particular, Condition* (i) *holds with*

$$P(x) := \frac{2|\sigma(0)|C_W \mathrm{Lip}(\sigma)}{2\sqrt{C_b}} x + C_W \mathrm{Lip}(\sigma)^2 x^2. \tag{24}$$

iii. *If* σ *is Lipschitz continuous and* $\sigma(0) = 0$, *then* σ *satisfies Conditions* (i) *and* (ii) *of Theorem* 5.1. *In particular, Condition* (i) *holds with*

$$P(x) := C_W \mathrm{Lip}(\sigma)^2 x^2. \tag{25}$$

iv. *If* σ *is such that* $\sigma(\cdot)^2$ *is Lipschitz continuous and* $C_b > 0$, *then* σ *satisfies Conditions* (i) *and* (ii) *of Theorem* 5.1. *In particular, Condition* (i) *holds with*

$$P(x) = \frac{\mathrm{Lip}(\sigma^2)C_W}{2\sqrt{C_b}} x. \tag{26}$$

The proof of the proposition is given later on in this section.

Remark 5.3. As a consequence of Proposition 5.2, we have that the most common activation functions satisfy the assumptions of Theorem 5.1. Indeed, one can easily prove that the ReLU function $\sigma(x) := x\mathbf{1}\{x \geq 0\}$, the hyperbolic tangent function $\sigma(x) := (e^{2x} - 1)/(e^{2x} + 1)$, the sin function $\sigma(x) := \sin(x)$, and the SWISH function $\sigma(x) := x/(1 + e^{-x})$ are Lipschitz continuous and equal to zero in zero. Moreover, if $C_b > 0$, then also the sigmoid function $\sigma(x) := (1 + e^{-x})^{-1}$ and the softplus function $\sigma(x) := \log(1 + e^x)$ satisfy the assumptions of Theorem 5.1, and indeed they are Lipschitz continuous. We emphasize that, if either the biases are nonnull or the activation function vanishes at 0, then the conditions on the activation function of Theorem 5.1 are more general than the one required in Basteri and Trevisan (2024), where $\sigma(\cdot)$ is assumed Lipschitz continuous (see Theorem 3.2 and Proposition 5.2 parts (ii) and (iii)). We also emphasize that the conditions on the activation function of Theorem 5.1 are satisfied by the perceptron function that is noncontinuous and therefore non-Lipschitz (see Proposition 5.2(i).

Another non-Lipschitz function that satisfies the conditions of Theorem 5.1 when the biases are nonnull I, for example, $\sigma(x) := \sqrt{x}\mathbf{1}\{x \geq 0\}$. Indeed, σ^2 is the ReLU function and therefore Lipschitz continuous (see Proposition 5.2(iv)).

Proof of Theorem 5.1. We consider separately the cases of the first hidden layer and that of the following ones.

Case $\ell = 1$.

Because the random variables $z_i^{(1)}$, $i = 1, \dots, n_1$, are independent and identically distributed with law $\mathcal{N}(0, \mathfrak{s}_0^2)$, we have

$$\begin{aligned}
\|\mathcal{O}_{n_1}^{(1)} - \mathcal{O}^{(1)}\|_{L^2} &= \sqrt{\mathbb{E}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\sigma(z_j^{(1)})^2 - \mathbb{E}\sigma(\mathfrak{s}_0 Z)^2\right)^2} \\
&= \sqrt{\mathbb{E}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}(\sigma(z_j^{(1)})^2 - \mathbb{E}\sigma(z_1^{(1)})^2)\right)^2} \\
&= \frac{1}{\sqrt{n_1}}\sqrt{\mathbb{V}\mathrm{ar}(\sigma(z_1^{(1)})^2)} \\
&\leq \frac{1}{\sqrt{n_1}}\sqrt{\mathbb{E}\sigma(\mathfrak{s}_0 Z)^4} \qquad (27) \\
&\leq \frac{c_1}{\sqrt{n_1}}. \qquad (28)
\end{aligned}$$

Note that this latter term is finite because of the assumption (ii).

Case $\ell = 2, \dots, L$.

Take $\ell \in \{2, \dots, L\}$. We have already noticed that, given $\mathcal{F}_{\ell-1}$, the random variables $\{z_i^{(\ell)}\}_{i=1,\dots,n_\ell}$ are independent with Gaussian law with mean zero and variance

$$\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}}^2 := C_b + C_W \mathcal{O}_{\mathbf{n}_{\ell-1}}^{(\ell-1)}.$$

Therefore, letting $Z_{\ell-1}$ denote a standard Gaussian random variable, independent of $\mathcal{F}_{\ell-1}$, we have

$$z_i^{(\ell)} \overset{d}{=} \mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}} Z_{\ell-1}, \quad i = 1, \dots, n_\ell$$

where the symbol $\overset{d}{=}$ denotes the equality in law (this relation immediately follows computing, for example, the characteristic function of both random variables). Therefore, letting $p(\cdot)$ denote the standard Gaussian density, we have

$$\mathbb{E}[\sigma(z_i^{(\ell)})^r | \mathcal{F}_{\ell-1}] \overset{d}{=} \int_{\mathbb{R}} \sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}} z)^r p(z)\mathrm{d}z, \quad r \in \{2,4\},\ i = 1, \dots, n_\ell \qquad (29)$$

and

$$\mathbb{E}\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} = \mathbb{E}\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}} Z_{\ell-1})^2.$$

By assumption (i) and the fact that $Z_{\ell-1}$ is independent of $\mathcal{F}_{\ell-1}$, we have

$$\begin{aligned}
|\mathbb{E}\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} - \mathcal{O}^{(\ell)}| &= |\mathbb{E}[\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}} Z_{\ell-1})^2] - \mathbb{E}[\sigma(\mathfrak{s}_{\ell-1} Z_{\ell-1})^2]| \\
&\leq \mathbb{E}P(|Z_{\ell-1}|)|\mathcal{O}_{\mathbf{n}_{\ell-1}}^{(\ell-1)} - \mathcal{O}^{(\ell-1)}| \\
&\leq \mathbb{E}P(|Z|)\|\mathcal{O}_{\mathbf{n}_{\ell-1}}^{(\ell-1)} - \mathcal{O}^{(\ell-1)}\|_{L^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} - \mathcal{O}^{(\ell)}\|_{L^2} &\leq \|\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} - \mathbb{E}\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)}\|_{L^2} + |\mathbb{E}\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} - \mathcal{O}^{(\ell)}| \\
&= \sqrt{\mathbb{V}\mathrm{ar}(\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)})} + |\mathbb{E}\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)} - \mathcal{O}^{(\ell)}| \\
&\leq \sqrt{\mathbb{V}\mathrm{ar}(\mathcal{O}_{\mathbf{n}_\ell}^{(\ell)})} + \mathbb{E}P(|Z|)\|\mathcal{O}_{\mathbf{n}_{\ell-1}}^{(\ell-1)} - \mathcal{O}^{(\ell-1)}\|_{L^2}. \qquad (30)
\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell}) &= \frac{1}{n_\ell^2}\left(\mathbb{E}\left(\sum_{j=1}^{n_\ell}\sigma(z_j^{(\ell)})^2\right)^2 - n_\ell^2(\mathbb{E}\sigma(z_1^{(\ell)})^2)^2\right)\\ &= \frac{1}{n_\ell^2}(n_\ell\mathbb{E}\sigma(z_1^{(\ell)})^4 + n_\ell(n_\ell-1)\mathbb{E}\sigma(z_1^{(\ell)})^2\sigma(z_2^{(\ell)})^2 - n_\ell^2(\mathbb{E}\sigma(z_1^{(\ell)})^2)^2)\\ &= \frac{1}{n_\ell^2}(n_\ell\mathbb{E}\sigma(z_1^{(\ell)})^4 - n_\ell\mathbb{E}\sigma(z_1^{(\ell)})^2\sigma(z_2^{(\ell)})^2 + n_\ell^2\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2))\\ &= \frac{1}{n_\ell^2}(n_\ell\mathbb{V}\mathrm{ar}(\sigma(z_1^{(\ell)})^2) + (n_\ell^2-n_\ell)\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2))\\ &= \frac{1}{n_\ell}\mathbb{V}\mathrm{ar}(\sigma(z_1^{(\ell)})^2) + \left(1-\frac{1}{n_\ell}\right)\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2). \end{aligned} \tag{31}$$

By (29), we have

$$\mathbb{E}[\sigma(z_1^{(\ell)})^4] = \mathbb{E}\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}}Z_{\ell-1})^4.$$

By assumption (i), we have

$$\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}}Z_{\ell-1})^2 \le P(|Z_{\ell-1}|)|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}| + \sigma(\mathfrak{s}_{\ell-1}Z_{\ell-1})^2, \quad \mathbb{P}\text{-a.s.}$$

and so (using that $Z_{\ell-1}$ is independent of $\mathcal{F}_{\ell-1}$ and the inequality $(a+b)^2 \le 2a^2+2b^2$, $a,b\in\mathbb{R}$),

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\sigma(z_1^{(\ell)})^2) &\le \mathbb{E}[\sigma(z_1^{(\ell)})^4]\\ &\le 2\mathbb{E}P(|Z|)^2\|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}\|_2^2 + 2\mathbb{E}\sigma(\mathfrak{s}_{\ell-1}Z)^4\\ &= A\|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}\|_{L^2}^2 + c_\ell^2, \end{aligned} \tag{32}$$

where $A := 2\mathbb{E}P(|Z|)^2 < \infty$. Note that the quantities $\mathcal{O}^{(\ell)}$, $\ell = 2,\dots,L$, are all finite because of the assumption (ii). Then, again by assumption (ii), we have

$$\mathbb{E}\sigma(\mathfrak{s}_{\ell-1}Z)^4 < \infty, \quad \text{for any } \ell = 2,\dots,L$$

and therefore, $c_\ell < \infty$, for any $\ell = 2,\dots,L$. By the conditional independence of the random variables $z_1^{(\ell)}$ and $z_2^{(\ell)}$, given $\mathcal{F}_{\ell-1}$, we have

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2) &= \mathbb{C}\mathrm{ov}(\mathbb{E}[\sigma(z_1^{(\ell)})^2 \mid \mathcal{F}_{\ell-1}],\mathbb{E}[\sigma(z_2^{(\ell)})^2 \mid \mathcal{F}_{\ell-1}])\\ &= \mathbb{E}[(\mathbb{E}[\sigma(z_1^{(\ell)})^2 \mid \mathcal{F}_{\ell-1}] - \mathbb{E}\sigma(z_1^{(\ell)})^2)(\mathbb{E}[\sigma(z_2^{(\ell)})^2 \mid \mathcal{F}_{\ell-1}] - \mathbb{E}\sigma(z_2^{(\ell)})^2)], \end{aligned}$$

and so, by the Cauchy-Schwarz inequality and (29) we have

$$|\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2)| \le \mathbb{E}[(\mathbb{E}[\sigma(z_1^{(\ell)})^2|\mathcal{F}_{\ell-1}] - \mathbb{E}\sigma(z_1^{(\ell)})^2)^2]. \tag{33}$$

Letting $\mathbb{P}_X$ denote the law of a random variable X and again using (29), we have that the random variable $\mathbb{E}[\sigma(z_1^{(\ell)})^2|\mathcal{F}_{\ell-1}] - \mathbb{E}\sigma(z_1^{(\ell)})^2$ has the same law as the random variable

$$\begin{aligned}&\int_{\mathbb{R}}\left(\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}}z)^2 - \int_{[0,\infty)}\sigma(z\sqrt{C_b+C_W y})^2\mathbb{P}_{\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}}}(\mathrm{d}y)\right)p(z)\mathrm{d}z\\ &= \int_{[0,\infty)\times\mathbb{R}}(\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}}z)^2 - \sigma(z\sqrt{C_b+C_W y})^2)\mathbb{P}_{\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}}}(\mathrm{d}y)p(z)\mathrm{d}z. \end{aligned}$$

Therefore, by (33) and Jensen's inequality, we have

$$|\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2)| \le \mathbb{E}\int_{[0,\infty)\times\mathbb{R}}(\sigma(\mathfrak{s}_{\ell-1,\mathbf{n}_{\ell-1}}z)^2 - \sigma(z\sqrt{C_b+C_W y})^2)^2\mathbb{P}_{\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}}}(\mathrm{d}y)p(z)\mathrm{d}z$$

By assumption (i), it then follows that

$$|\mathbb{C}\mathrm{ov}(\sigma(z_1^{(\ell)})^2,\sigma(z_2^{(\ell)})^2)| \le (\mathbb{E}P(|Z|))^2\int_{[0,\infty)}\mathbb{E}|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - y|^2\mathbb{P}_{\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}}}(\mathrm{d}y) \le A\mathbb{V}\mathrm{ar}\left(\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}}\right), \tag{34}$$

where the latter relation follows by the definition of the constant A. Combining (31), (32), and (34), we have

$$\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell}) \le A(\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}}) + \|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}_{\sigma^2}\|^2_{L^2}) + \frac{c_\ell^2}{n_\ell}.$$

Iterating this inequality, we have

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell}) &\le A\left(A(\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell-2)}_{\mathbf{n}_{\ell-2}}) + \|\mathcal{O}^{(\ell-2)}_{\mathbf{n}_{\ell-2}} - \mathcal{O}^{(\ell-2)}\|^2_{L^2}) + \frac{c_{\ell-1}^2}{n_{\ell-1}}\right) + A\|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}\|^2_{L^2} + \frac{c_\ell^2}{n_\ell} \\
&= A^2\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell-2)}_{\mathbf{n}_{\ell-2}}) + A^2\|\mathcal{O}^{(\ell-2)}_{\mathbf{n}_{\ell-2}} - \mathcal{O}^{(\ell-2)}\|^2_{L^2} + A\|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}\|^2_{L^2} + A\frac{c_{\ell-1}^2}{n_{\ell-1}} + \frac{c_\ell^2}{n_\ell} \\
&\le A^3\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell-3)}_{\mathbf{n}_{\ell-3}}) + A^3\|\mathcal{O}^{(\ell-3)}_{\mathbf{n}_{\ell-3}} - \mathcal{O}^{(\ell-3)}\|^2_{L^2} + A^2\|\mathcal{O}^{(\ell-2)}_{\mathbf{n}_{\ell-2}} - \mathcal{O}^{(\ell-2)}\|^2_{L^2} \\
&\quad + A\|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}\|^2_{L^2} + A^2\frac{c_{\ell-2}^2}{n_{\ell-2}} + A\frac{c_{\ell-1}^2}{n_{\ell-1}} + \frac{c_\ell^2}{n_\ell} \\
&\le A^{\ell-1}\mathbb{V}\mathrm{ar}(\mathcal{O}^{(1)}_{n_1}) + \sum_{k=1}^{\ell-1} A^{\ell-k}\|\mathcal{O}^{(k)}_{\mathbf{n}_k} - \mathcal{O}^{(k)}\|^2_{L^2} + \sum_{k=2}^{\ell} A^{\ell-k}\frac{c_k^2}{n_k} \\
&= 2A^{\ell-1}\mathbb{V}\mathrm{ar}(\mathcal{O}^{(1)}_{n_1}) + \sum_{k=2}^{\ell-1} A^{\ell-k}\|\mathcal{O}^{(k)}_{\mathbf{n}_k} - \mathcal{O}^{(k)}\|^2_{L^2} + \sum_{k=2}^{\ell} A^{\ell-k}\frac{c_k^2}{n_k},
\end{aligned}$$

for any $\ell = 2, \dots, L$, where the latter equality follows noticing that $\mathbb{E}\mathcal{O}^{(1)}_{n_1} = \mathcal{O}^{(1)}$. Here, we adopt the usual convention $\sum_{k=k_1}^{k_2} \cdots = 0$ if $k_1 > k_2$. Note that by (27), we have

$$\mathbb{V}\mathrm{ar}(\mathcal{O}^{(1)}_{n_1}) \le \frac{1}{n_1}\mathbb{E}\sigma(\mathfrak{s}_0 Z)^4,$$

and so,

$$2\mathbb{V}\mathrm{ar}(\mathcal{O}^{(1)}_{n_1}) \le \frac{c_1^2}{n_1}.$$

Consequently,

$$\mathbb{V}\mathrm{ar}(\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell}) \le \sum_{k=2}^{\ell-1} A^{\ell-k}\|\mathcal{O}^{(k)}_{\mathbf{n}_k} - \mathcal{O}^{(k)}\|^2_{L^2} + \sum_{k=1}^{\ell} A^{\ell-k}\frac{c_k^2}{n_k}.$$

Combining this inequality with (30) and using the elementary relation $\sqrt{a_1 + a_2} \le \sqrt{a}_1 + \sqrt{a}_2$, $a_1, a_2 \ge 0$, for any $\ell = 2, \dots, L$, we have

$$\begin{aligned}
\|\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell} - \mathcal{O}^{(\ell)}\|_{L^2} &\le \sum_{k=2}^{\ell-1} A^{(\ell-k)/2}\|\mathcal{O}^{(k)}_{\mathbf{n}_k} - \mathcal{O}^{(k)}\|_{L^2} + \sum_{k=1}^{\ell} A^{(\ell-k)/2}\frac{c_k}{\sqrt{n_k}} + \mathbb{E}P(|Z|)\|\mathcal{O}^{(\ell-1)}_{\mathbf{n}_{\ell-1}} - \mathcal{O}^{(\ell-1)}\|_{L^2} \\
&\le \sum_{k=2}^{\ell-1} B^{(\ell-k)/2}\|\mathcal{O}^{(k)}_{\mathbf{n}_k} - \mathcal{O}^{(k)}\|_{L^2} + \sum_{k=1}^{\ell} B^{(\ell-k)/2}\frac{c_k}{\sqrt{n_k}},
\end{aligned} \tag{35}$$

where we used that $\mathbb{E}P(|Z|) \le A^{1/2}$, and we set $B := 4A$. Recalling that $\sqrt{A} = \sqrt{2}\|P(|Z|)\|_2$ and the definition of B, one easily realizes that the claim reads as

$$\|\mathcal{O}^{(\ell)}_{\mathbf{n}_2} - \mathcal{O}^{(\ell)}\|_{L^2} \le \sum_{j=1}^{\ell} (2\sqrt{B})^{\ell-j}\frac{c_j}{\sqrt{n_j}}, \quad \ell = 2, \dots, L. \tag{36}$$

Now, we use the relation (35) to prove (36) by induction on $\ell = 2, \dots, L$. Taking $\ell = 2$ in (35), we have

$$\|\mathcal{O}^{(2)}_{\mathbf{n}_2} - \mathcal{O}^{(2)}\|_{L^2} \le B^{1/2}\frac{c_1}{\sqrt{n_1}} + \frac{c_2}{\sqrt{n_2}} \le 2B^{1/2}\frac{c_1}{\sqrt{n_1}} + \frac{c_2}{\sqrt{n_2}},$$

for example, (36) with $\ell = 2$. Now, suppose that

$$\|\mathcal{O}^{(k)}_{\mathbf{n}_k} - \mathcal{O}^{(k)}\|_{L^2} \le \sum_{j=1}^{k} (2\sqrt{B})^{k-j} \frac{c_j}{\sqrt{n_j}}, \quad \text{for any } k = 2, \dots, \ell - 1.$$

Then, by (35), we have

$$\begin{aligned}
\|\mathcal{O}^{(\ell)}_{\mathbf{n}_\ell} - \mathcal{O}^{(\ell)}\|_{L^2} &\le \sum_{k=2}^{\ell-1} B^{(\ell-k)/2} \sum_{j=1}^{k} 2^{k-j} B^{(k-j)/2} \frac{c_j}{\sqrt{n_j}} + \sum_{k=1}^{\ell} B^{(\ell-k)/2} \frac{c_k}{\sqrt{n_k}} \\
&\le \sum_{k=2}^{\ell-1} \sum_{j=1}^{k} 2^{k-j} B^{(\ell-j)/2} \frac{c_j}{\sqrt{n_j}} + \sum_{k=1}^{\ell} B^{(\ell-k)/2} \frac{c_k}{\sqrt{n_k}} \\
&\le \sum_{j=1}^{\ell-1} \sum_{k=j}^{\ell-1} 2^{k-j} B^{(\ell-j)/2} \frac{c_j}{\sqrt{n_j}} + \sum_{k=1}^{\ell} B^{(\ell-k)/2} \frac{c_k}{\sqrt{n_k}} \\
&= \sum_{j=1}^{\ell-1} \left(\sum_{k=j}^{\ell-1} 2^{k-j} + 1 \right) B^{(\ell-j)/2} \frac{c_j}{\sqrt{n_j}} + \frac{c_\ell}{\sqrt{n_\ell}} \\
&= \sum_{j=1}^{\ell-1} (2\sqrt{B})^{\ell-j} \frac{c_j}{\sqrt{n_j}} + \frac{c_\ell}{\sqrt{n_\ell}} = \sum_{j=1}^{\ell} (2\sqrt{B})^{\ell-j} \frac{c_j}{\sqrt{n_j}},
\end{aligned}$$

where we used that

$$\sum_{k=j}^{\ell-1} 2^{k-j} + 1 = \sum_{s=0}^{\ell-j-1} 2^s + 1 = 2^{\ell-j}, \quad \text{for any } j = 1, \dots, \ell - 1$$

because $\{2^s\}_{s \ge 0}$ is a geometric progression. The proof is completed. □

Proof of Proposition 5.2.

Proof of (i). The claim immediately follows noticing that, because of the nonnegativity of the quantities $\sqrt{C_b + C_W a_2}$ and $\sqrt{C_b + C_W a_1}$, we have $|\sigma(x\sqrt{C_b + C_W a_2})^2 - \sigma(x\sqrt{C_b + C_W a_1})^2| = 0$, for any $x \in \mathbb{R}$.

Proof of (ii). Because σ is Lipschitz continuous, we have

$$\begin{aligned}
&|\sigma(x\sqrt{C_b + C_W a_2})^2 - \sigma(x\sqrt{C_b + C_W a_1})^2| \\
&\quad = |\sigma(x\sqrt{C_b + C_W a_2}) - \sigma(x\sqrt{C_b + C_W a_1})||\sigma(x\sqrt{C_b + C_W a_2}) + \sigma(x\sqrt{C_b + C_W a_1})| \\
&\quad \le \mathrm{Lip}(\sigma)|x|\|\sqrt{C_b + C_W a_2} - \sqrt{C_b + C_W a_1}|(|\sigma(x\sqrt{C_b + C_W a_2})| + |\sigma(x\sqrt{C_b + C_W a_1})|) \\
&\quad \le \mathrm{Lip}(\sigma)|x|\|\sqrt{C_b + C_W a_2} - \sqrt{C_b + C_W a_1}|[2|\sigma(0)| \\
&\qquad + \mathrm{Lip}(\sigma)|x|(\sqrt{C_b + C_W a_2} + \sqrt{C_b + C_W a_1})],
\end{aligned} \tag{37}$$

where the latter inequality follows noticing that (again by the Lipschitz continuity of σ) for any $x, \kappa \in \mathbb{R}$,

$$|\sigma(\kappa x)| \le |\sigma(0)| + \mathrm{Lip}(\sigma(\cdot))|\kappa\|x|. \tag{38}$$

Multiplying and dividing the term in (37) by $\sqrt{C_b + C_W a_2} + \sqrt{C_b + C_W a_1}$, we easily have that

$$\begin{aligned}
&|\sigma(x\sqrt{C_b + C_W a_2})^2 - \sigma(x\sqrt{C_b + C_W a_1})^2| \\
&\quad \le C_W \mathrm{Lip}(\sigma)|x| \left(\frac{2|\sigma(0)|}{\sqrt{C_b + C_W a_2} + \sqrt{C_b + C_W a_1}} + \mathrm{Lip}(\sigma)|x| \right) |a_2 - a_1| \\
&\quad \le P(|x|)|a_2 - a_1|,
\end{aligned}$$

where $P(x)$ is defined by (24). As far as Condition (ii) of Theorem 5.1 is concerned, we note that by (38)

$$\sigma(\kappa Z)^4 \le (|\sigma(0)| + \mathrm{Lip}(\sigma)|\kappa\|Z|)^4, \quad \text{almost surely}$$

and so the claim is an immediate consequence of the fact that the absolute moments of Z are finite.

Proof of (iii). The proof is similar to the proof of (ii) and therefore is omitted.

Proof of (iv). If $\sigma(\cdot)^2$ is Lipschitz continuous, then

$$\begin{aligned}
&|\sigma(x\sqrt{C_b + C_W a_2})^2 - \sigma(x\sqrt{C_b + C_W a_1})^2| \\
&\quad \le \mathrm{Lip}(\sigma^2)|x|\,|\sqrt{C_b + C_W a_2} - \sqrt{C_b + C_W a_1}| \\
&\quad = \mathrm{Lip}(\sigma^2)|x|\,|\sqrt{C_b + C_W a_2} - \sqrt{C_b + C_W a_1}|\,\frac{\sqrt{C_b + C_W a_2} + \sqrt{C_b + C_W a_1}}{\sqrt{C_b + C_W a_2} + \sqrt{C_b + C_W a_1}} \\
&\quad = \mathrm{Lip}(\sigma^2)|x|\,|(C_b + C_W a_2) - (C_b + C_W a_1)|\,\frac{1}{\sqrt{C_b + C_W a_2} + \sqrt{C_b + C_W a_1}} \\
&\quad \le \frac{\mathrm{Lip}(\sigma^2) C_W}{2\sqrt{C_b}}\,|x|\,|a_2 - a_1|,
\end{aligned}$$

which shows that Condition (i) of Theorem 5.1 holds with $P(x)$ given by (26). Furthermore, using (38) with $\sigma(\cdot)^2$ in place of $\sigma(\cdot)$, for any $\kappa \in \mathbb{R}$,

$$\sigma(\kappa Z)^4 \le (|\sigma(0)|^2 + \mathrm{Lip}(\sigma(\cdot)^2)|\kappa|\,|Z|)^2, \quad \text{almost surely.}$$

Therefore, Condition (ii) of Theorem 5.1 is an immediate consequence of this latter relation and the fact that the absolute moments of Z are finite. □

6. Normal Approximation of Deep Random Gaussian NNs

6.1. Normal Approximation of Deep Random Gaussian NNs in the Convex Distance

The following theorem holds.

Theorem 6.1. *Let* $\mathrm{GNN}(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma; \mathbf{x}; \mathbf{b}, \mathbf{W})$ *be a deep random Gaussian NN, and let the notation and assumptions of Theorem* 5.1 *prevail. Then,*

$$\begin{aligned}
d_c(\mathbf{z}^{(L+1)}, \mathbf{z}) &\le C_1 \left(\sum_{k=1}^{L} [4\sqrt{2}\|P(|Z|)\|_{L^2}]^{L-k} \frac{c_k}{\sqrt{n_k}} \right) \\
&\le C_i \left(\sum_{k=1}^{L} [4\sqrt{2}\|P(|Z|)\|_{L^2}]^{L-k} \frac{c_k}{\sqrt{n_k}} \right), \quad i = 2, 3
\end{aligned}$$

where the constants c_k, $k = 1, \ldots, L$, *are defined by* (23),

$$C_1 = C_1(n_0, n_{L+1}, \mathbf{x}, \sigma, C_b, C_W) := C_W \left(\frac{80}{\mathfrak{s}_{L+1}^3} + \frac{48}{\mathfrak{s}_{L+1}^2} + 20\sqrt{2} \right) n_{L+1}^{59/24}, \tag{39}$$

$$C_2 = C_2(n_{L+1}, C_b, C_W) := C_W \left(\frac{80}{C_b^{3/2}} + \frac{48}{C_b} + 20\sqrt{2} \right) n_{L+1}^{59/24}$$

and

$$C_3 = C_3(n_0, n_{L+1}, \mathbf{x}, \sigma, C_W) := C_W \left(\frac{80}{(C_W \mathcal{O}^{(L+1)})^{3/2}} + \frac{48}{C_W \mathcal{O}^{(L+1)}} + 20\sqrt{2} \right) n_{L+1}^{59/24}.$$

Here, $\mathfrak{s}_{L+1} := \sqrt{C_b + C_W \mathcal{O}^{(L+1)}}$, *and clearly the upper bound with constant* C_2 *holds under the additional assumption that the biases are nonnull.*

Remark 6.2. Let $\mathrm{GNN}(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma; \mathbf{x}; \mathbf{b}, \mathbf{W})$ be a deep random Gaussian NN. Theorem 3.5 in Favaro et al. (2023) shows that if

$$c_2 n \le n_1, \ldots, n_L \le c_1 n, \quad \text{for some constants } c_1 \ge c_2 > 0 \text{ and some } n \ge 1 \tag{40}$$

and σ is polynomially bounded to order $r \ge 1$, then there exists a constant C_0 such that

$$d_c(\mathbf{z}^{(L+1)}, \mathbf{z}) \le C_0 n^{-1/2}.$$

We stress that the bounds in Theorem 6.1 provide (under different assumptions on σ and without any condition on the widths of the hidden layers) a detailed description of the analytical dependence of the upper estimate on the parameters of the model.

Proof. Throughout this proof, for ease of notation, we put

$$\kappa := d_c(\mathbf{z}^{(L+1)}, \mathbf{z}) \quad \text{and} \quad \gamma := C_W \|\mathcal{O}^{(L)}_{\mathbf{n}_L} - \mathcal{O}^{(L)}\|_{L^2}.$$

We preliminarily note that it suffices to prove that

$$\kappa \le (80\|\Sigma^{-1}_{n_{L+1}}\|^{3/2}_{op} + 48\|\Sigma^{-1}_{n_{L+1}}\|_{op} + 20\sqrt{2})n^{59/24}_{L+1}\gamma. \tag{41}$$

Indeed, the claim then follows by Theorem 5.1, noticing that

$$\|\Sigma^{-1}_{n_{L+1}}\|_{op} = \frac{1}{\mathfrak{s}_L} \tag{42}$$

and that $C_1 \le C_i$, $i = 2, 3$.

If $\gamma > 1/\mathrm{e}$, then the inequality (41) holds because $\kappa \le 1$ (which follows by the definition of the convex distance) and

$$20\sqrt{2}n^{59/24}_{L+1}\gamma > 20\sqrt{2}\gamma > 20\sqrt{2}/3 > 1.$$

From now on, we assume $\gamma \le 1/\mathrm{e}$. Let $h(\cdot) := \mathbf{1}_C(\cdot)$, where C is an arbitrarily fixed measurable convex set in $\mathbb{R}^{n_{L+1}}$, and let

$$h_t(\mathbf{y}) := \mathbb{E}h(\sqrt{t}\mathbf{z} + \sqrt{1-t}\mathbf{y}), \quad t \in (0,1), \quad \mathbf{y} \in \mathbb{R}^{n_{L+1}}.$$

For any $t \in (0,1)$, by Lemma 2.8(i),

$$f_{t,h}(\mathbf{y}) := \frac{1}{2}\int_t^1 \frac{1}{1-s}(\mathbb{E}h(\sqrt{t}\mathbf{z} + \sqrt{1-t}\mathbf{y}) - \mathbb{E}h(\mathbf{z}))\mathrm{d}s, \quad \mathbf{y} \in \mathbb{R}^{n_{L+1}}$$

solves the Stein Equation (5) with n_{L+1}, h_t, $\Sigma_{n_{L+1}}$ defined by (12), and $\mathbf{z}$, in place of d, g, Σ and $\mathbf{N}_\Sigma$, respectively, that is,

$$h_t(\mathbf{y}) - \mathbb{E}[h_t(\mathbf{z})] = \langle \mathbf{y}, \nabla f_{t,h}(\mathbf{y})\rangle_{n_{L+1}} - \langle \Sigma_{n_{L+1}}, \mathrm{Hess}\, f_{t,h}(\mathbf{y})\rangle_{H.S.}, \quad \mathbf{y} \in \mathbb{R}^{n_{L+1}}. \tag{43}$$

By Lemma 2.9, it follows that

$$\kappa \le \frac{4}{3}\sup_{h \in \mathcal{I}_{n_{L+1}}} |\mathbb{E}h_t(\mathbf{z}^{(L+1)}) - \mathbb{E}h_t(\mathbf{z})| + \frac{20n_{L+1}}{\sqrt{2}}\frac{\sqrt{t}}{1-t}, \quad t \in (0,1). \tag{44}$$

Without loss of generality, hereafter we assume that $\mathbf{z}$ is independent of $\mathcal{F}_L$. Therefore, by (43), we have

$$\mathbb{E}[h_t(\mathbf{z}^{(L+1)}) - h_t(\mathbf{z}) \mid \mathcal{F}_L] = \sum_{i=1}^{n_{L+1}} \mathbb{E}[z^{(L+1)}_i \partial_i f_{t,h}(\mathbf{z}^{(L+1)}) \mid \mathcal{F}_L] - \sum_{i,j=1}^{n_{L+1}} \Sigma_{n_{L+1}}(i,j)\mathbb{E}[\partial^2_{ij} f_{t,h}(\mathbf{z}^{(L+1)}) \mid \mathcal{F}_L]. \tag{45}$$

By Lemma 2.8(i), for any $t \in (0,1)$, the mapping $\mathbf{y} \longmapsto \partial_i f_{t,h}(\mathbf{y})$ is in $C^1(\mathbb{R}^{n_{L+1}})$ and has bounded first-order derivatives. Then, because $\mathbf{z}^{(L+1)}|\mathcal{F}_L$ is a centered Gaussian random vector with covariance matrix

$$\Sigma'_{n_{L+1}} := (C_b + C_W \mathcal{O}^{(L)}_{\mathbf{n}_L})\mathrm{Id}_{n_{L+1}}, \tag{46}$$

by Lemma 2.10(ii), we have

$$\mathbb{E}[z^{(L+1)}_i \partial_i f_{t,h}(\mathbf{z}^{(L+1)}) | \mathcal{F}_L] = \sum_{j=1}^{n_{L+1}} \Sigma'_{n_{L+1}}(i,j)\mathbb{E}[\partial^2_{ij} f_{t,h}(\mathbf{z}^{(L+1)}) | \mathcal{F}_L].$$

On combining this relation with (45) and taking the expectation, we have

$$\begin{aligned}
\mathbb{E}[h_t(\mathbf{z}^{(L+1)}) - h_t(\mathbf{z})] &= \sum_{i,j=1}^{n_{L+1}} \mathbb{E}[(\Sigma'_{n_{L+1}}(i,j) - \Sigma_{n_{L+1}}(i,j))\mathbb{E}[\partial^2_{ij} f_{t,h}(\mathbf{z}^{(L+1)}) \mid \mathcal{F}_L]] \\
&= \sum_{i,j=1}^{n_{L+1}} \mathbb{E}[\mathbb{E}[(\Sigma'_{n_{L+1}}(i,j) - \Sigma_{n_{L+1}}(i,j))\partial^2_{ij} f_{t,h}(\mathbf{z}^{(L+1)}) \mid \mathcal{F}_L]] \\
&= \mathbb{E}\langle \Sigma'_{n_{L+1}} - \Sigma_{n_{L+1}}, \mathrm{Hess}(f_{t,h}(\mathbf{z}^{(L+1)}))\rangle_{H.S.}
\end{aligned} \tag{47}$$

where in the second equality we used the $\mathcal{F}_L$-measurability of $\mathcal{O}^{(L)}_{\mathbf{n}_L}$. Taking the modulus on the relation (47) and applying the Cauchy-Schwarz inequality, we have

$$|\mathbb{E}[h_t(\mathbf{z}^{(L+1)}) - h_t(\mathbf{z})]| \le \sqrt{\mathbb{E}\|\Sigma'_{n_{L+1}} - \Sigma_{n_{L+1}}\|^2_{H.S.}}\sqrt{\mathbb{E}\|\mathrm{Hess}(f_{t,h}(\mathbf{z}^{(L+1)}))\|^2_{H.S.}}. \tag{48}$$

By Lemma 2.8(ii), for any $h \in \mathcal{I}_{n_{L+1}}$, we have

$$\mathbb{E}\|\mathrm{Hess}(f_{t,h}(\mathbf{z}^{(L+1)}))\|_{H.S.}^2 \le \|\Sigma_{n_{L+1}}^{-1}\|_{op}^2(n_{L+1}^2(\log t)^2\kappa + 530n_{L+1}^{17/6}), \quad t \in (0,1).$$

Moreover,

$$\mathbb{E}\|\Sigma'_{n_{L+1}} - \Sigma_{n_{L+1}}\|_{H.S.}^2 = n_{L+1}C_W^2\|\mathcal{O}_{\mathbf{n}_L}^{(L)} - \mathcal{O}^{(L)}\|_{L^2}^2. \tag{49}$$

On combining these relations and using the elementary inequality $\sqrt{a_1+a_2} \le \sqrt{a}_1 + \sqrt{a}_2$, $a_1, a_2 \ge 0$, we have

$$\begin{aligned}\sup_{C\in\mathcal{C}_{n_{L+1}}} &|\mathbb{P}(\sqrt{t}\mathbf{z} + \sqrt{1-t}\mathbf{z}^{(L+1)} \in C) - \mathbb{P}(\sqrt{t}\mathbf{z} + \sqrt{1-t}\mathbf{z} \in C)| \\ &\le C_W\|\Sigma_{n_{L+1}}^{-1}\|_{op}(n_{L+1}^{3/2}|\log t|\sqrt{\kappa} + 24n_{L+1}^{23/12})\|\mathcal{O}_{\mathbf{n}_L}^{(L)} - \mathcal{O}^{(L)}\|_{L^2}.\end{aligned}$$

On combining this latter inequality with (44), we have

$$\kappa \le \frac{4}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}(n_{L+1}^{3/2}|\log t|\sqrt{\kappa} + 24n_{L+1}^{23/12})\gamma + \frac{20n_{L+1}}{\sqrt{2}}\frac{\sqrt{t}}{1-t}, \quad t \in (0,1). \tag{50}$$

Because $\kappa \le 1$, by this relation, we have

$$\kappa \le \frac{4}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}(n_{L+1}^{3/2}|\log t| + 24n_{L+1}^{23/12})\gamma + \frac{20n_{L+1}}{\sqrt{2}}\frac{\sqrt{t}}{1-t}, \quad t \in (0,1).$$

Setting $t = \gamma^2$ in this latter inequality (note that this choice of the parameter t is admissible because $\gamma \le 1/\mathrm{e} < 1$), we have

$$\begin{aligned}\kappa &\le \frac{4}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}(2n_{L+1}^{3/2}|\log\gamma| + 24n_{L+1}^{23/12})\gamma + \frac{20n_{L+1}}{\sqrt{2}}\frac{\gamma}{1-\gamma^2} \\ &\le \frac{4}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}(2n_{L+1}^{3/2}|\log\gamma| + 24n_{L+1}^{23/12})\gamma + 20\sqrt{2}n_{L+1}\gamma,\end{aligned} \tag{51}$$

where in the latter inequality we used the relation

$$\frac{1}{\sqrt{2}}\frac{\gamma}{1-\gamma^2} \le \sqrt{2}\gamma, \tag{52}$$

which holds because $\gamma \le 1/\mathrm{e} < 1/\sqrt{2}$. We rewrite the inequality (51) as

$$\kappa \le \frac{8}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}n_{L+1}^{3/2}\gamma|\log\gamma| + (32n_{L+1}^{23/12}\|\Sigma_{n_{L+1}}^{-1}\|_{op} + 20\sqrt{2}n_{L+1})\gamma.$$

Taking the square root and multiplying by $|\log\gamma|$, we have

$$\begin{aligned}|\log\gamma|\sqrt{\kappa} \le &\sqrt{\frac{8}{3}}\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2}n_{L+1}^{3/4}\gamma^{1/2}|\log\gamma|^{3/2} \\ &+ (32n_{L+1}^{23/12}\|\Sigma_{n_{L+1}}^{-1}\|_{op} + 20\sqrt{2}n_{L+1})^{1/2}\gamma^{1/2}|\log\gamma|.\end{aligned}$$

Because $\max\{\sup_{y\in(0,1/\mathrm{e}]}y^{1/2}|\log y|^{3/2}, \sup_{y\in(0,1/\mathrm{e}]}y^{1/2}|\log y|^{1/2}\} \le 4$, we have

$$\begin{aligned}|\log\gamma|\sqrt{\kappa} &\le 4\sqrt{\frac{8}{3}}\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2}n_{L+1}^{3/4} + 4(32n_{L+1}^{23/12}\|\Sigma_{n_{L+1}}^{-1}\|_{op} + 20\sqrt{2}n_{L+1})^{1/2} \\ &= 8\frac{\sqrt{6}}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2}n_{L+1}^{3/4} + 4(32n_{L+1}^{23/12}\|\Sigma_{n_{L+1}}^{-1}\|_{op} + 20\sqrt{2}n_{L+1})^{1/2} \\ &\le 8\frac{\sqrt{6}}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2}n_{L+1}^{3/4} + 16\sqrt{2}n_{L+1}^{23/24}\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2} + \sqrt{20\sqrt{2}}n_{L+1}^{1/2} \\ &\le \left[\left(8\frac{\sqrt{6}}{3} + 16\sqrt{2}\right)\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2} + \sqrt{20\sqrt{2}}\right]n_{L+1}^{23/24} \\ &\le (30\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2} + 6)n_{L+1}^{23/24}.\end{aligned} \tag{53}$$

By (50) with $t = \gamma^2$, (52), and (53), we finally have (41); indeed,

$$\begin{aligned}\kappa &\le \frac{4}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}(2n_{L+1}^{3/2}\,|\log\gamma|\,\sqrt{\kappa} + 24n_{L+1}^{23/12})\gamma + 20\sqrt{2}n_{L+1}\gamma \\ &\le \frac{4}{3}\|\Sigma_{n_{L+1}}^{-1}\|_{op}[(60\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{1/2} + 12)n_{L+1}^{59/24} + 24n_{L+1}^{23/12}]\gamma + 20\sqrt{2}n_{L+1}\gamma \\ &\le (80\|\Sigma_{n_{L+1}}^{-1}\|_{op}^{3/2} + 48\|\Sigma_{n_{L+1}}^{-1}\|_{op} + 20\sqrt{2})n_{L+1}^{59/24}\gamma.\end{aligned}$$

The proof is completed. □

6.2. Normal Approximation of Deep Random Gaussian NNs in the 1-Wasserstein Distance

The following theorem holds.

Theorem 6.3. *Let* $\mathrm{GNN}(L, n_0, n_{L+1}, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ *be a deep random Gaussian NN, and let the notation and assumptions of Theorem* 5.1 *prevail. Then,*

$$\begin{aligned}d_{W_1}(\mathbf{z}^{(L+1)}, \mathbf{z}) &\le K_1\left(\sum_{k=1}^{L}[4\sqrt{2}\|P(|Z|)\|_{L^2}]^{L-k}\frac{c_k}{\sqrt{n_k}}\right) \\ &\le K_i\left(\sum_{k=1}^{L}[4\sqrt{2}\|P(|Z|)\|_{L^2}]^{L-k}\frac{c_k}{\sqrt{n_k}}\right), \quad i = 2,3\end{aligned}$$

where the constants c_k, $k = 1, \ldots, L$, *are defined by* (23),

$$\begin{aligned}K_1 &= K_1(n_0, n_{L+1}, \mathbf{x}, \sigma, C_b, C_W) := \frac{n_{L+1}C_W}{\mathfrak{s}_L}, \\ K_2 &= K_2(n_{L+1}, C_b, C_W) := \frac{n_{L+1}C_W}{\sqrt{C_b}}\end{aligned}$$

and

$$K_3 = K_3(n_0, n_{L+1}, \mathbf{x}, \sigma, C_W) := \frac{n_{L+1}\sqrt{C_W}}{\sqrt{\mathcal{O}^{(L)}}}.$$

Here again, the upper estimate with the constant K_2 *holds if* $C_b > 0$.

Remark 6.4. Let $\mathrm{GNN}(L, n_0, 1, \mathbf{n}_L, \sigma, \mathbf{x}, \mathbf{b}, \mathbf{W})$ be a deep random Gassian NN with univariate output, widths of the hidden layers satisfying (40), and a polynomially bounded to order $r \ge 1$ activation function σ. Then, by theorem 3.3 in Favaro et al. (2023), we have that there exist two constants $C, C_0 > 0$ such that

$$\frac{C_0}{n} \le d_{W_1}(z^{(L+1)}, z) \le \frac{C}{n}.$$

Clearly, this inequality shows the optimality of the rate $1/n$. Here again, we note that the corresponding bound provided by Theorem 6.3 gives (under different assumptions on σ and without any condition on the widths of the hidden layers) a detailed description of the analytical dependence of the upper estimate on the parameters of the model.

Proof. Let $g \in \mathcal{L}_{n_{L+1}}(1)$ be arbitrarily fixed. Without loss of generality, we assume that $\mathbf{z}$ is independent of $\mathcal{F}_L$. By Lemma 2.7, we then have

$$\mathbb{E}[g(\mathbf{z}^{(L+1)}) - g(\mathbf{z}) \mid \mathcal{F}_L] = \sum_{i=1}^{n_{L+1}}\mathbb{E}[z_i^{(L+1)}\partial_i f_g(\mathbf{z}^{(L+1)}) \mid \mathcal{F}_L] - \sum_{i,j=1}^{n_{L+1}}\Sigma_{n_{L+1}}(i,j)\mathbb{E}[\partial_{ij}^2 f_g(\mathbf{z}^{(L+1)}) \mid \mathcal{F}_L], \tag{54}$$

where

$$f_g(\mathbf{y}) := \int_0^\infty \mathbb{E}[g(\mathbf{z}) - g(\mathrm{e}^{-t}\mathbf{y} + \sqrt{1-\mathrm{e}^{-2t}}\mathbf{z})]\,\mathrm{d}t, \quad \mathbf{y} \in \mathbb{R}^{n_{L+1}}.$$

Again, by Lemma 2.7, we have that the mapping $\mathbf{y} \longmapsto \partial_i f_g(\mathbf{y})$ is in $C^1(\mathbb{R}^{n_{L+1}})$ and has bounded first-order derivatives. Applying Lemma 2.10(ii) exactly as in the proof of Theorem 6.1 (see a few lines before Equation (47)), we have

$$\mathbb{E}[g(\mathbf{z}^{(L+1)}) - g(\mathbf{z})] = \mathbb{E}\langle\Sigma'_{n_{L+1}} - \Sigma_{n_{L+1}}, \mathrm{Hess}(f_g(\mathbf{z}^{(L+1)}))\rangle_{H.S.},$$

where the matrix $\Sigma'_{n_{L+1}}$ is defined by (46). Therefore, applying the Cauchy-Schwarz inequality as in (48), we have

$$|\mathbb{E}[g(\mathbf{z}^{(L+1)}) - g(\mathbf{z})]| \le \sqrt{\mathbb{E}\|\Sigma'_{n_{L+1}} - \Sigma_{n_{L+1}}\|_{H.S.}^2}\sqrt{\mathbb{E}\|\mathrm{Hess}(f_g(\mathbf{z}^{(L+1)}))\|_{H.S.}^2} \tag{55}$$

By Lemma 2.7, we have

$$\sqrt{\mathbb{E}\|\mathrm{Hess}(f_g(\mathbf{z}^{(L+1)}))\|_{H.S.}^2} \le \sup_{\mathbf{y}\in\mathbb{R}^{n_{L+1}}} \|\mathrm{Hess} f_g(\mathbf{y})\|_{H.S.} \le \sqrt{n_{L+1}}\|\Sigma_{n_{L+1}}^{-1}\|_{op}\|\Sigma_{n_{L+1}}\|_{op}^{1/2},$$

On combining these relations with (49), we have

$$|\mathbb{E}[g(\mathbf{z}^{(L+1)}) - g(\mathbf{z})]| \le n_{L+1} C_W \|\Sigma_{n_{L+1}}^{-1}\|_{op}\|\Sigma_{n_{L+1}}\|_{op}^{1/2}\|\mathcal{O}_{\mathbf{n}_L}^{(L)} - \mathcal{O}^{(L)}\|_{L^2}.$$

The claim follows taking the supremum over g on this inequality and then using Theorem 5.1, relation (42), and the fact that $\|\Sigma_{n_{L+1}}\|_{op} = \mathfrak{s}_L^2$. □

7. Localization of the Output

In this section, we want to show the potentiality of the obtained results for practical applications. Indeed, both Theorem 4.1(ii) and Theorem 6.1 allow one to explicitly estimate the probability that the output of a random Gaussian NN evaluated at the input $\mathbf{x}$ belongs to a suitable set without resorting to computationally expensive Monte Carlo simulations. This is what we call output localization, which can suggest a suitable architecture design to estimate a target function f. Indeed, in statistical learning, given a training set

$$\{(\mathbf{x}_i, f(\mathbf{x}_i)\}_{i\in I} \subset \mathbb{R}^{n_0} \times \mathbb{R}^{n_{L+1}},$$

the goal is to estimate the unknown function f by a NN with a certain fixed architecture. This is performed by minimizing an empirical risk function over the space of NN's parameters, that is, the biases and the weights. Such kinds of minimization problems are nonconvex and are usually addressed by a gradient descent procedure; the latter stabilizes toward a solution that depends on the initialization point. Because in practical applications the initialization point is given by a realization of a random Gaussian NN, it can be convenient to choose L, $\mathbf{n}_L$, σ, C_b, and C_W by exploiting output localization, that is, in such a way to have a good estimate of the probability $\mathbb{P}(\mathbf{z}^{(L+1)}(\mathbf{x}_i) \in V_i)$, $i \in I$, where V_i is an appropriate neighborhood of $f(\mathbf{x}_i)$.

Hence, in the following we want to show how to use the upper bound on the total variation distance provided in Theorem 4.1(ii), for a shallow random Gaussian NN with univariate output, and the upper bound on the convex distance provided by Theorem 6.1, for a deep random Gaussian NN, to localize the output. Indeed, given a measurable convex set $V \subset \mathbb{R}^{n_{L+1}}$, by the definitions of both the total variation and the convex distances, we have

$$\mathbb{P}(\mathbf{z} \in V) - C_{bound} \le \mathbb{P}(\mathbf{z}^{(L+1)} \in V) \le \mathbb{P}(\mathbf{z} \in V) + C_{bound},$$

where $L = 1$, $n_2 = 1$, and

$$C_{bound} := 2\frac{C_W\sqrt{\mathbb{V}\mathrm{ar}(\sigma(\mathfrak{s}_0 Z)^2)}}{C_b + C_W\mathbb{E}\sigma(\mathfrak{s}_0 Z)^2}\frac{1}{\sqrt{n_1}}, \tag{56}$$

for the case of a shallow NN, whereas

$$C_{bound} := C_1\left(\sum_{k=1}^{L}[4\sqrt{2}\|P(|Z|)\|_{L^2}]^{L-k}\frac{c_k}{\sqrt{n_k}}\right), \tag{57}$$

for the case of a deep NN, where C_1 is given by (39) and the constants c_k, $k = 1, \dots, L$, are defined by (23).

Let $V := \prod_{i=1}^{n_{L+1}}[r_i, s_i]$ be a rectangle of $\mathbb{R}^{n_{L+1}}$. Because $\mathbf{z}$ is an n_{L+1}-dimensional centered Gaussian random vector with covariance matrix (12), we have

$$\mathbb{P}(\mathbf{z} \in V) = \prod_{i=1}^{n_{L+1}}\left(\mathbb{P}\left(Z \le \frac{s_i}{\mathfrak{s}_L}\right) - \mathbb{P}\left(Z \le \frac{r_i}{\mathfrak{s}_L}\right)\right).$$

Now, we furnish numerical values of the constant C_{bound} in (56) and (57); in both cases, we take $\sigma(x) := x\mathbf{1}\{x \ge 0\}$, that is, a ReLU activation function. Because the ReLU function is Lipschitz continuous with Lipschitz constant equal to 1, $\mathbb{E}Z^2 = 1$ and $\mathbb{E}Z^4 = 3$, by Proposition 5.2(ii), we have

$$\|P(|Z|)\|_{L^2} = C_W\sqrt{\mathbb{E}Z^4\mathbf{1}\{Z \ge 0\}} = C_W\sqrt{3/2}.$$

By the expression of the constants c_ℓ in Theorem 5.1, we have

$$c_\ell = \mathfrak{s}_{\ell-1}^2\sqrt{2\mathbb{E}Z^4\mathbf{1}\{Z \ge 0\}} = \mathfrak{s}_{\ell-1}^2\sqrt{3}, \quad \ell = 1, \dots, L.$$

Table 2. Values of C_{bound} in (56) for Different Inputs $\mathbf{x}$ for a Shallow Random Gaussian NN with Architecture $L = 1$, $n_0 = 4$, $n_1 = n$, $n_2 = 1$, and ReLU Activation Function

$\mathbf{x} = (0, 0, 0, 0)$																		
n	1			10			10^2			10^3			10^4			10^5		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.02	0.21	1.49	0.01	0.07	0.47	0.00	0.02	0.15	0.00	0.01	0.05	0.00	0.00	0.01	0.00	0.00	0.00
10	0.01	0.07	0.47	0.00	0.02	0.15	0.00	0.01	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$\mathbf{x} = (0.1, 0.1, 0.1, 0.1)$																		
n	1			10			10^2			10^3			10^4			10^5		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.02	0.21	1.49	0.01	0.07	0.47	0.00	0.02	0.15	0.00	0.01	0.05	0.00	0.00	0.01	0.00	0.00	0.00
10	0.01	0.07	0.47	0.00	0.02	0.15	0.00	0.01	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$\mathbf{x} = (0.5, -0.5, 0.5, -0.5)$																		
n	1			10			10^2			10^3			10^4			10^5		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.02	0.22	1.54	0.01	0.07	0.49	0.00	0.02	0.15	0.00	0.01	0.05	0.00	0.00	0.02	0.00	0.00	0.00
10	0.01	0.07	0.47	0.00	0.02	0.15	0.00	0.01	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$\mathbf{x} = (10, 10, 10, 10)$																		
n	1			10			10^2			10^3			10^4			10^5		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.03	0.48	0.44	0.01	0.15	0.14	0.00	0.05	0.04	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
10	0.01	0.09	0.36	0.00	0.03	0.11	0.00	0.01	0.04	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Table 3. Values of C_{bound} in (57) for Different Inputs $\mathbf{x}$ for a Deep Random Gaussian NN with Architecture $L = 3$, $n_0 = 4$, $n_1 = n_2 = n_3 = n$, $n_4 = 1$, and ReLU Activation Function

$\mathbf{x} = (0, 0, 0, 0)$																		
n	10^4			10^5			10^6			10^7			10^8			10^9		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.03	0.58	88.59	0.01	0.18	28.01	0.00	0.06	8.86	0.00	0.02	2.80	0.00	0.01	0.89	0.00	0.00	0.28
10	0.07	1.38	331.57	0.02	0.44	104.85	0.01	0.14	33.16	0.00	0.04	10.49	0.00	0.01	3.32	0.00	0.00	1.05
$\mathbf{x} = (0.1, 0.1, 0.1, 0.1)$																		
n	10^4			10^5			10^6			10^7			10^8			10^9		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.03	0.58	89.30	0.01	0.18	28.24	0.00	0.06	8.93	0.00	0.02	2.82	0.00	0.01	0.89	0.00	0.00	0.28
10	0.07	1.38	331.85	0.02	0.44	104.94	0.01	0.14	33.18	0.00	0.04	10.49	0.00	0.01	3.32	0.00	0.00	1.05
$\mathbf{x} = (0.5, -0.5, 0.5, -0.5)$																		
n	10^4			10^5			10^6			10^7			10^8			10^9		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.03	0.58	106.14	0.01	0.18	33.56	0.00	0.06	10.61	0.00	0.02	3.36	0.00	0.01	1.06	0.00	0.00	0.34
10	0.07	1.38	338.62	0.02	0.44	107.08	0.01	0.14	33.86	0.00	0.04	10.71	0.00	0.01	3.39	0.00	0.00	1.07
$\mathbf{x} = (10, 10, 10, 10)$																		
n	10^4			10^5			10^6			10^7			10^8			10^9		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	0.03	1.90	2,998.50	0.01	0.60	948.21	0.00	0.19	299.85	0.00	0.06	94.82	0.00	0.02	29.99	0.00	0.01	9.48
10	0.07	1.69	3,027.47	0.02	0.54	957.37	0.01	0.17	302.75	0.00	0.05	95.74	0.00	0.02	30.27	0.00	0.01	9.57

Table 4. Values of C_1 in (39) for Different Inputs $\mathbf{x}$ for a Deep Random Gaussian NN with Architecture $L = 3$, $n_0 = 4$, $n_1 = n_2 = n_3 = n$, $n_4 = 1$, and ReLU Activation Function

	$\mathbf{x} = (0,0,0,0)$			$\mathbf{x} = (0.1,0.1,0.1,0.1)$			$\mathbf{x} = (0.5,-0.5,0.5,-0.5)$			$\mathbf{x} = (10,10,10,10)$		
C_b \ C_W	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
1	1.55	14.80	85.04	1.55	14.80	85.00	1.55	14.80	83.86	1.55	14.78	33.09
10	0.36	3.52	31.83	0.36	3.52	31.83	0.36	3.52	31.82	0.36	3.52	30.28

By the definition of the quantities $\mathcal{O}^{(\ell)}$, $\ell = 1,\dots,L$, (see (10)), we have

$$\mathcal{O}^{(\ell)} = \mathfrak{s}_{\ell-1}^2 \mathbb{E}Z^2\mathbf{1}\{Z \geq 0\} = \mathfrak{s}_{\ell-1}^2/2, \quad \ell = 1,\dots,L$$

with $\mathcal{O}^{(0)}$ given by (11). Therefore,

$$\mathcal{O}^{(\ell)} = \frac{C_b}{2}\sum_{k=0}^{\ell-1}\frac{C_W^k}{2^k} + \frac{C_W^\ell}{2^\ell}\mathcal{O}^{(0)}.$$

Table 2 gives the values of C_{bound} in (56) in the case of a shallow random Gaussian NN with architecture $L = 1$, $n_0 = 4$, $n_1 = n$, and $n_2 = 1$ for different values of $n \in \{1, 10, 10^2, 10^3, 10^4, 10^5\}$. Table 3 gives the values of C_{bound} in (57) in the case of a deep random Gaussian NN with architecture $L = 3$, $n_0 = 4$, $n_1 = n_2 = n_3 = n$, and $n_4 = 1$, for different values of $n \in \{10^4, 10^5, 10^6, 10^7, 10^8, 10^9\}$. In both tables, we consider four different inputs,

$$\mathbf{x} \in \{(0,0,0,0),(0.1,0.1,0.1,0.1),(0.5,-0.5,0.5,-0.5),(10,10,10,10)\},$$

whose Euclidean norm is, respectively, 0, strictly less than 1, 1 and strictly larger than 1, and

$$C_b \in \{1, 10\} \quad \text{and} \quad C_W \in \{0.01, 0.1, 1\}.$$

It is clear from Table 2 that for a shallow random Gaussian NN, the Gaussian approximation of the output is very good for any of the choices of the parameters n, C_b, and C_W and of the input $\mathbf{x}$. It is also clear from Table 3 that for a deep random Gaussian NN, such as the considered NN with three hidden layers, the Gaussian approximation of the output is very good for some choices of the parameters C_b, C_W, and n and of the input $\mathbf{x}$ (also when the number n of neurons in the hidden layers is not excessively large), but it is very poor for other choices of these quantities. In order to analyze the influence of the parameters C_b and C_W and of the input $\mathbf{x}$ on the value of C_{bound} in (57), in Table 4 we reported the value of the constant C_1 that appears in (57), and it is explicitly given in (39). As it can be seen from this table, the value of C_{bound} strongly depends on the value of C_1, which, in turn, is closely related to the choice of the parameters C_b and C_W and does not depend much on the norm of the input vector $\mathbf{x}$.

References

Balasubramanian K, Goldstein L, Ross N, Salim A (2024) Gaussian random field approximation via Stein's method with applications to wide random neural networks. *Appl. Comput. Harmon. Anal.* 72:1–38.

Basteri A, Trevisan D (2024) Quantitative Gaussian approximation of randomly initialized deep neural networks. *Machine Learning* 113:6373–6393.

Benktus V (2003) On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Planning Inference* 113:385–402.

Bordino A, Favaro S, Fortini S (2024) Non-asymptotic approximations of Gaussian neural networks via second-order Poincaré inequalities. *Proc. 6th Sympos. Adv. Approximate Bayesian Inference*, 45–78.

Cammarota V, Marinucci D, Salvi M, Vigogna S (2024) A quantitative functional central limit theorem for shallow neural networks. *Modern Stochastics Theory Appl.* 11:1–24.

Chen LHY, Goldstein L, Shao QM (2011) *Normal Approximation by Stein's Method* (Springer, Heidelberg, Germany).

Eldan R, Mikulincer D, Schramm T (2021) Non-asymptotic approximations of neural networks by Gaussian processes. *Conf. Learn. Theory*, 1754–1775.

Favaro S, Hanin B, Marinucci D, Nourdin I, Peccati G (2023) Quantitative CLTs in deep networks networks. Preprint, submitted July 12, https://arxiv.org/pdf/2307.06092.

Hanin B (2023) Random neural networks in the infinite width limit as Gaussian processes. *Ann. Appl. Probab.* 33:4798–4819.

Klukowski A (2022) Rate of convergence of polynomial networks to Gaussian processes. *Conf. Learn. Theory*, 701–722.

Lee J, Bahri Y, Novak R, Schoenholz SS, Pennington J, Sohl-Dickstein J (2018) Deep neural networks as Gaussian processes. *Internat. Conf. Learn. Representations* (ICLR, Appleton, WI).

Macci C, Pacchiarotti B, Torrisi GL (2024) Large and moderate deviations for Gaussian neural networks. Preprint, submitted June 25, https://arxiv.org/pdf/2401.01611.

Matthews AGG, Rowland M, Hron J, Turner RE, Ghahramani Z (2018) Gaussian process behaviour in wide deep neural networks. *Internat. Conf. Learn. Representations* (ICLR, Appleton, WI), 4:77–86.

Nazarov F (2004) On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. Morel J-M, Teissier B, eds. *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics (Springer, Berlin), 169–187.

Neal RM (1996) Priors for infinite networks. Bickel P, Diggle P, Fienberg S, Krickeberg K, Olkin I, Wermuth N, Zeger S, eds. *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, vol. 118 (Springer, New York), 29–53.

Nourdin I, Peccati G (2012) *Normal Approximations with Malliavin Calculus* (Cambridge University Press, Cambridge, UK).

Nourdin I, Peccati G, Yang X (2022) Multivariate normal approximations on the Wiener space: New bounds in the convex distance. *J. Theoret. Probab.* 35:2020–2037.

Roberts DA, Yaida S, Hanin B (2022) *The Principles of Deep Learning Theory. An Effective Theory Approach to Understanding Neural Networks* (Cambridge University Press, Cambridge, MA).

Schulte M, Yukich JE (2019) Multivariate second order Poincaré inequalities for Poisson functionals. *Electron. J. Probab.* 24:1–42.

Stein C (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Berkeley Symp. Math. Statist. Prob.* 6.2:583–602.

Torrisi GL (2023) Quantitative multidimensional central limit theorems for means of the Dirichlet-Ferguson measure. *ALEA - Latin Amer. J. Prob. Math. Statist.* 20:825–860.

Villani C (2009) *Optimal Transport: Old and New* (Springer, New York).

Yaida S (2019) Non-Gaussian processes and neural networks at finite widths. Preprint, submitted September 30, https://arxiv.org/abs/1910.00019.

Yang G (2019) Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 32 (Curran Associates, Red Hook, NY).