GAUSSIAN RANDOM FIELD APPROXIMATION VIA STEIN'S METHOD WITH APPLICATIONS TO WIDE RANDOM NEURAL NETWORKS

Krishnakumar Balasubramanian
University of California, Davis
kbala@ucdavis.edu

Larry Goldstein
University of Southern California
larry@usc.edu

Nathan Ross
University of Melbourne
nathan.ross@unimelb.edu.au

Adil Salim
Microsoft Research
adilsalim@microsoft.com

May 2, 2024

Abstract

We derive upper bounds on the Wasserstein distance (W_1), with respect to sup-norm, between any continuous $\mathbb{R}^d$ valued random field indexed by the n-sphere and the Gaussian, based on Stein's method. We develop a novel Gaussian smoothing technique that allows us to transfer a bound in a smoother metric to the W_1 distance. The smoothing is based on covariance functions constructed using powers of Laplacian operators, designed so that the associated Gaussian process has a tractable Cameron-Martin or Reproducing Kernel Hilbert Space. This feature enables us to move beyond one dimensional interval-based index sets that were previously considered in the literature. Specializing our general result, we obtain the first bounds on the Gaussian random field approximation of wide random neural networks of any depth and Lipschitz activation functions at the random field level. Our bounds are explicitly expressed in terms of the widths of the network and moments of the random weights. We also obtain tighter bounds when the activation function has three bounded derivatives.

Keywords: Distributional approximation, Gaussian random field, Stein's method, Laplacian-based smoothing, Deep neural networks.

CONTENTS

1 INTRODUCTION

Random fields that arise in a variety of applications related to deep learning (Neal, 1996; Lee et al., 2018; de G. Matthews et al., 2018; Yang, 2019; Hanin, 2023) and stochastic optimization (Benveniste et al., 2012; Sirignano and Spiliopoulos, 2020; Chen et al., 2020; Rotskoff and Vanden-Eijnden, 2022; Balasubramanian et al., 2023) can exhibit limiting Gaussian behavior, rigorously understood through the theory of weak convergence. Combining this asymptotic behavior with the comprehensive theory of Gaussian random fields leads to insights about the qualitative and quantitative behavior of the random field of interest. In order to justify the accuracy of the approximation of quantities of interest by those of their limits, it is important to quantify the error in the Gaussian random field approximation. Indeed, in the standard multivariate central limit theorem, Berry-Esseen bounds precisely determine when the Gaussian behavior "kicks-in". Our main goal in this work is to develop such quantitative Berry-Esseen-type bounds for Gaussian random field approximations via Stein's method. We focus in particular on bounds in the Wasserstein metric (W_1) with respect to sup-norm, and highlight that convergence of these bounds to zero implies asymptotic weak convergence. Moreover, such bounds immediately imply Wasserstein bounds between important statistics of the fields, such as finite-dimensional distributions and extrema.

Stein's method has been extensively developed to provide quantitative distributional approximation bounds in both the Gaussian and non-Gaussian settings; we refer to Chen, Goldstein, and Shao (2011); Ross (2011); Nourdin and Peccati (2012) for a detailed treatment of the former. Recent works (see, for example, Barbour et al. (2024, Section 1.1)) have focused on developing Stein's method to derive Gaussian process approximations results. These works pertain to random process indexed by the interval $[0, T]$, for some $T < \infty$. As is common in Stein's method, bounds are first developed in some "smooth" metric and are then transferred to the metric of interest, such as the Wasserstein, Lévy-Prokhorov or Kolmogorov metrics, via various smoothing techniques.

For instance, Barbour et al. (2024, Lemma 1.10) develops an infinite-dimensional analog of a widely-used finite-dimensional Gaussian smoothing technique. Based on this foundation, the authors establish Gaussian process approximation bounds for processes indexed by the interval $[0, T]$, in the W_1 and Lévy-Prokhorov metrics. However, their smoothing technique is restricted to random processes indexed by some subset of the real line, as it relies on a detailed understanding of the Cameron-Martin space of one-dimensional Brownian motion. As there are no canonical Gaussian random fields *indexed* by more general sets, e.g., the n-sphere, which have explicit Cameron-Martin spaces, new ideas are required to adapt these smoothing techniques to this setting.

A main contribution of this work is the development of a novel smoothing technique which can be used in conjunction with Stein's method to derive Gaussian random field approximation bounds in the W_1 metric. The smoothing technique is based on the construction of a Gaussian random field with an explicit Cameron-Martin space via Laplacian operators. Though we focus on the case of random fields indexed by the n-sphere $\mathcal{S}^n$, our approach is generally applicable to random fields indexed by any compact metric measure space $\mathcal{M}$, subject to increased technical complexity.

We apply our general result to derive quantitative bounds for the W_1 distance between the output of a wide random neural network indexed by inputs in $\mathcal{S}^n$ and the corresponding Gaussian random field. Though wide random neural networks produce highly complicated random fields, such bounds allow them to be studied via their more tractable limiting Gaussian behavior. In the one hidden layer case, Neal (1996) argues that wide random neural networks asymptotically behave as Gaussian

Notation	Description
$(\mathcal{M}, \mathtt{d}, \nu)$	Metric space $(\mathcal{M}, \mathtt{d})$ equipped with a measure ν
$\mathcal{S}^n$	n-sphere
$\mathrm{C}(\mathcal{M}; \mathbb{R}^d)$	Banach space of continuous functions equipped with sup-norm
(ε, δ)	Regularization and smoothing parameters respectively
F, H	Random fields in $\mathrm{C}(\mathcal{M}; \mathbb{R}^d)$
G	Gaussian Random Field used to approximate $F \in \mathrm{C}(\mathcal{M}; \mathbb{R}^d)$
S	Smoothing Gaussian Random Field
$\mathtt{d}_{\mathcal{H}}(F, H)$	Integral probability metric over a class of test functions $\mathcal{H}$
D^k	k-th order Fréchet derivative

Table 1: Summary of some main notations used.

random fields. The works of de G. Matthews et al. (2018) and Lee et al. (2018) give heuristic and empirical evidence that general depth neural networks exhibit Gaussian random field limits. Very recently, Hanin (2023) proves that deep neural networks converge weakly to appropriately defined Gaussian random fields as the layer widths tend to infinity. At a high-level, one proceeds here by first establishing convergence of finite dimensional distributions, which typically follows directly from the multivariate CLT. Weak convergence then follows from tightness results. In a different but related direction, Li et al. (2022) provide a characterization of the limiting covariance matrix of the output of the neural network when evaluated at a finite-set of points, as the depth and width tends to infinity at the same rate.

From a quantitative point of view, the question of how wide a random neural network has to be in order that the limiting Gaussian random field provides a good approximation is left unanswered by results that only demonstrate weak convergence. Works that addresses this gap include Eldan et al. (2021); Basteri and Trevisan (2022); Klukowski (2022); Bordino et al. (2023b), discussed in more detail in Section 1.2.1. However, results currently known to us have at least one of the following drawbacks: they (i) work in weaker topologies, such as Wasserstein metrics with respect to integral (e.g., L^2) distances, rather than the sup-norm, (ii) only provide approximation bounds for finite dimensional distributions, and not at the random field level, (iii) require Gaussian or similar restrictive assumptions on the random weights, (iv) consider special cases like one hidden-layer neural networks or use restricted activation functions, such as polynomials. In contrast, our work provides precise quantitative bounds for the error in approximating wide random neural networks with Gaussian random fields, without any of the above-mentioned restrictions.

In the remainder of the introduction, we state and discuss our main results. Section 1.1 is devoted to our smoothing result, Theorem 1.1. Section 1.2 contains our Gaussian approximation results for wide random neural networks, Theorems 1.2 and 1.4.

1.1 Bounds for random field approximations

We now formally describe our setting and main result. Consider a compact metric space $(\mathcal{M}, \mathtt{d})$, equipped with a finite Borel measure ν that is positive on open balls. Let $\mathrm{C}(\mathcal{M}; \mathbb{R}^d)$ denote the (separable) Banach space of continuous functions $f : \mathcal{M} \to \mathbb{R}^d$, equipped with the sup-norm $\|f\|_\infty := \sup_{x \in \mathcal{M}} \|f(x)\|_2$, where $\|\cdot\|_2$ is the usual Euclidean norm in $\mathbb{R}^d$. For two random fields $F, H \in \mathrm{C}(\mathcal{M}; \mathbb{R}^d)$, we are interested in the distributional approximation of the random field F by H in appropriate distances, which we introduce next.

For a function $\zeta : \mathrm{C}(\mathcal{M}; \mathbb{R}^d) \to \mathbb{R}$, we denote taking Fréchet derivatives by $D, D^2, \ldots$, and let the operator norm $\|\cdot\|$ be defined for a $k(\geqslant 1)$-linear form A on $\mathrm{C}(\mathcal{M}; \mathbb{R}^d)$ by $\|A\| :=$

$\sup_{\|f\|_\infty=1}|A[f,\dots,f]|$. The (integral) probability distances we consider are given by the supremum of the differences $|\,\mathbb{E}\zeta(F)-\mathbb{E}\zeta(H)|$ taken over all functions in some class $\mathcal{H}$ of *test functions* that map $\mathrm{C}(\mathcal{M};\mathbb{R}^d)\to\mathbb{R}$:

$$\mathsf{d}_{\mathcal{H}}(F,H)\coloneqq\sup_{\zeta\in\mathcal{H}}\big|\mathbb{E}[\zeta(F)]-\mathbb{E}[\zeta(H)]\big|.$$

In particular, we are interested in the case where the role of $\mathcal{H}$ is played by

$$\mathcal{W}\coloneqq\left\{\zeta:\mathrm{C}(\mathcal{M};\mathbb{R}^d)\to\mathbb{R}:\sup_{f\neq h}\frac{|\zeta(f)-\zeta(h)|}{\|f-h\|_\infty}\leqslant 1\right\},$$

the class of 1-Lipschitz functions, in which case the distance is called as the Wasserstein metric (W_1), denoted by $\mathsf{d}_{\mathcal{W}}(F,H)$; convergence in this metric is known to imply weak convergence in (the Polish space) $(\mathrm{C}(\mathcal{M};\mathbb{R}^d),\|\cdot\|_\infty)$; see Dudley (2018, Theorem 11.3.3).

To proceed, we introduce the following weaker metric based on the class of "smooth" test functions

$$\mathcal{F}\coloneqq\big\{\zeta:\mathrm{C}(\mathcal{M};\mathbb{R}^d)\to\mathbb{R}:\ \sup_f\|D^k\zeta(f)\|\leqslant 1,k=1,2;\sup_{f\neq h}\frac{\|D^2\zeta(f)-D^2\zeta(h)\|}{\|f-h\|_\infty}\leqslant 1\big\}. \tag{1.1}$$

The metric $\mathsf{d}_{\mathcal{F}}$ is well-suited to Stein's method, but, in contrast to analogous metrics in the finite dimensional case, it does not directly imply weak convergence, or provide bounds on more informative metrics such as the Wasserstein or Lévy-Prokhorov. Conceptually speaking, this disconnect can occur because it is not established that the test functions in $\mathcal{F}$ capture tightness, and practically speaking, it can occur because the technical tools used in finite dimensions (approximation by smoother functions and boundary measure inequalities) do not generally directly carry over to infinite dimensions. Using our novel Laplacian-based smoothing method, we non-trivially adapt the techniques of Barbour et al. (2024), and prove the following general approximation result in the W_1 metric for random fields indexed by the sphere.

Theorem 1.1. *[Master Theorem] Let $F,H\in\mathrm{C}(\mathcal{S}^n;\mathbb{R}^d)$ be random fields, where $\mathcal{S}^n$ is the unit sphere in $\mathbb{R}^{n+1}$ for some finite integer n. Then for any $\varepsilon,\delta\in(0,1)$ and $\iota>0$,*

$$\mathsf{d}_{\mathcal{W}}(F,H)\leqslant C\Big(d\,\delta^{-2}\varepsilon^{-2(n+\iota)}\underbrace{\boxed{\mathsf{d}_{\mathcal{F}}(F,H)}}_{\text{Section 4}}+\underbrace{\boxed{\mathbb{E}\|F-F_\varepsilon\|_\infty}+\boxed{\mathbb{E}\|H-H_\varepsilon\|_\infty}}_{\text{Section 5}}+\underbrace{\boxed{\delta\sqrt{d}}}_{\text{Section 3}}\Big), \tag{1.2}$$

where F_ε and H_ε are ε-regularizations of F and H defined at (2.9) below, and C is a constant depending only on n and ι.

To explain the terms appearing in the bound, we first give the basic idea behind the proof of Theorem 1.1. Given a function $\zeta:\mathrm{C}(\mathcal{S}^n;\mathbb{R}^d)\to\mathbb{R}$ which is Lipschitz, we define a (ε,δ)-regularized version $\zeta_{\varepsilon,\delta}$ such that for $k=1,2$, $D^k\zeta_{\varepsilon,\delta}(f)$ exists and has norm bounded uniformly in f of order smaller than $\delta^{-2}\varepsilon^{-2(n+\iota)}$, and $D^2\zeta_{\varepsilon,\delta}$ is Lipschitz with respect to the operator norm, with constant of order $\delta^{-2}\varepsilon^{-2(n+\iota)}$. In particular, there is a constant c such that, $c\,\delta^2\varepsilon^{2(n+\iota)}\zeta_{\varepsilon,\delta}\in\mathcal{F}$. Applying the triangle inequality yields

$$\big|\mathbb{E}[\zeta(F)]-\mathbb{E}[\zeta(H)]\big|\leqslant\big|\mathbb{E}[\zeta_{\varepsilon,\delta}(F)]-\mathbb{E}[\zeta_{\varepsilon,\delta}(H)]\big|+\big|\mathbb{E}[\zeta(F)]-\mathbb{E}[\zeta_{\varepsilon,\delta}(F)]\big|+\big|\mathbb{E}[\zeta_{\varepsilon,\delta}(H)]-\mathbb{E}[\zeta(H)]\big|.$$

Because $c\,\delta^2\varepsilon^{2(n+\iota)}\zeta_{\varepsilon,\delta}\in\mathcal{F}$, the first term is bounded of order $\delta^{-2}\varepsilon^{-2(n+\iota)}\times\mathsf{d}_{\mathcal{F}}$. In Theorem 4.1 we bound $\mathsf{d}_{\mathcal{F}}(F,H)$ when H is a continuous and centered $\mathbb{R}^d$ valued Gaussian random field, denoted by $(G(x))_{x\in\mathcal{M}}$, having non-negative definite covariance kernel $C_{ij}(x,y)=\mathbb{E}[G_i(x)G_j(y)]$. This

result follows from a development of Stein's method closely related to that of Barbour et al. (2023), following Barbour (1990).

In contrast to the first term in (1.2), the remaining three terms decay as ε and δ become small, and in particular, the second and third terms become small because ζ and $\zeta_{\varepsilon,\delta}$ become close. The quantity $\|F - F_\varepsilon\|_\infty$ is closely related to the modulus of continuity of F (see Definition 5.1), and hence the term $\mathbb{E}\|F - F_\varepsilon\|_\infty$ can be further bounded using classical quantitative tightness arguments, which we present in Lemma 5.3. The optimal choice of ε and δ is the one having the best tradeoff between the first and the remaining terms, which may in applications depend on the rate of decay of $\mathtt{d}_{\mathcal{F}}(F, H)$ as a function of 'sample' or 'network' size, and which mitigates its prefactor tending to infinity.

While this approach is a standard way to parlay a preliminary bound in a smooth metric into a stronger one, the crux of the problem at the random field level is: *how does one construct* $\zeta_{\varepsilon,\delta}$? In finite dimensions, a fruitful regularization takes a function ζ and replaces it with $\zeta_\delta(x) = \mathbb{E}[\zeta(x+\delta S)]$, where S is a "smoothing" standard Gaussian. The smoothness of ζ_δ follows by making a change of measure and using the smoothness of the Gaussian density. See, for example, Raič (2018) and references therein for additional details.

For random fields *indexed* by $\mathcal{M}$ (or even $\mathcal{S}^n$ with $n \geqslant 2$), there is no "standard" Gaussian and in choosing an appropriate smoothing Gaussian S there are two related potential difficulties. The first is that Cameron-Martin change of measure formulas involve Paley-Wiener integrals, which in general do not have closed form expressions. Moreover, the Cameron-Martin (or Reproducing Kernel Hilbert) space where the change of measure formula holds is typically restricted to a strict subset of $\mathrm{C}(\mathcal{M}; \mathbb{R}^d)$, meaning that $\mathscr{L}(f + \delta S)$ and $\mathscr{L}(\delta S)$ will be singular for many reasonable f. Following the strategy of Barbour et al. (2024), one approach is to define a smoothing Gaussian random field $S : \mathcal{M} \to \mathbb{R}^d$, where the Cameron-Martin space is a subset of smooth functions. In the simpler setting of Barbour et al. (2024) where $\mathcal{M} = [0, T]$, S is taken to be Brownian motion with a random Gaussian initial value, and the Cameron-Martin space is well known to be absolutely continuous functions equipped with L^2-derivative inner product. In our more general setting of random fields indexed by $\mathcal{M}$, there is no canonical Gaussian process like Brownian motion with a well-understood Cameron-Martin space.

In our construction of a smoothing Gaussian random field indexed by $\mathcal{S}^n$, the associated Cameron-Martin space contains a class of functions in the domain of a certain fractional Laplacian and whose images are L^2 bounded, and thus can be equipped with a related L^2 inner product. With this function class in hand, there is still the issue that not all functions $f \in \mathrm{C}(\mathcal{S}^n; \mathbb{R}^d)$ are in the domain of a fractional Laplacian, and so we use a second ε-regularization, now of f, given by $f_\varepsilon(x) = \mathbb{E} f(B_\varepsilon^{(x)})$, where $(B_t^{(x)})_{t \geqslant 0}$ is a Brownian motion on $\mathcal{S}^n$ started from x. Now defining $\zeta_{\varepsilon,\delta}(f) = \mathbb{E}[\zeta(f_\varepsilon + \delta S)]$, bounds on derivatives of $\zeta_{\varepsilon,\delta}$ can be derived from quantitative information on the spectrum of the Laplacian, which is available in detail for $\mathcal{M} = \mathcal{S}^n$. This procedure is elaborated in Section 2.

Although Theorem 4.1 for bounding $\mathtt{d}_{\mathcal{F}}(F, G)$ holds for any compact metric measure space $(\mathcal{M}, \mathtt{d}, \nu)$, specializing to the case of $\mathcal{M} = \mathcal{S}^n$ in Theorem 1.1 allows us to obtain explicit bounds in terms of the problem parameters (i.e., n and d, etc.). The technology of our Laplacian-based smoothing approach applies in more general settings. Explicit bounds can be obtained using our approach anytime there are appropriate estimates for the heat kernel (to prove analogs of Lemma 5.3) and the spectrum of the Laplacian (to construct covariance functions analogous to Section 2.1). For general Riemannian manifolds, such quantitative spectral estimates are well studied; for example, see Grigor'yan (2009), Grieser (2002) and Zelditch (2017). Even more generally, understanding Gaussian random fields and Laplacian operators on general metric measure spaces is an active area;

see, for example, Sturm (1998) and Burago et al. (2019). We highlight that our proofs would also work with functionals of the Laplacian other than fractional powers, as long as they would ensure the required smoothness conditions are satisfied. This flexibility in our proof technique might turn out to be crucial in cases when $\mathcal{M}$ is not the n-sphere.

1.2 Application to wide random neural networks

We now show how Theorem 1.1 is used to obtain quantitative bounds on the distributional approximation of wide random neural networks by appropriately defined Gaussian random fields. Our first motivation to do so is as follows. In practice, widely used training algorithms like stochastic gradient descent are initialized randomly. In light of that, an interesting question was raised by Golikov and Yang (2022): *Does the distribution of the initial weights matter for the training process?* The authors demonstrate that for a large class of distributions of the initial weights, wide random neural networks are Gaussian random fields in the limit. Based on this outcome, they argue that as long as the distribution of the weights are from this universality class, the answer to the above question is *no*. Our results in this section could be used to quantify this phenomenon.

Our second motivation is to initiate the study of the training dynamics of neural networks for prediction problems, at the random field level. Several works (Sirignano and Spiliopoulos, 2020; Chen et al., 2020; Rotskoff and Vanden-Eijnden, 2022) demonstrate that when neural networks are trained by gradient descent with small order step-sizes, certain functionals exhibit limiting Gaussian behavior along the training trajectory. Under larger order step-sizes, the works (Damian et al., 2022; Ba et al., 2022; Abbe et al., 2022) demonstrate that neural networks behave differently than Gaussian-process based prediction methods (including certain classes of kernel methods), thus suggesting the existence of a phase transition from Gaussian to non-Gaussian limits. Our result in this section, along with the associated proof techniques, take a first step towards understanding the above phenomena at the random field level, by developing quantitative information about the setting where the Gaussian behavior is observed.

Formally, we consider a fully connected L-layer neural network that is defined recursively through random fields $F^{(\ell)} : \mathcal{M} \to \mathbb{R}^{n_\ell}$, $\ell = 1, \dots, L$, where $n_1, \dots, n_L$ are positive integers corresponding to the widths of the network, with n_L assumed constant. We also assume that $\mathcal{M} \subset \mathbb{R}^{n_0}$. The random fields are generated by a collection of random matrices $(W^{(\ell)})_{\ell=0}^{L-1}$ where $W^{(\ell)} : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_{\ell+1}}$, with $W^{(0)}$ having i.i.d. rows, $W^{(\ell)}$ having independent entries for $1 \leqslant \ell \leqslant L-1$, and a collection $(b^{(\ell)})_{\ell=0}^{L-1}$ of centered Gaussian "bias" vectors. For $x \in \mathcal{M}$, we define

$$\begin{aligned} F^{(1)}(x) &= W^{(0)}x + b^{(0)}, \\ F^{(\ell)}(x) &= W^{(\ell-1)}\sigma\big(F^{(\ell-1)}(x)\big) + b^{(\ell-1)}, \quad \ell = 2, \dots, L, \end{aligned}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function that we apply to vectors coordinate-wise. We assume that

$$\mathrm{Var}(W_{ij}^{(\ell)}) = \frac{c_w^{(\ell)}}{n_\ell}, \quad \text{and} \quad \mathrm{Var}(b_i^{(\ell)}) = c_b^{(\ell)}.$$

The limiting Gaussian random field is defined inductively as follows. First let $G^{(1)} = F^{(1)}$, which in general is not a Gaussian random field (since $W_{ij}^{(0)}$ is not assumed Gaussian), and has covariance

$$C_{ij}^{(1)}(x, y) = \delta_{ij}\left(\frac{c_w^{(0)}}{n_0}\langle x, y\rangle + c_b^{(0)}\right),$$

where δ_{ij} is the Kronecker delta, and $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner product. Given the distribution of $G^{(\ell)}$ for some $\ell \geqslant 1$, we define $G^{(\ell+1)}$ to be a centered Gaussian random field with covariance

$$C_{ij}^{(\ell+1)}(x,y) = \delta_{ij}\left(c_w^{(\ell)} \, \mathbb{E}\Big[\sigma\big(G_1^{(\ell)}(x)\big)\sigma\big(G_1^{(\ell)}(y)\big)\Big] + c_b^{(\ell)} \right).$$

As the rows of $W^{(\ell)}$ are assumed i.i.d. and the network is fully connected, the components of $F^{(\ell)}, \ell \geqslant 1$ are exchangeable, so in particular identically distributed. Additionally, the covariance functions of $F^{(\ell+1)}$ obey the same recurrence as the one above, with $G_1^{(\ell)}$ replaced by $F_1^{(\ell)}$, and hence have uncorrelated components. Consequently, paralleling the covariance structure of $F^{(\ell)}$, the components of the Gaussian weighted network $G^{(\ell)}$ are, additionally, made independent.

We now state our results for neural networks that have Lipschitz activation function. Widely used activation functions such as the ReLU, sigmoid, softmax, and tanh satisfy this assumption.

Theorem 1.2. *Let $\mathcal{S}^n \subset \mathbb{R}^{n+1} =: \mathbb{R}^{n_0}$ be the n-dimensional sphere, and $G^{(\ell)}, F^{(\ell)} : \mathcal{S}^n \to \mathbb{R}^{n_\ell}$, $\ell = 1, \dots, L$ be defined as above. Assume σ is Lipschitz with constant Lip_σ. If there is a $p > n$ and constants $B^{(\ell)}$, $\ell = 0, \dots, L-1$, independent of $n_1, \dots, n_{L-1}$ such that*

$$\mathbb{E}\big[(W_{ij}^{(\ell)})^{2p}\big] \leqslant \left(\frac{c_w^{(\ell)}}{n_\ell}\right)^p (B^{(\ell)})^{p/2}, \tag{1.3}$$

then for any $\iota > 0$, there is a constant c depending only on $(c_w^{(\ell)}, c_b^{(\ell)}, B^{(\ell)})_{\ell=0}^L, n, p, \sigma(0), \iota$ such that

$$\begin{aligned} &\mathsf{d}_{\mathcal{W}}(F^{(L)}, G^{(L)}) \\ &\quad \leqslant c\,(1+\mathrm{Lip}_\sigma)^{3(L-1)} \sum_{\ell=1}^{L-1} \left(n_{\ell+1}^{1/2} \left(\frac{n_{\ell+1}^4}{n_\ell}\right)^{(1-\frac{n}{p})/(6(1-\frac{n}{p})+8(n+\iota))} \log(n_\ell / n_{\ell+1}^4) \right) \prod_{j=\ell+1}^{L-1} \mathbb{E}\|W^{(j)}\|_{\mathrm{op}}, \end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the matrix operator norm with respect to Euclidean distance.

To the best of our knowledge, Theorem 1.2 provides the first result in the literature for bounding the law of wide random neural networks of any depth and Lipschitz activation functions, to that of a Gaussian. We emphasize in particular that the stated bounds are at the random field level, with the metric being the W_1 metric under the stronger sup-norm topology; see Section 1.2.1 for comparisons to prior works.

Remark 1.3. To understand the bound in Theorem 1.2, first note that in terms of the layer widths, the bound can be small only if

$$n_{\ell+1}^{7+\frac{4(n+\iota)}{1-n/p}} \ll n_\ell, \tag{1.4}$$

so each layer must go to infinity polynomially faster than the next for the bound to go to zero. For the operator norm terms, if $W_{ij}^{(\ell)} = n_\ell^{-1/2} \widetilde{W}_{ij}^{(\ell)}$, with $\widetilde{W}_{ij}^{(\ell)}$ sub-Gaussian, then according to Vershynin (2018, Exercises 4.4.6 and 4.4.7), we have

$$\mathbb{E}\|W^{(\ell)}\|_{\mathrm{op}} = \Theta\big(1 + \sqrt{n_{\ell+1}/n_\ell}\big).$$

Note that under the same assumption, (1.3) is satisfied for all $p \geqslant 2$. Thus, assuming n_ℓ goes to infinity fast enough relative to $n_{\ell+1}$ so that (1.4) is satisfied for large p, the final bound has rate

$$\sum_{\ell=1}^{L-1} n_{\ell+1}^{1/2} \left(\frac{n_{\ell+1}^4}{n_\ell}\right)^{\frac{1}{8n+6}-\varepsilon},$$

for any $\varepsilon > 0$. In the case $n = 1$ and $L = 2$, the rate of $n_1^{-\frac{1}{14}+\varepsilon}$ matches that given in the similar but simpler setting in Barbour et al. (2024, Remark 1.9), which suggests the exponent is linked to our method, and is most certainly not optimal. (One would hope for the central limit rate of $n_1^{-1/2}$, at least up to log factors.)

Regarding the dependence on the widths tending to infinity, it is known from Hanin (2023) that asymptotic convergence to a Gaussian process holds as long as the widths all go to infinity, regardless of relative size. Therefore, our bound's dependence on the scaling of the widths is sub-optimal and is applicable in the so-called *sequential limit* setting (see Section 1.2.1 for more details). The source of this sub-optimality is an artifact of our proof. Specifically, under a sub-Gaussian assumption on the entries of $W^{(\ell)}$, the result in Theorem 1.2 actually provides a bound for the Gaussian approximation for every layer, i.e., for the quantity $\sum_{\ell=1}^{L} \mathsf{d}_{\mathcal{W}}(F^{(\ell)}, G^{(\ell)})$. Indeed, for sufficiently wide neural networks, using results by Vershynin (2018, Exercise 4.4.7) as outlined in Remark 1.3, the bound in Theorem 1.2, with a potentially different constant c having the same dependencies, applies to all layers simultaneously. For such a bound, polynomial dependence on the next layer is likely inevitable, as is the case in the classical CLT in high dimensions.

The moment bounds on the weights are used to control the modulus of continuity terms in (1.2). The operator norm of the weights appear because we are working directly in the sup-norm. The reliance of the bound on these is likely not optimal, and improvements may be achievable using our method with better technical manipulations. We emphasize that the question of the optimal reliance on the widths and moments of the weights is totally open, and our bound is the first and currently the only available bound at the process level that sheds light on the answer.

We next give a high-level idea behind the proof of Theorem 1.2. Note that, conditional on the ℓ-th layer, layer $(\ell + 1)$ is a sum of n_ℓ random fields and a Gaussian:

$$F_i^{(\ell+1)} = \sum_{j=1}^{n_\ell} W_{ij}^{(\ell)} \sigma\big(F_j^{(\ell)}\big) + b_i^{(\ell)}, \quad i = 1, \dots, n_{\ell+1}.$$

Inductively, assuming an appropriate bound on the distributional distance between $F^{(\ell)}$ and $G^{(\ell)}$, we can bound the error made in the approximation

$$F_i^{(\ell+1)} \approx \sum_{j=1}^{n_\ell} W_{ij}^{(\ell)} \sigma\big(G_j^{(\ell)}\big) + b_i^{(\ell)}, \quad i = 1, \dots, n_{\ell+1}.$$

The field on the right-hand side has the same covariance as $G^{(\ell+1)}$, and hence the approximation bound in Theorem 1.2 follows by recursive application of the Stein's method approximation Theorem 4.1 (summarized in Lemma 6.1) to bound $\mathsf{d}_{\mathcal{F}}$, combined with Theorem 1.1.

As detailed in Remark 1.5, our next result shows that one gains an improved dependence on the widths under the assumption that the activation function σ is three-times differentiable, and when smoothing is only performed in the final stage, in contrast to the result of Theorem 1.2, which is obtained by smoothing at each step of the recursion.

Theorem 1.4. *Instantiate the conditions of Theorem 1.2 and assume in addition that the activation function σ has three bounded derivatives. Then, for any $\iota > 0$, there is a constant c depending only on $(c_w^{(\ell)}, c_b^{(\ell)}, B^{(\ell)})_{\ell=0}^{L}$, n, p, ι, and $\|\sigma^{(k)}\|_\infty$, the supremum of the kth derivative of σ, $k = 1, 2, 3$, such that*

$$\mathsf{d}_{\mathcal{W}}(F^{(L)}, G^{(L)}) \leqslant c\sqrt{n_L}(n_L \beta_L^2)^{(1-\frac{n}{p})/(6(1-\frac{n}{p})+8(n+\iota))} \sqrt{\log(1/(n_L\beta_L^2))},$$

where

$$\beta_L := \sum_{\ell=1}^{L-1} \frac{n_{\ell+1}^{3/2}}{\sqrt{n_\ell}} \prod_{j=\ell+1}^{L-1} \max\{1, \mathbb{E}[\|W^{(j)}\|_{\mathrm{op}}^3]\}. \tag{1.5}$$

Remark 1.5. Under the same setting as in Remark 1.3, with C_L a constant depending only on L, we have

$$\beta_L^2 \leqslant C_L \sum_{\ell=1}^{L-1} \frac{n_{\ell+1}^3}{n_\ell} \quad \text{and hence} \quad \mathsf{d}_{\mathcal{W}}(F^{(L)}, G^{(L)}) \leqslant \sqrt{n_L} \left(n_L \sum_{\ell=1}^{L-1} \frac{n_{\ell+1}^3}{n_\ell} \right)^{\frac{1}{(8n+6)}-\varepsilon},$$

for any $\varepsilon > 0$, which will tend to zero as long as $n_{\ell+1}^3 \ll n_\ell$, demonstrating the improvement on the width dependence obtained by Theorem 1.4 (specifically, in comparison to (1.4)).

1.2.1. Comparison to related works. Eldan et al. (2021) studied Gaussian random field approximation bounds for the case of $L = 2$ with Gaussian weights with three specific choices of activation functions. They used the Wasserstein-2 distance with respect to L^p topology on the sphere. For polynomial activations they work with $p = \infty$, and for ReLU and tanh they work with $p < \infty$. Following that work, Klukowski (2022) derived improved bounds in the Wasserstein-2 distance with respect to L^2 topology, assuming the rows of the weight matrix are drawn uniformly from the sphere. We remark that weak convergence with respect to integral norms (such as L^p with $p < \infty$) does not imply weak convergence of finite dimensions, or of other natural statistics such as the maximum.

Basteri and Trevisan (2022) gives rate of convergence of finite dimensional distributions for general depth fully connected networks and Gaussian weights. The metric is Wasserstein-2 with respect to Euclidean norm. The bound of Basteri and Trevisan (2022) exhibits multivariate convergence as long as n_ℓ tends to infinity for each $\ell = 1, \ldots, L-1$, in any order. This phenomenon is a consequence of a very good relationship between the dimension and the number of terms for the rate of convergence in the multivariate CLT, stemming from the metric used there, and the Gaussian assumptions on the weights.

Bordino et al. (2023b) used Stein's method to derive bounds for univariate distributional approximation for one-layer neural networks with Gaussian weights in the W_1, Kolmogorov and total variation metrics, and bounds for the error of the approximation of a multivariate output of the network by a Gaussian, in the W_1 metric. Their approach is based on a straight-forward but laborious application of a Gaussian approximation result for functions of Gaussian random variables in Vidotto (2020), which is a multivariate refinement of the second-order Poincaré inequality version of Stein's method introduced by Chatterjee (2009).

More recently, Favaro et al. (2023b, Theorem 3.13), in a paper posted to arXiv roughly two weeks after the posting of our draft, prove a rate of convergence in the sup-norm under the assumptions that (i) the weights W_{ij} are Gaussian, (ii) the activation function σ is infinitely smooth with polynomially bounded derivatives, and (iii) the width of the layers all tend to infinity at the same rate. The setting in (iii) is called as the *simultaneous limit* setting in the literature, whereas the setting for our results in Theorems 1.2 and 1.4 is known as the *sequential limit* setting; see Lee et al. (2018); de G. Matthews et al. (2018); Bahri and Hanin (2023) for the applications and differences between the two settings. While our bound is not informative under the *simultaneous limit* setting (iii), it becomes informative under the *sequential limit* setting, and applies when assuming significantly much less than (i) and (ii). In particular, condition (ii) renders Favaro et al. (2023b, Theorem 3.13) inapplicable for the widely used ReLU activation functions, in contrast to our results. The proof of Theorem 3.13 of Favaro et al. (2023b) follows from the observation (also used in Basteri and Trevisan

(2022)) that due to the weights being Gaussian, the distribution of one layer conditional on the previous one is a Gaussian field, but with a conditional covariance. In transport metrics with respect to Sobolev topologies, the distance between two Gaussians is determined by a distance between their covariances, which can be controlled using results of Hanin (2022). While this approach leads to a remarkable result for the Gaussian weights with infinitely smooth activation functions, the method of proof does not appear to generalize beyond this specific setting, thereby providing compelling motivation for our alternative general approach.

1.2.2. Future directions. Obtaining a deeper understanding of the weak convergence of wide random neural networks to Gaussian (and non-Gaussian) random fields is an active area of research. Here, we highlight a few interesting directions which can be pursued based on our work.

Rate improvements: There are at least two directions to explore for improving the bounds of Theorem 1.2 and Theorem 1.4. The first, is in the case of Gaussian weights, is to understand whether the proof approach in Basteri and Trevisan (2022) for the multivariate setting could be extended to the random-field level. The second is to develop improved rates (in potentially weaker topologies, but still at the random field level) by combining our techniques with those in Hanin (2023). Both directions are intriguing but appear to be non-trivial at the random field level, and we leave them as future work to investigate.

Heavy-tailed weights: Motivated by constructing priors for Bayesian inference for neural networks, Neal (1996, Section 2.2) also heuristically examined the limits of single layer neural networks with the entries of the weight matrices being stable random variables. Recently, several works (Der and Lee, 2005; Jung et al., 2023; Bordino et al., 2023a; Lee et al., 2023; Fortuin et al., 2022; Favaro et al., 2023a; Bordino et al., 2023a) showed that the limits of such neural networks (including deep ones) converge weakly to appropriately defined stable random fields. An interesting question that arises is whether one can establish quantitative distributional approximation bounds in the heavy-tailed setting. Our work provides a step in this direction. Indeed, our main result in Theorem 1.1 is immediately applicable. The remaining challenge will be in establishing a version of Theorem 4.1 for stable random fields. This could potentially be accomplished by extending recent works, for example, Xu (2019); Arras and Houdré (2019, 2022); Chen et al. (2023), on multivariate stable approximations to the random field setting.

The rest of the article is organized as follows. Section 2 defines and develops properties of our smoothing Gaussian process, which are then used in Section 3 to prove our general smoothing result, Theorem 1.1. Section 4 develops Stein's method for Gaussian processes, culminating in Theorem 4.1, which is used to bound $\mathtt{d}_{\mathcal{F}}$. Section 5 uses classical quantitative chaining arguments along with heat kernel bounds to prove Lemma 5.3, which gives an easily applied method for bounding $\mathbb{E}\|F - F_\varepsilon\|_\infty$. Finally, Section 6 uses the theory developed in the previous sections to prove our wide neural network approximation results, Theorems 1.2 and 1.4.

Acknowledgments. We thank Volker Schlue for discussions regarding certain differential geometric aspects and Max Fathi for the suggestion to look at the Laplacian for smoothing. This project originated at the "Stein's Method: The Golden Anniversary" workshop organized by the Institute for Mathematical Sciences at the National University of Singapore in June–July, 2022. We thank the institute for the hospitality and the organizers for putting together the stimulating workshop. KB was supported in part by National Science Foundation (NSF) grant DMS-2053918.

2 GAUSSIAN SMOOTHING FOR RANDOM FIELDS INDEXED BY THE SPHERE

We begin by constructing our Gaussian smoothing random field, with its covariance defined based on the powers of Laplacian operators, and specifying its Cameron-Martin space.

2.1 Constructing a covariance and Cameron-Martin space from the Laplacian

To define our smoothing Gaussian random field, we construct a covariance function based on the Laplacian on the n-sphere $\mathcal{S}^n$, which we view as embedded in $\mathbb{R}^{n+1}$,

$$\mathcal{S}^n = \{x \in \mathbb{R}^{n+1} : \|x\|_2 = 1\}.$$

A standard way to define the Laplacian on the sphere is to "lift" functions $f : \mathcal{S}^n \to \mathbb{R}$ to $\tilde{f} : \mathbb{R}^{n+1} \setminus \{0\} \to \mathbb{R}$ by

$$\tilde{f}(x) = f(x/\|x\|_2).$$

Letting $\widetilde{\Delta}$ denote the usual Laplacian on $\mathbb{R}^{n+1}$, we can then define the Laplacian Δ acting on twice differential functions on $\mathcal{S}^n$ by

$$\Delta f(x) = \widetilde{\Delta}\tilde{f}(x), \ x \in \mathcal{S}^n;$$

see, for example, Dai and Xu (2013, Corollary 1.4.3). The negative of the Laplacian $(-\Delta)$ is a positive definite operator on $L^2(\mathcal{S}^n; \mathbb{R})$ and has an orthonormal basis given by *spherical harmonics*. The eigenvalues of $(-\Delta)$ are

$$\lambda_k = k(k+n-1), \ \ k = 0, 1, 2, \ldots, \tag{2.1}$$

and an orthonormal basis for the eigenspace associated to λ_k is given by a collection of polynomials $\mathscr{H}_k = \{\varphi_k^{(1)}, \ldots, \varphi_k^{(d_k)}\}$, with

$$d_k := \dim \mathscr{H}_k = \frac{2k+n-1}{k}\binom{n+k-2}{k-1}; \tag{2.2}$$

see, for example, Dai and Xu (2013, Corollary 1.1.4). The union $\bigcup_{k\geqslant 0} \mathscr{H}_k$, of the sets of all basis vectors for the k^{th} eigenspace, gives an orthonormal basis for $L^2(\mathcal{S}^n; \mathbb{R})$. From here, we define the *zonal harmonics*

$$Z_k(x, y) = \sum_{j=1}^{d_k} \varphi_k^{(j)}(x)\varphi_k^{(j)}(y), \tag{2.3}$$

that for $n \geqslant 2$ satisfy

$$Z_k(x, y) = \frac{\Gamma\big((n+1)/2\big)(2k+n-1)}{2\pi^{(n+1)/2}(n-1)} C_k^{(n-1)/2}\big(\langle x, y\rangle\big), \tag{2.4}$$

where $C_k^\lambda, \lambda > 0, k \geqslant -1$ are the *Gegenbauer polynomials* defined by the three term recurrence, for $x \in [-1, 1]$,

$$C_{k+1}^\lambda(x) = \frac{2(k+\lambda)}{k+1} x C_k^\lambda(x) - \frac{k+2\lambda-1}{k+1} C_{k-1}^\lambda(x) \quad \text{for } k \geqslant 1,$$

with initial values $C_{-1}^\lambda \equiv 0$ and $C_0^\lambda \equiv 1$. For $n = 1$, $Z_k(x, y) = \pi^{-1}\cos\big(k(\theta_x - \theta_y)\big)$, where θ_x, θ_y are the polar angles of x, y, i.e., $x = (\cos(\theta_x), \sin(\theta_x))$. For our purposes, the key property of Z_k is that

$$|Z_k(x, y)| \leqslant Z_k(x, x) = \frac{\Gamma\big((n+1)/2\big)}{2\pi^{(n+1)/2}} d_k, \tag{2.5}$$

see Dai and Xu (2013, Corollary 1.2.7), noting their different normalization at (1.1.1) of the inner product on the sphere. Thus, for any $\iota > 0$, we can define the kernel $C^{(\iota)} = (C_{ij}^{(\iota)})_{i,j=1}^d$ on $\mathcal{S}^n$ by

$$C_{ij}^{(\iota)}(x, y) = \delta_{ij} \sum_{k\geqslant 1} \frac{Z_k(x, y)}{\lambda_k^{n_\iota}}, \tag{2.6}$$

where $n_\iota := (n+\iota)/2$. Because[1] $\lambda_k \asymp k^2$ and $d_k \asymp k^{n-1}$, by (2.5) we see that $|Z_k(x,y)|$ is $\mathrm{O}(k^{n-1})$ uniformly, hence the sum (2.6) converges absolutely and uniformly. Since each Z_k is continuous and the sphere is compact, $C^{(\iota)}$ is continuous and positive definite due to the decomposition (2.3). We fix $n \geqslant 2$ and $\iota > 0$, and set $C = C^{(\iota)}$ for the remaining part of this section. With our covariance kernel in hand, we define our smoothing random field S and its Cameron-Martin space.

Definition 2.1 (The Smoothing Gaussian random field S and its Cameron-Martin space)**.** Let S be the centered $\mathbb{R}^d$-valued Gaussian random field indexed by $\mathcal{S}^n$ with covariance function C given by (2.6). Let $\mathbf{e}_i \in \mathbb{R}^d$, for $i = 1 \ldots, d$ be the standard basis vectors for $\mathbb{R}^d$. The associated orthonormal decomposition for an $h \in L^2(\mathcal{S}^n; \mathbb{R}^d)$ is given by

$$h = \sum_{i=1}^{d} \mathbf{e}_i \sum_{k \geqslant 1} \sum_{j=1}^{d_k} h_{k,i}^{(j)} \varphi_k^{(j)} \quad \text{where} \quad h_{k,i}^{(j)} = \int_{\mathcal{S}^n} h_i(x) \varphi_k^{(j)}(x) \mathrm{d}x, \tag{2.7}$$

where $\mathrm{d}x$ is the volume measure on the sphere. We define the *Cameron-Martin* or *Reproducing Kernel Hilbert* space H of S to be the subset of $L^2(\mathcal{S}^n; \mathbb{R}^d)$ defined by

$$H = \left\{ h \in L^2(\mathcal{S}^n; \mathbb{R}^d) : \sum_{k \geqslant 1} \lambda_k^{n_\iota} \sum_{i=1}^{d} \sum_{j=1}^{d_k} (h_{k,i}^{(j)})^2 < \infty \right\}, \tag{2.8}$$

equipped with inner product

$$\langle h, g \rangle_H := \sum_{k \geqslant 1} \lambda_k^{n_\iota} \sum_{i=1}^{d} \sum_{j=1}^{d_k} h_{k,i}^{(j)} g_{k,i}^{(j)}.$$

There is the following alternative description of the Cameron-Martin space and inner product. We define the fractional Laplacian operator $(-\Delta)^\alpha$ for any $\alpha > 0$ through the orthonormal basis $(-\Delta)^\alpha \varphi_k^{(j)} := \lambda_k^\alpha \varphi_k^{(j)}$, and for $h : \mathcal{S}^n \to \mathbb{R}^d$ we write $(-\Delta)^\alpha h$ for the fractional-Laplacian applied coordinate-wise.

Proposition 2.2. *If $h, g \in L^2(\mathcal{S}^n; \mathbb{R}^d)$ are such that $(-\Delta)^{\frac{1}{2}n_\iota} h, (-\Delta)^{\frac{1}{2}n_\iota} g \in L^2(\mathcal{S}^n; \mathbb{R}^d)$, then $h, g \in H$ and*

$$\langle h, g \rangle_H = \left\langle (-\Delta)^{\frac{1}{2}n_\iota} h, (-\Delta)^{\frac{1}{2}n_\iota} g \right\rangle_{L^2(\mathcal{S}^n; \mathbb{R}^d)}.$$

Proof. First note that

$$\left\langle (-\Delta)^{\frac{1}{2}n_\iota} h, (-\Delta)^{\frac{1}{2}n_\iota} g \right\rangle_{L^2(\mathcal{S}^n; \mathbb{R}^d)} = \sum_{i=1}^{d} \int_{\mathcal{S}^n} (-\Delta)^{\frac{1}{2}n_\iota} h_i(x) (-\Delta)^{\frac{1}{2}n_\iota} g_i(x) \, \mathrm{d}x.$$

Thus, by additivity, it suffices to show the result for $d = 1$. Since $(-\Delta)^{\frac{1}{2}n_\iota} h \in L^2(\mathcal{S}^n)$, we can compute the coefficients in its $L^2(\mathcal{S}^n)$ expansion (2.7) as

$$\begin{aligned} \int (-\Delta)^{\frac{1}{2}n_\iota} h(x) \varphi_k^{(j)}(x) \mathrm{d}x &= \sum_{\ell \geqslant 1} \lambda_\ell^{\frac{1}{2}n_\iota} \sum_{i=1}^{d_k} h_\ell^{(i)} \int \varphi_\ell^{(i)}(x) \varphi_k^{(j)}(x) \mathrm{d}x \\ &= \lambda_k^{\frac{1}{2}n_\iota} h_k^{(j)}, \end{aligned}$$

[1] For two functions f, g, by $f \asymp g$ means that there exists absolute constants $c, C > 0$ such that $c|g| \leqslant |f| \leqslant C|g|$.

where the second equality follows from orthonormality. Thus, we have that

$$\big\langle (-\Delta)^{\frac{1}{2}n_\iota} h, (-\Delta)^{\frac{1}{2}n_\iota} g \big\rangle_{L^2(\mathcal{S}^n;\mathbb{R}^d)} = \sum_{k\geqslant 1} \lambda_k^{n_\iota} \sum_{j=1}^{d_k} h_k^{(j)} g_k^{(j)} = \langle h, g\rangle_H.$$

□

To explicitly state the Cameron-Martin change of measure formula for S, we first provide its Karhunen-Loeve expansion; see Adler and Taylor (2007, Chapter 3).

Theorem 2.3 (Karhunen-Loeve Expansion for the Smoothing Gaussian random field S)**.** *There exists* $(\mathcal{Z}_{k,i}^{(j)} : k \geqslant 1, 1 \leqslant j \leqslant d_k, 1 \leqslant i \leqslant d)$ *independent centered normal random variables with* $\mathrm{Var}(\mathcal{Z}_{k,i}^{(j)}) = \lambda_k^{-n_\iota}$ *such that*

$$S_i = \sum_{k\geqslant 1} \sum_{j=1}^{d_k} \mathcal{Z}_{k,i}^{(j)} \varphi_k^{(j)},$$

where the convergence holds in L^2 *and almost surely, uniformly on* $\mathcal{S}^n$*.*

With this result, we have the following natural definition.

Definition 2.4 (Paley-Wiener integral)**.** For $h \in H$ with L^2 expansion (2.7), the Paley-Wiener integral with respect to S is the centered normal random variable with variance $\langle h, h\rangle_H$ given by

$$\langle h, S\rangle_H := \sum_{i=1}^{d} \sum_{k\geqslant 1} \lambda_k^{n_\iota} \sum_{j=1}^{d_k} \mathcal{Z}_{k,i}^{(j)} h_{k,i}^{(j)},$$

where the $\mathcal{Z}_{k,i}^{(j)}$ are as in Theorem 2.3.

We can now formally state the Cameron-Martin change of measure formula for S, which follows from an application of a theorem of Kakutani (1948) for absolute continuity of infinite product measure.

Theorem 2.5 (Cameron-Martin change of measure for the Smoothing Gaussian random field S)**.** *For any* $h \in H$*,* $\mathscr{L}(h + S)$ *has Radon-Nikodym derivative with respect to* $\mathscr{L}(S)$*, given by*

$$\frac{d\mathscr{L}(h+S)}{d\mathscr{L}(S)} = \exp\big\{\langle h, S\rangle_H - \tfrac{1}{2}\langle h, h\rangle_H\big\}.$$

2.2 Regularization to the Cameron-Martin space

In the previous section, we provided the Cameron-Martin change of measure formula for our smoothing Gaussian random field S, but it only applies to functions $f \in \mathrm{C}(\mathcal{S}^n;\mathbb{R}^d)$ that are in the Cameron-Martin space H defined at (2.8), or, according to Proposition 2.2, that are sufficiently smooth. Thus, we define the ε-regularization of f by

$$f_\varepsilon(x) = \big(f_{\varepsilon,i}(x)\big)_{i=1}^d = \big(e^{\varepsilon\frac{\Delta}{2}} f_i(x)\big)_{i=1}^d = \sum_{i=1}^{d} \mathbf{e}_i \sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} \sum_{j=1}^{d_k} f_{k,i}^{(j)} \varphi_k^{(j)}(x). \tag{2.9}$$

The ε-regularized $f_\varepsilon(x)$ equals $\mathbb{E}\big[f(B_\varepsilon^{(x)})\big]$, where $(B_t^{(x)})_{t\geqslant 0}$ is a d-dimensional Brownian motion run on the sphere started from x; see Bakry et al. (2014). The next proposition uses this representation of f_ε in terms of the "heat kernel" for Brownian motion, which will be useful to derive smoothness properties.

Proposition 2.6. *Let*

$$p(x,y;\varepsilon) = \sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} \sum_{j=1}^{d_k} \varphi_k^{(j)}(x)\varphi_k^{(j)}(y) = \sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} Z_k(x,y), \tag{2.10}$$

using the definition of Z_k in (2.3). Then for any bounded and measurable $f : \mathcal{S}^n \to \mathbb{R}^d$,

$$f_{\varepsilon,i}(x) = \int_{\mathcal{S}^n} p(x,y;\varepsilon) f_i(y)\mathrm{d}y. \tag{2.11}$$

Proof. Dropping the subscript i, we have by (2.10),(2.5), (2.2), and Fubini's theorem that

$$\begin{aligned}\int_{\mathcal{S}^n} p(x,y;\varepsilon) f(y)\mathrm{d}y &= \sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} \int_{\mathcal{S}^n} f(y) Z_k(x,y)\mathrm{d}y \\ &= \sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} \sum_{j=1}^{d_k} f_k^{(j)} \varphi_k^{(j)}(x),\end{aligned}$$

which is the same as (2.9). □

We are now in position to derive bounds on $\|(-\Delta)^\alpha f_\varepsilon\|$.

Proposition 2.7. *If $f : \mathcal{S}^n \to \mathbb{R}^d$ is bounded and measurable, then for any $\alpha > 0$, $(-\Delta)^\alpha f_\varepsilon$ exists and $(-\Delta)^\alpha f_\varepsilon \in L^2(\mathcal{S}^n;\mathbb{R}^d)$. Moreover, there is a constant $c = c(n,\alpha)$ depending only on n and α such that*

$$\|(-\Delta)^\alpha f_{\varepsilon,i}\|_\infty \leqslant c\,\|f_i\|_\infty \varepsilon^{-(2\alpha+n)/2}.$$

Proof. By (2.4) each (lifted) $Z_k(\cdot,y)$ is infinitely differentiable, with derivatives growing in absolute value at most polynomially in k. Thus, using (2.10), that $\lambda_k \asymp k^2$, and dominated convergence, $(-\Delta_x)^\alpha p(x,y;\varepsilon)$ is well-defined and

$$(-\Delta_x)^\alpha p(x,y;\varepsilon) = \sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} (-\Delta_x)^\alpha Z_k(x,y).$$

Now, dropping the i subscript, using (2.10) and (2.3), followed by (2.5), (2.2) to find $d_k = \mathrm{O}(k^{n-1})$ and (2.1), so as to apply dominated convergence, we have

$$\begin{aligned}\big|(-\Delta)^\alpha f_\varepsilon(x)\big| &= \left|\int_{\mathcal{S}^n} (-\Delta_x)^\alpha p(x,y;\varepsilon) f(y)\mathrm{d}y\right| \\ &\leqslant \|f\|_\infty \int_{\mathcal{S}^n} \big|(-\Delta_x)^\alpha p(x,y;\varepsilon)\big|\mathrm{d}y \\ &= \|f\|_\infty \int_{\mathcal{S}^n} \left|\sum_{k\geqslant 1} e^{-\frac{\varepsilon\lambda_k}{2}} (-\Delta_x)^\alpha Z_k(x,y)\right|\mathrm{d}y \\ &= \|f\|_\infty \int_{\mathcal{S}^n} \left|\sum_{k\geqslant 1} \lambda_k^\alpha e^{-\frac{\varepsilon\lambda_k}{2}} Z_k(x,y)\right|\mathrm{d}y \\ &\leqslant \|f\|_\infty \sum_{k\geqslant 1} \lambda_k^\alpha e^{-\frac{\varepsilon\lambda_k}{2}} d_k \\ &\leqslant c\|f\|_\infty \sum_{k\geqslant 1} k^{2\alpha+n-1} e^{-\frac{\varepsilon k^2}{2}}.\end{aligned}$$

By comparing this sum with

$$\int_0^\infty (\varepsilon/2)^{(2\alpha+n)/2} x^{2\alpha+n-1} e^{-\varepsilon x^2/2} \mathrm{d}x = \tfrac{1}{2}\Gamma(\alpha + n/2),$$

we find

$$\left|(-\Delta)^\alpha f_\varepsilon(x)\right| \leqslant c\, \varepsilon^{-(2\alpha+n)/2} \|f\|_\infty,$$

where c is a constant depending only on n and α, as desired. □

Propositions 2.2 and 2.7 imply that $f_\varepsilon \in H$ for bounded and measurable f, and also give the following lemma bounding $|\langle f_\varepsilon, g_\varepsilon\rangle_H|$, whose proof is straightforward.

Lemma 2.8. *If f, g are bounded and measurable functions $\mathcal{S}^n \to \mathbb{R}^d$, then there is a constant $c = c(n, \iota)$ depending only on n and ι such that*

$$\left|\langle f_\varepsilon, g_\varepsilon\rangle_H\right| = \left|\left\langle (-\Delta)^{\frac{1}{2}n_\iota} f_\varepsilon, (-\Delta)^{\frac{1}{2}n_\iota} g_\varepsilon \right\rangle_{L^2(\mathcal{S}^n;\mathbb{R}^d)}\right| \leqslant c\, d \|f\|_\infty \|g\|_\infty \varepsilon^{-(2n+\iota)}.$$

2.3 Smoothing using S and regularization

We now use the f_ε regularization given in the last section to define a (ε, δ)-regularized version of a test function ζ. The following result is an analog of Barbour et al. (2024, Lemma 1.10).

Theorem 2.9. *Let $\zeta : \mathrm{C}(\mathcal{S}^n; \mathbb{R}^d) \to \mathbb{R}$ and, for $f : \mathcal{S}^n \to \mathbb{R}^d$ bounded and measurable, define*

$$\zeta_{\varepsilon,\delta}(f) := \mathbb{E}[\zeta(f_\varepsilon + \delta S)],$$

where f_ε is the ε-regularization defined at (2.9). If ζ is bounded or Lipschitz, then $\zeta_{\varepsilon,\delta}$ is infinitely differentiable. Moreover, for every $k \geqslant 0$ there is a constant c depending only on k, n and ι, such that if ζ is bounded, then

$$\|D^k \zeta_{\varepsilon,\delta}\| \leqslant c\, d^{k/2} \delta^{-k} \varepsilon^{-k(n+\iota)} \|\zeta\|_\infty,$$

and if ζ is 1-Lipschitz and $h : \mathcal{S}^n \to \mathbb{R}^d$ is bounded and measurable, then

$$\|D^k \zeta_{\varepsilon,\delta}(f) - D^k \zeta_{\varepsilon,\delta}(h)\| \leqslant c\, d^{k/2} \delta^{-k} \varepsilon^{-k(n+\iota)} \|f - h\|_\infty. \tag{2.12}$$

Proof. The proof is closely related to that of Barbour et al. (2024, Lemma 1.10), where the Cameron-Martin inner product and ε-regularization are simpler. Intuition behind the manipulations below can be found there.

Firstly, $\zeta_{\varepsilon,\delta}$ is clearly well-defined if ζ is bounded. If ζ is C-Lipschitz, then

$$\left|\zeta_{\varepsilon,\delta}(f) - \zeta(f_\varepsilon)\right| \leqslant C\delta\, \mathbb{E}\|S\|_\infty < \infty,$$

where the last inequality is Fernique's theorem (Fernique, 1970). Moreover, $\zeta_{\varepsilon,\delta}$ is measurable since, from (2.11), $f \mapsto f_\varepsilon$ is continuous with respect to sup-norm, as is $(f, g) \mapsto f + g$ in product topology. Thus, $(f, s) \mapsto \zeta(f_\varepsilon + \delta s)$ is measurable with respect to product topology.

We claim that for ζ bounded, $k \geqslant 1$ and $g^{(i)} \in \mathrm{C}(\mathcal{S}^n; \mathbb{R}^d)$, $i = 1, \ldots, k$, we have

$$D^k \zeta_{\varepsilon,\delta}(f)[g^{(1)}, \ldots, g^{(k)}] = \mathbb{E}\left[\zeta(\delta S) e^{\Psi_\varepsilon(f)} \sum_{\pi \in \mathcal{P}_{k,2}} \prod_{b \in \pi} D^{|b|} \Psi_\varepsilon(f)[g^{(b)}]\right], \tag{2.13}$$

where

$$\Psi_\varepsilon(f) = \tfrac{1}{\delta}\langle f_\varepsilon, S\rangle_H - \tfrac{1}{2\delta^2}\langle f_\varepsilon, f_\varepsilon\rangle_H.$$

In (2.13) $\mathcal{P}_{k,2}$ is the set of all partitions of $\{1,\dots,k\}$, whose blocks have at most 2 elements; $b \in \pi$ means that b is a block of π, and we denote its cardinality by $|b|$. When $b = \{i\}$ the expression $D^{|b|}\Psi_\varepsilon(f)[g^{(b)}]$ is defined as

$$D^{|b|}\Psi_\varepsilon(f)[g^{(b)}] = D\Psi_\varepsilon(f)[g^{(i)}] = \delta^{-1}\big\langle g_\varepsilon^{(i)}, S - \delta^{-1} f_\varepsilon\big\rangle_H,$$

and when $|b| = 2$ is given by $b = \{i_1, i_2\}$, then

$$D^{|b|}\Psi_\varepsilon(f)[g^{(b)}] = D^2\Psi_\varepsilon(f)[g^{(i_1)}, g^{(i_2)}] = -\delta^{-2}\langle g_\varepsilon^{(i_1)}, g_\varepsilon^{(i_2)}\rangle_H,$$

which we note does not depend on f. Compare to Barbour et al. (2024, Equation (2.11)) with $\Theta \equiv 0$. Assuming (2.13), the Cameron-Martin Theorem 2.5 implies that

$$D^k\zeta_{\varepsilon,\delta}(f)[g^{(1)},\dots,g^{(k)}] = \mathbb{E}\bigg[\zeta(f_\varepsilon + \delta S)\sum_{\pi\in\mathcal{P}_{k,2}}\prod_{b\in\pi}\widehat{D}^{|b|}\Psi_\varepsilon(f)[g^{(b)}]\bigg], \tag{2.14}$$

where $\widehat{D}^2\Psi_\varepsilon(f) = D^2\Psi_\varepsilon(f)$, and

$$\widehat{D}\Psi_\varepsilon(f)[g] = \delta^{-1}\big\langle g_\varepsilon^{(i)}, S\big\rangle_H \sim \mathrm{Normal}\big(0, \delta^{-2}\langle g_\varepsilon^{(i)}, g_\varepsilon^{(i)}\rangle_H\big),$$

and we note that $\widehat{D}^{|b|}\Psi_\varepsilon(f)$ does not depend on f for $|b| \in \{1,2\}$. Technically, we are applying the Cameron-Martin change of measure formula to the joint distribution of the random variables $\big(\langle g_\varepsilon^{(i)}, S - \delta^{-1}f_\varepsilon\rangle_H\big)_{i=1}^k$ and $(S - \delta^{-1}f_\varepsilon)$, which follows in a straightforward way from Kakutani's theorem and the definition of the Paley-Wiener integral.

By Lemma 2.8, we have

$$\big|\langle g_\varepsilon^{(i_1)}, g_\varepsilon^{(i_2)}\rangle_H\big| \leqslant c\,d\|g^{(i_1)}\|_\infty\|g^{(i_2)}\|_\infty\varepsilon^{-(2n+\iota)}, \tag{2.15}$$

where c is a constant depending only on n_ι. Thus, if ζ is bounded, we have

$$\big|D^k\zeta_{\varepsilon,\delta}(f)[g^{(1)},\dots,g^{(k)}]\big| \leqslant \|\zeta\|_\infty\sum_{\pi\in\mathcal{P}_{k,2}}\prod_{\substack{b\in\pi\\|b|=2}}\big|\widehat{D}^2\Psi_\varepsilon(f)[g^{(b)}]\big|\,\mathbb{E}\prod_{\substack{b\in\pi\\|b|=1}}\big|\widehat{D}\Psi_\varepsilon(f)[g^{(b)}]\big|,$$

and then the definition of $\widehat{D}^k\Psi_\varepsilon$, (2.15), and Hölder's inequality imply

$$\big|D^k\zeta_{\varepsilon,\delta}(f)[g^{(1)},\dots,g^{(k)}]\big| \leqslant c\,d^{k/2}\delta^{-k}\varepsilon^{-k(n+\iota)}\|\zeta\|_\infty\prod_{i=1}^k\|g^{(i)}\|_\infty,$$

where c depends on k (through the sum over $\mathcal{P}_{k,2}$ and the absolute moments up to order k of standard normal variables) and n_ι, as desired.

Assume now ζ is 1-Lipshitz, and letting $f, h \in \mathrm{C}(\mathcal{S}^n;\mathbb{R}^d)$ and recalling that $\widehat{D}^{|b|}\Psi_\varepsilon(f)[g^{(b)}] = \widehat{D}^{|b|}\Psi_\varepsilon(h)[g^{(b)}]$, (2.14) implies

$$\begin{aligned}D^k\zeta_{\varepsilon,\delta}(f)[g^{(1)},\dots,g^{(k)}] - D^k\zeta_{\varepsilon,\delta}(h)[g^{(1)},\dots,g^{(k)}]&\\ = \mathbb{E}\bigg[\big(\zeta(f_\varepsilon + \delta S) - \zeta(h_\varepsilon + \delta S)\big)\sum_{\pi\in\mathcal{P}_{k,2}}\prod_{b\in\pi}\widehat{D}^{|b|}\Psi_\varepsilon(f)[g^{(b)}]\bigg],&\end{aligned}$$

and using that ζ is Lipschitz and (2.11), we have

$$\big|\zeta(f_\varepsilon + \delta S) - \zeta(h_\varepsilon + \delta S)\big| \leqslant \|f_\varepsilon - h_\varepsilon\|_\infty \leqslant \|f - h\|_\infty.$$

With this, (2.12) follows in exactly the same way as the bounded case.

To establish (2.13), we use induction. For $k = 1$, the Cameron-Martin Theorem 2.5 implies

$$\zeta_{\varepsilon,\delta}(f+g) - \zeta_{\varepsilon,\delta}(f) = \mathbb{E}\big[\zeta(\delta S)(e^{\Psi_\varepsilon(f+g)} - e^{\Psi_\varepsilon(f)})\big],$$

so by the bounded- or Lipschitz-ness of ζ and the Cauchy-Schwarz inequality, it is enough to show that

$$\mathbb{E}\Big[\big(e^{\Psi_\varepsilon(f+g)-\Psi_\varepsilon(f)} - 1 - D\Psi_\varepsilon(f)[g]\big)^2\Big] = \mathrm{o}\big(\|g\|_\infty^2\big). \tag{2.16}$$

But

$$\Psi_\varepsilon(f+g) - \Psi_\varepsilon(f) = D\Psi_\varepsilon(f)[g] - \frac{1}{2\delta^2}\langle g_\varepsilon, g_\varepsilon\rangle_H,$$

with $D\Psi_\varepsilon(f)[g] \sim \mathrm{Normal}(-\delta^{-2}\langle f_\varepsilon, g_\varepsilon\rangle_H, \delta^{-2}\langle g_\varepsilon, g_\varepsilon\rangle_H)$, and so straightforward computing shows

$$\begin{aligned}
&\mathbb{E}\Big[\big(e^{\Psi_\varepsilon(f+g)-\Psi_\varepsilon(f)} - 1 - D\Psi_\varepsilon(f)[g]\big)^2\Big] \\
&= e^{\delta^{-2}\langle g_\varepsilon, g_\varepsilon\rangle_H - 2\delta^{-2}\langle f_\varepsilon, g_\varepsilon\rangle_H} + \left(1 - \frac{\langle f_\varepsilon, g_\varepsilon\rangle_H}{\delta^2}\right)^2 + \frac{\langle g_\varepsilon, g_\varepsilon\rangle_H}{\delta^2} \\
&\quad - 2e^{-\delta^{-2}\langle f_\varepsilon, g_\varepsilon\rangle_H}\left(1 - \frac{\langle f_\varepsilon, g_\varepsilon\rangle_H}{\delta^2} + \frac{\langle g_\varepsilon, g_\varepsilon\rangle_H}{\delta^2}\right),
\end{aligned}$$

which, using Lemma 2.8, is easily seen to be $\mathrm{o}\big(\|g\|_\infty^2\big)$, as desired. Compare to Barbour et al. (2024, (2.12-15)).

Assuming (2.13) holds for k, we want to show it holds for $k+1$. We write

$$\begin{aligned}
&D^k\zeta_{\varepsilon,\delta}(f+g)[g^{(1)},\dots,g^{(k)}] - D^k\zeta_{\varepsilon,\delta}(f)[g^{(1)},\dots,g^{(k)}] \\
&\qquad = \mathbb{E}\left[\zeta(\delta S)\big(e^{\Psi_\varepsilon(f+g)} - e^{\Psi_\varepsilon(f)}\big)\sum_{\pi\in\mathcal{P}_{k,2}}\prod_{b\in\pi} D^{|b|}\Psi_\varepsilon(f)[g^{(b)}]\right] &(2.17)\\
&\qquad\quad + \mathbb{E}\left[\zeta(\delta S)e^{\Psi_\varepsilon(f)}\sum_{\pi\in\mathcal{P}_{k,2}}\Big(\prod_{b\in\pi} D^{|b|}\Psi_\varepsilon(f+g)[g^{(b)}] - \prod_{b\in\pi} D^{|b|}\Psi_\varepsilon(f)[g^{(b)}]\Big)\right] &(2.18)
\end{aligned}$$

Because of (2.16), the term (2.17) is equal to

$$\mathbb{E}\left[\zeta(\delta S)D\Psi_\varepsilon(f)[g]\sum_{\pi\in\mathcal{P}_{k,2}}\prod_{b\in\pi} D^{|b|}\Psi_\varepsilon(f)[g^{(b)}]\right] + \mathrm{o}\big(\|g\|_\infty\big). \tag{2.19}$$

Now working on (2.18), noting that $D^2\Psi_\varepsilon(f+g) = D^2\Psi_\varepsilon(f)$ and $D\Psi_\varepsilon(f+g)[h] = D\Psi_\varepsilon(f)[h] + D^2\Psi_\varepsilon(f)[h,g]$, we find

$$\begin{aligned}
&\sum_{\pi\in\mathcal{P}_{k,2}}\Big(\prod_{b\in\pi} D^{|b|}\Psi_\varepsilon(f+g)[g^{(b)}] - \prod_{b\in\pi} D^{|b|}\Psi_\varepsilon(f)[g^{(b)}]\Big) \\
&\qquad = \sum_{\pi\in\mathcal{P}_{k,2}}\prod_{\substack{b\in\pi\\|b|=2}} D^2\Psi_\varepsilon(f)[g^{(b)}]\Big\{\prod_{\substack{b\in\pi\\|b|=1}}\Big(D\Psi_\varepsilon(f)[g^{(b)}] + D^2\Psi_\varepsilon(f)[g^{(b)},g]\Big) - \prod_{\substack{b\in\pi\\|b|=1}} D\Psi_\varepsilon(f)[g^{(b)}]\Big\} \\
&\qquad = \sum_{\pi\in\mathcal{P}_{k,2}}\prod_{\substack{b\in\pi\\|b|=2}} D^2\Psi_\varepsilon(f)[g^{(b)}]\Big\{\sum_{\substack{b\in\pi\\|b|=1}} D^2\Psi_\varepsilon(f)[g^{(b)},g]\prod_{\substack{b\neq a\in\pi\\|a|=1}} D\Psi_\varepsilon(f)[g^{(a)}]\Big\} + \mathbf{o}\big(\|g\|_\infty\big),
\end{aligned}$$

where $\mathbf{o}\big(\|g\|_\infty\big)$ is a random variable, say $X = X(g)$ depending on g, such that $\mathbb{E}\big[|X|^p\big]^{1/p} = \mathrm{o}\big(\|g\|_\infty\big)$ for any $p \geqslant 2$. This is because $D\Psi_\varepsilon(f)[g^{(b)}]$ is Gaussian, and $D^2\Psi_\varepsilon(f)[g^{(b)}, g] = \mathrm{O}\big(\|g\|_\infty\big)$, by Lemma 2.8. Thus, up to a $\mathrm{o}\big(\|g\|_\infty\big)$ term, (2.18) is equal to

$$\mathbb{E}\bigg[\zeta(\delta S)e^{\Psi_\varepsilon(f)} \sum_{\pi\in\mathcal{P}_{k,2}} \prod_{\substack{b\in\pi\\|b|=2}} D^2\Psi_\varepsilon(f)[g^{(b)}]\bigg\{\sum_{\substack{b\in\pi\\|b|=1}} D^2\Psi_\varepsilon(f)[g^{(b)}, g] \prod_{\substack{b\neq a\in\pi\\|a|=1}} D\Psi_\varepsilon(f)[g^{(a)}]\bigg\}\bigg]. \tag{2.20}$$

Combining (2.19) and (2.20) completes the induction. □

3 PROOF OF THE WASSERSTEIN BOUND

Armed with Theorem 2.9, we follow the strategy described in Section 1.1 to prove our master theorem.

Proof of Theorem 1.1. To achieve a bound in the Wasserstein distance, let $\zeta : \mathrm{C}(\mathcal{S}^n; \mathbb{R}^d)$ be a Lipschitz function and let $\zeta_{\varepsilon,\delta}$ be defined as in Theorem 2.9. The triangle inequality yields

$$\begin{aligned}&|\mathbb{E}\zeta(F) - \mathbb{E}\zeta(H)|\\ &\quad\leqslant \big|\mathbb{E}[\zeta_{\varepsilon,\delta}(F)] - \mathbb{E}[\zeta_{\varepsilon,\delta}(H)]\big| + \big|[\mathbb{E}\zeta(F)] - \mathbb{E}[\zeta_{\varepsilon,\delta}(F)]\big| + \big|[\mathbb{E}[\zeta(H)] - \mathbb{E}\zeta_{\varepsilon,\delta}(H)]\big|.\end{aligned} \tag{3.1}$$

For the first term of (3.1), we use (2.12) of Theorem 2.9 with $k = 2$ and the definition (1.1) of $\mathcal{F}$ to find

$$\big|\mathbb{E}[\zeta_{\varepsilon,\delta}(F)] - \mathbb{E}[\zeta_{\varepsilon,\delta}(H)]\big| \leqslant c\, d\, \delta^{-2}\varepsilon^{-2(n+\iota)}\mathsf{d}_{\mathcal{F}}(F, H).$$

For the second term, using the definition of $\zeta_{\varepsilon,\delta}$ and that ζ is Lipschitz implies

$$\big|\mathbb{E}[\zeta(F)] - \mathbb{E}[\zeta_{\varepsilon,\delta}(F)]\big| = \big|\mathbb{E}[\zeta(F)] - \mathbb{E}[\zeta(F_\varepsilon + \delta S)]\big| \leqslant \mathbb{E}\|F - F_\varepsilon\|_\infty + \delta\,\mathbb{E}\|S\|_\infty,$$

where $S : \mathcal{S}^n \to \mathbb{R}^d$ is the smoothing Gaussian random field defined at (2.6). Since S has independent components and, by (2.6) and (2.5), has covariance uniformly bounded in absolute value of order $c_n \sum_{k\geqslant 1} k^{-1-\iota}$ for some constant c_n depending only on n, Fernique's theorem implies

$$\mathbb{E}\|S\|_\infty \leqslant c_{n,\iota}\,\sqrt{d},$$

where $c_{n,\iota}$ constant depending only on n and ι. Thus, we find

$$\big|\mathbb{E}[\zeta(F)] - \mathbb{E}[\zeta_{\varepsilon,\delta}(F)]\big| \leqslant \mathbb{E}\|F - F_\varepsilon\|_\infty + c_{n,\iota}\delta\sqrt{d}.$$

The same reasoning shows that this same inequality holds with H replacing F. Substituting these bounds in (3.1) verifies that the desired bound in (1.2) holds. □

4 PROPERTIES OF SOLUTION TO THE STEIN EQUATION

Applications of Theorem 1.1 require bounds on the first three terms of the right-hand side of (1.2). In this section, we bound the first term for the case when $H = G$, the approximating Gaussian field. We handle the second and the third terms in the following section, see Lemma 5.3 in particular.

We start with the following result, which extends the work of Barbour et al. (2023, Section 2) and provides properties of solutions to infinite-dimensional versions of Stein's equation. Specifically, Barbour et al. (2023) worked with Stein's equations for Gaussian processes indexed by an interval $[0, T]$, whereas here we work with random fields indexed by a compact measured metric space.

Theorem 4.1 (Bounds on solutions of the Stein equation)**.** *For a Gaussian random field* $G \in \mathrm{C}(\mathcal{M};\mathbb{R}^d)$*, define the operator* $\mathcal{A} = \mathcal{A}_G$ *acting on* $\zeta : \mathrm{C}(\mathcal{M};\mathbb{R}^d) \to \mathbb{R}$ *with*

$$\max_{k=1,2} \sup_{g\in \mathrm{C}(\mathcal{M};\mathbb{R}^d)} \|D^k\zeta(g)\| < \infty,$$

by

$$\mathcal{A}\zeta(f) := \mathbb{E}\big[D^2\zeta(f)[G,G]\big] - D\zeta(f)[f],$$

where $f \in \mathrm{C}(\mathcal{M};\mathbb{R}^d)$*, and* D *denotes Frechét derivative. Then for any such* ζ*, there exists an* $\eta = \eta_\zeta$ *satisfying*

$$\mathcal{A}\eta(f) = \zeta(f) - \mathbb{E}[\zeta(G)]. \tag{4.1}$$

Moreover, in the operator norm, for any $k = 1, 2$*, or* $k \geqslant 3$ *with* $\sup_{g\in \mathrm{C}(\mathcal{M};\mathbb{R}^d)} \|D^k\zeta(g)\| < \infty$*, we have*

$$\|D^k\eta(f)\| \leqslant \frac{1}{k} \sup_{g\in \mathrm{C}(\mathcal{M};\mathbb{R}^d)} \|D^k\zeta(g)\|, \tag{4.2}$$

and for any $\zeta \in \mathcal{F}$ *and all* $f, h \in C(\mathcal{M};\mathbb{R}^d)$*, we have*

$$\|D^2\eta(f) - D^2\eta(h)\| \leqslant \frac{1}{3}\|f - h\|_\infty. \tag{4.3}$$

Remark 4.2. The operator $\mathcal{A}$ defined in Theorem 4.1 plays the role of the left-hand side of the 'random field' version

$$\mathbb{E}\big[D^2\eta(f)[G,G]\big] - D\eta(f)[f] = \zeta(f) - \mathbb{E}[\zeta(G)] \tag{4.4}$$

of the finite dimensional Stein equation for a centered Gaussian G with covariance matrix $\Sigma = (\sigma_{ij})$ given by

$$\sum_{i,j} \sigma_{ij}\partial_{ij}\eta(x) - \sum_i x_i\partial_i\eta(x) = \zeta(x) - \mathbb{E}[\zeta(G)],$$

where ∂_i and ∂_{ij} denote the first and second partial derivatives respectively in coordinates i, j . With the Stein equation (4.4) and the bounds on its solution provided by Theorem 4.1, the standard steps of Stein's method can be implemented. In particular, the integral probability metric bound to G over some given function class $\mathcal{H}$ can be computed by bounding the absolute expectation of the right-hand side of (4.4) for $\zeta \in \mathcal{H}$ by taking absolute expectations on the left-hand side, given in terms of the solution η. In particular, uniformly bounding $|\mathbb{E}\mathcal{A}\eta_\zeta(F)|$ for all solutions $\eta_\zeta, \zeta \in \mathcal{F}$ to (4.4) yields a bound on $\mathtt{d}_{\mathcal{F}}(F, G)$, the first term on the right-hand side of (1.2). See Lemma 6.1 and its proof for the implementation.

Remark 4.3. The term $\mathbb{E}\big[D^2\zeta(f)[G,G]\big]$ implicitly depends on the covariance C of G. As discussed in Remark 4.2, if G is finite-dimensional, then this term evaluates explicitly to $\nabla^\top\Sigma\nabla\eta(x)$, where Σ is the covariance matrix of G. When G is a random field indexed by an uncountable set, in general it is not clear how to write this term solely in terms of C. In applications, the term should be rewritten in a form that matches the particular application and does not involve an expectation against G. Typically, this form involves the covariance structure of G and is most easily found using some structure of the random field F to determine the first order term in a Taylor expansion of $\mathbb{E}\big[D\eta(F)[F]\big]$. See Section 6 for further details from our application to wide random neural networks, and also Barbour et al. (2023) for applications that provide additional relevant examples.

Proof of Theorem 4.1. The result essentially follows from the work of Barbour et al. (2023, Section 2), building off Barbour (1990) and Kasprzak et al. (2017), in the setting where the index set of the process f is the interval $[0, T]$.

Fix ζ with two bounded derivatives. For $f \in \mathrm{C}(\mathcal{M}; \mathbb{R}^d)$ define $h_f : \mathbb{R}_+ \to \mathbb{R}$ by $h_f(t) := \mathbb{E}[\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)]$, and $\eta = \eta_\zeta$ to be

$$\eta(f) = -\int_0^\infty \big(h_f(t) - \mathbb{E}[\zeta(G)]\big)\mathrm{d}t.$$

The integral is well-defined since ζ has bounded derivative, $\|f\|_\infty$ is finite by continuity and compactness of $\mathcal{M}$, and $\|G\|_\infty$ has finite moments, by Gaussianity and path continuity. That $\mathcal{A}$ can be applied to η (meaning it has two bounded derivatives) follows essentially from Barbour (1990), see also Kasprzak et al. (2017), since dominated convergence implies that if $\sup_g \|D^k\zeta(g)\| < \infty$, we have the well-defined expressions

$$D^k\eta(f)[g_1, \ldots, g_k] = -\int_0^\infty e^{-kt}\,\mathbb{E}\big\{D^k\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[g_1, \ldots, g_k]\big\}\mathrm{d}t, \tag{4.5}$$

from which the bounds in (4.2) easily follow. For (4.3), if $\zeta \in \mathcal{F}$ applying (4.5) with $k = 2$ and (1.1), that assures elements of $\mathcal{F}$ to have a Lipschitz-1 second derivative, imply

$$\begin{aligned}
&\big|D^2\eta(f)[g,g] - D^2\eta(h)[g,g]\big| \\
&\qquad\leqslant \int_0^\infty e^{-2t}\,\mathbb{E}\big\{\big|D^2\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[g,g] - D^2\zeta(e^{-t}h + \sqrt{1-e^{-2t}}G)[g,g]\big|\big\}\mathrm{d}t \\
&\qquad\leqslant \|g\|_\infty^2 \int_0^\infty e^{-2t}\|e^{-t}f - e^{-t}h\|_\infty \mathrm{d}t \\
&\qquad\leqslant \frac{1}{3}\|f-h\|_\infty\|g\|_\infty^2.
\end{aligned}$$

We now show (4.1). We have

$$\begin{aligned}
\zeta(f) - \mathbb{E}[\zeta(G)] &= -\int_0^\infty h_f'(t)\mathrm{d}t \\
&= \int_0^\infty e^{-t}\,\mathbb{E}\big[D\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[f]\big]\mathrm{d}t \\
&\qquad - \int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\,\mathbb{E}\big[D\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[G]\big]\mathrm{d}t \\
&= -D\eta(f)[f] - \int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\,\mathbb{E}\big[D\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[G]\big]\mathrm{d}t,
\end{aligned}$$

where the third equality follows by (4.5). Comparing to (4.1), we will have shown the first claim if we can demonstrate that

$$\mathbb{E}\big[D^2\eta(f)[G,G]\big] = -\int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\,\mathbb{E}\big[D\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[G]\big]\mathrm{d}t.$$

Evaluating (4.5) for $k = 2$, the left-hand side of the previous display can be expressed as

$$-\int_0^\infty e^{-2t}\,\mathbb{E}\big[D^2\zeta(e^{-t}f + \sqrt{1-e^{-2t}}G)[G', G']\big]\mathrm{d}t,$$

where G' is an independent copy of G. But the integrands are equal since, for fixed t and f and $\varphi(g) \coloneqq \zeta(e^{-t}f + \sqrt{1-e^{-2t}}g)$, Barbour et al. (2023, Proof of Proposition 2.1) implies

$$\mathbb{E}\big[D\varphi(G)[G]\big] = \mathbb{E}\big[D^2\varphi(G)[G', G']\big],$$

and

$$D\varphi(g)[g_1] = \sqrt{1-e^{-2t}}D\zeta(e^{-t}f + \sqrt{1-e^{-2t}}g)[g_1],$$
$$D^2\varphi(g)[g_1, g_2] = (1-e^{-2t})D^2\zeta(e^{-t}f + \sqrt{1-e^{-2t}}g)[g_1, g_2].$$

Note that here we are using the Karhunen-Loeve expansion which is the part of the argument that uses the Borel measure ν on our metric space $(\mathcal{M}, \mathtt{d})$. □

There have been a number of recent works developing Stein's method for processes, predominantly in the context of distributional approximation by interval-indexed Gaussian processes, and especially Brownian motion; though see Gan and Ross (2021) for an exception. Building from the seminal work of Barbour (1990), Shih (2011) develops Stein's method in the very general setting of a Gaussian measure on a separable Banach space. However, the bounds there are too abstract to be evaluated explicitly in practice. Closely following Shih (2011), the works Coutin and Decreusefond (2013, 2020); Bourguin and Campese (2020) provide more concrete results in the less general setting of a Gaussian measure on a Hilbert space. However, the associated probability metrics are with respect to the Hilbert space topology, e.g., L^2 and Sobolev, which are quite weak and do not see fundamental natural statistics such as finite dimensional distributions and extrema. The works of Kasprzak (2020a,b); Dobler and Kasprzak (2021), based on Barbour (1990), are more closely related to our work, but work only in smooth function metrics like $\mathtt{d}_{\mathcal{F}}$. We refer to Barbour et al. (2024, Section 1.1) for additional details and comparisons.

5 CHAINING ARGUMENTS FOR MODULUS OF CONTINUITY

We now present results for bounding the second and third terms in (1.2) that arise from the smoothing process. We start with a proposition that is useful for obtaining probabilistic bounds on the modulus of continuity of an $\mathbb{R}^d$-valued random field on a compact metric space $(\mathcal{M}, \mathtt{d})$.

Definition 5.1 (Modulus of Continuity)**.** The modulus of continuity of a function $J : \mathcal{M} \to \mathbb{R}^d$ at level $\theta > 0$ is defined as $\omega_J(\theta) \coloneqq \sup\big\{\|J(x) - J(y)\|_2 : x, y \in \mathcal{M}, \mathtt{d}(x,y) < \theta\big\}$.

While the proofs below leverage standard chaining arguments, existing results seem not to provide the form of the results we require, as those mainly focus on expectation bounds and the case of $d = 1$.

Define the *covering number* $\mathcal{N}(\mathcal{M}, \mathtt{d}, \varepsilon)$ (or just $\mathcal{N}(\varepsilon)$ when $(\mathcal{M}, \mathtt{d})$ is clear from context) of $(\mathcal{M}, \mathtt{d})$ at level $\varepsilon > 0$ as the smallest cardinality over finite collections of points $U \subseteq \mathcal{M}$ so that every point of $\mathcal{M}$ is within ε of some point of U (i.e., U is an ε-net).

Proposition 5.2. *Let $(\mathcal{M}, \mathtt{d})$ be a compact metric space and let $(H(x))_{x\in\mathcal{M}}$ be an $\mathbb{R}^d$-valued random field with continuous paths and write $H = (H_1, \dots, H_d)$. Suppose there exist positive constants c_0, β, γ and c_1 such that for any $x, y \in \mathcal{M}$ and $i = 1, \dots, d$, we have*

$$\mathbb{P}\big(|H_i(x) - H_i(y)| \geqslant \lambda\big) \leqslant c_0 \frac{\mathtt{d}(x,y)^\beta}{\lambda^\gamma} \quad \text{for all } \lambda > 0, \tag{5.1}$$

and for every $\varepsilon > 0$ the covering numbers satisfy

$$\mathcal{N}(\varepsilon) \leqslant c_1 \varepsilon^{-\alpha}. \tag{5.2}$$

Then, if $\alpha < \beta/2$ there is a constant c depending only on $\operatorname{diam}(\mathcal{M})$, $\alpha, \beta, \gamma, c_0, c_1$ *such that for all* $i = 1, \dots, d$ *and* $\theta > 0$,

$$\mathbb{P}\big(\omega_{H_i}(\theta) > \lambda\big) \leqslant c\,\frac{\theta^{\beta-2\alpha}}{\lambda^{\gamma}}, \tag{5.3}$$

and for any $0 < k < \gamma$,

$$\mathbb{E}\big[\omega_H(\theta)^k\big] \leqslant c\,d^{k/2}\,\theta^{k(\beta-2\alpha)/\gamma}. \tag{5.4}$$

Proof. Following Pollard (1984, Chapter VII, Section 2, 9 Chaining Lemma), we can construct a nested sequence of subsets $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \cdots \subseteq \mathcal{M}$ such that every $t \in \mathcal{M}$ is within $\operatorname{diam}(\mathcal{M})2^{-i}$ of a point of $\mathcal{M}_i$, and

$$|\mathcal{M}_i| \leqslant \mathcal{N}\big(\operatorname{diam}(\mathcal{M})2^{-(i+1)}\big) \leqslant c_1\frac{2^{\alpha(i+1)}}{\operatorname{diam}(\mathcal{M})^{\alpha}}. \tag{5.5}$$

For each $x \in \mathcal{M}$, there is a sequence $(x_i)_{i\geqslant 1}$ with $x_i \in \mathcal{M}_i$ and $\lim_{i\to\infty} x_i = x$; in particular $\mathcal{M}^* := \bigcup_i \mathcal{M}_i$ is dense in $\mathcal{M}$.

We first show that for all $\theta > 0$,

$$\mathbb{P}(w_H(\theta) > \lambda) = \mathbb{P}(w_H^*(\theta) > \lambda),$$

where $w_H^*(\theta) = \sup\big\{\|H(x) - H(y)\|_2 : x, y \in \mathcal{M}^*, \mathtt{d}(x,y) \leqslant \theta\big\}$ is the modulus of continuity of $H : \mathcal{M}^* \to \mathbb{R}$ at level θ, that is, only considering points in $\mathcal{M}^*$. The equality holds as the events are equal. Clearly $\{w_H^*(\theta) > \lambda\} \subseteq \{w_H(\theta) > \lambda\}$ since $\mathcal{M}^* \subseteq \mathcal{M}$. For the other direction, if there are $x, y \in \mathcal{M}$ with $\mathtt{d}(x,y) < \theta$ and $\|H(x) - H(y)\|_2 > \lambda$, then letting $x_i, y_i \in \mathcal{M}_i$ such that $x_i \to x$ and $y_i \to y$, then $\mathtt{d}(x_i, y_i) \to \mathtt{d}(x,y) < \theta$ and continuity of H implies that $\|H(x_i) - H(y_i)\|_2 \to \|H(x) - H(y)\|_2 > \theta$ and so there must be some i with $\mathtt{d}(x_i, y_i) < \theta$ and $\|H(x_i) - H(y_i)\|_2 > \lambda$.

To bound $w_H^*(\theta)$, let $\theta > 0$ be fixed and let x, y be arbitrary points in $\mathcal{M}^*$ satisfying $\mathtt{d}(x,y) < \theta$. Since the $\mathcal{M}_i$'s are nested, there exists n such that $x, y \in \mathcal{M}_{n+1}$. Further, there are sequences $(x_i)_{i=0}^n, (y_i)_{i=0}^n$ such that $x_i, y_i \in \mathcal{M}_i$, and $\mathtt{d}(x_i, x_{i+1}) \vee \mathtt{d}(y_i, y_{i+1}) \leqslant \operatorname{diam}(\mathcal{M})2^{-i+1}$, letting $x_{n+1} := x$ and $y_{n+1} := y$. These sequences can be constructed sequentially, e.g., set x_n to be the nearest point in $\mathcal{M}_n$ to x, which must be within $\operatorname{diam}(\mathcal{M})2^{-n}$ since $\mathcal{M}_n$ is a $\operatorname{diam}(\mathcal{M})2^{-n}$-net. Given that we know x_{i+1}, we choose x_i to be the point in $\mathcal{M}_{i+1}$ that is closest to x_i. Since $\mathcal{M}_i \subseteq \mathcal{M}_{i+1}$, there must be such a point with distance no greater than $\operatorname{diam}(\mathcal{M})2^{-i+1}$.

Denoting the maximum change in H over points in $\mathcal{M}_i$ that are within $\operatorname{diam}(\mathcal{M})\rho$ by

$$D_i(\rho) := \sup\big\{\|H(u) - H(v)\|_2 : u, v \in \mathcal{M}_i,\, \mathtt{d}(u,v) \leqslant \operatorname{diam}(\mathcal{M})\rho\big\}.$$

Set $m = \lfloor -\log_2(\theta/\operatorname{diam}(\mathcal{M}))\rfloor$, implying in particular that $\theta \leqslant \operatorname{diam}(\mathcal{M})2^{-m}$. The triangle inequality implies that

$$\begin{aligned}\|H(x) - H(y)\|_2 &\leqslant \|H(x_m) - H(y_m)\|_2 + \sum_{i=m}^{n}\Big(\|H(x_{i+1}) - H(x_i)\|_2 + \|H(y_{i+1}) - H(y_i)\|_2\Big)\\ &\leqslant D_m\big(2^{-m+2}\big) + 2\sum_{i=m}^{\infty} D_i\big(2^{-i+1}\big),\end{aligned} \tag{5.6}$$

where the 2^{-m+2} in the second inequality follows by the triangle inequality

$$\mathtt{d}(x_m, y_m) \leqslant d(x,y) + \sum_{i=m}^{n}\mathtt{d}(x_i, x_{i+1}) + \sum_{i=m}^{n}\mathtt{d}(y_i, y_{i+1}) < \theta + \sum_{i=m}^{n}\mathtt{d}(x_i, x_{i+1}) + \sum_{i=m}^{n}\mathtt{d}(y_i, y_{i+1})$$

$$\leqslant \operatorname{diam}(\mathcal{M})2^{-m} + 2\operatorname{diam}(\mathcal{M})\sum_{i=m}^{n} 2^{-i+1} \leqslant 3\operatorname{diam}(\mathcal{M})2^{-m} \leqslant \operatorname{diam}(\mathcal{M})2^{-m+2}.$$

Noting (5.6), we set $\lambda_i = (1-a)a^{i-m}(\lambda/3)$ for $i \geqslant m$, for some $a \in (0,1)$ to be chosen later, and applying the union bound we have

$$\mathbb{P}\big(\omega_H^*(\theta) > \lambda\big) \leqslant \mathbb{P}\big(D_m\big(2^{-m+2}\big) > \lambda/3\big) + \sum_{i=m}^{\infty} \mathbb{P}\big(D_i\big(2^{-i+1}\big) > \lambda_i\big). \tag{5.7}$$

Now, again using a union bound, (5.1) and (5.5) yields that

$$\begin{aligned}
\mathbb{P}\big(D_i(\rho) > \lambda\big) &\leqslant \sum_{\substack{u,v\in\mathcal{M}_i \\ \mathtt{d}(u,v)\leqslant \operatorname{diam}(\mathcal{M})\rho}} \mathbb{P}\big(\|H(u) - H(v)\|_2 > \lambda\big) \\
&\leqslant c_0|\mathcal{M}_i|^2 \frac{\operatorname{diam}(\mathcal{M})^\beta \rho^\beta}{\lambda^\gamma} \\
&\leqslant c_0\, c_1^2\, 2^{2\alpha(i+1)} \frac{\operatorname{diam}(\mathcal{M})^{\beta-2\alpha}\rho^\beta}{\lambda^\gamma},
\end{aligned}$$

and so for a constant c' depending on $\operatorname{diam}(\mathcal{M})$, $\alpha, \beta, \gamma, c_0, c_1$, using that the first term in (5.7) can be bounded by a constant depending only on α, β, γ and a times the bound on the first term of the sum that follows, we have

$$\mathbb{P}\big(\omega_H^*(\theta) > \lambda\big) \leqslant c'(1-a)^{-\gamma}\lambda^{-\gamma} \sum_{i=m}^{\infty} \left(\frac{2^{2\alpha-\beta}}{a^\gamma}\right)^i.$$

Since $2\alpha - \beta < 0$, it is possible to choose a such that $r := 2^{2\alpha-\beta}/a^\gamma < 1$. So doing, we obtain

$$\sum_{i=m}^{\infty} \left(\frac{2^{2\alpha-\beta}}{a^\gamma}\right)^i = (1-r)^{-1}r^m = (1-r)a^{-\gamma m}2^{(2\alpha-\beta)m} \leqslant (1-r)2^{(2\alpha-\beta)m},$$

where we have used that $a \in (0,1)$ and is being raised to a positive power, so can be bounded by 1. Recalling that $m = \lfloor -\log_2(\theta/\operatorname{diam}(\mathcal{M}))\rfloor$, we hence observe that there is a constant c depending on $\operatorname{diam}(\mathcal{M})$, $\alpha, \beta, \gamma, c_0, c_1$, such that

$$\mathbb{P}\big(\omega_H^*(\theta) > \lambda\big) \leqslant c\,\frac{\theta^{\beta-2\alpha}}{\lambda^\gamma}.$$

Now, we proceed to prove (5.4) starting with the case $d = 1$. Letting $\tilde{c}$ be a constant that may vary from line to line, but will at most only depend on $\operatorname{diam}(\mathcal{M}), \alpha, \beta, \gamma, c_0, c_1$, the result easily follows from (5.3), since under our hypotheses that $0 < k < \gamma$ we have

$$\begin{aligned}
\mathbb{E}\big[\omega_F(\theta)^k\big] &= \int_0^\infty \mathbb{P}\big(\omega_F(\theta) > \lambda^{1/k}\big)\mathrm{Leb}(\mathrm{d}\lambda) \\
&= \int_0^{\theta^{k(\beta-2\alpha)/\gamma}} \mathbb{P}\big(\omega_F(\theta) > \lambda^{1/k}\big)\mathrm{Leb}(\mathrm{d}\lambda) + \int_{\theta^{k(\beta-2\alpha)/\gamma}}^\infty \mathbb{P}\big(\omega_F(\theta) > \lambda^{1/k}\big)\mathrm{Leb}(\mathrm{d}\lambda) \\
&\leqslant \theta^{k(\beta-2\alpha)/\gamma} + \tilde{c}\int_{\theta^{k(\beta-2\alpha)/\gamma}}^\infty \frac{\theta^{\beta-2\alpha}}{\lambda^{\gamma/k}}\mathrm{Leb}(\mathrm{d}\lambda) \\
&\leqslant \tilde{c}\,\theta^{k(\beta-2\alpha)/\gamma},
\end{aligned} \tag{5.8}$$

as desired.

Now, for general $d \geqslant 1$, it is clear from the definition of the modulus of continuity that

$$\omega_F^2(\theta) \leqslant \sum_{i=1}^{d} \omega_{F_i}^2(\theta). \tag{5.9}$$

Raising both sides of (5.9) to any positive power $k \geqslant 1$, and using that $(\sum_i a_i)^k \leqslant d^{k-1} \sum_i a_i^k$ for non-negative a_i, we have

$$\omega_F^{2k}(\theta) \leqslant \left(\sum_{i=1}^{d} \omega_{F_i}^2(\theta) \right)^k \leqslant d^{k-1} \sum_{i=1}^{d} \omega_{F_i}^{2k}(\theta).$$

Taking expectation on both sides and applying (5.8), we have

$$\Big(\mathbb{E}[\omega_F^k(\theta)] \Big)^2 \leqslant \mathbb{E}[\omega_F^{2k}(\theta)] \leqslant d^{k-1} \sum_{i=1}^{d} \mathbb{E}[\omega_{F_i}^{2k}(\theta)] \leqslant \tilde{c} d^k \, \theta^{2k(\beta-2\alpha)/\gamma},$$

and taking square roots yields the desired inequality. □

Lemma 5.3. *Let $\mathcal{M} = \mathcal{S}^n \subset \mathbb{R}^{n+1}$ for some $n \geqslant 2$, with natural geodesic metric $\mathtt{d}$, and $H = (H_1, \ldots, H_d) : \mathcal{S}^n \to \mathbb{R}^d$ be a random field with continuous paths, and H_ε be the ε-regularization of H defined at (2.9) for a fixed $0 < \varepsilon < 1$. If for all $i = 1, \ldots, d$, for all $x, y \in \mathcal{S}^n$, some constant $\hat{c}$, and some $p > n$ we have*

$$\mathbb{E}\big[\big(H_i(x) - H_i(y) \big)^{2p} \big] \leqslant \hat{c} \, \mathtt{d}(x,y)^{2p}, \tag{5.10}$$

then there is a constant c depending only on $\hat{c}, n$, and p, such that

$$\mathbb{E}\|H - H_\varepsilon\|_\infty \leqslant c \sqrt{d} \, \varepsilon^{\frac{1}{2}(1-\frac{n}{p})} \sqrt{\log(1/\varepsilon)}.$$

Proof. Using the alternative expression for H_ε given at (2.11) in Proposition 2.6, for any given $\theta > 0$ we immediately have

$$H(x) - H_\varepsilon(x) = \int_{y : d(x,y) \leqslant \theta} p(x,y;\varepsilon) \big(H(x) - H(y) \big) \mathrm{d}y + \int_{y : d(x,y) > \theta} p(x,y;\varepsilon) \big(H(x) - H(y) \big) \mathrm{d}y,$$

where $\mathrm{d}y$ is the volume element on the sphere. It is easy to see that $\omega_H(\theta)$ is finite because H is continuous and the sphere is compact. Hence, we can further bound

$$\begin{aligned} \big\| H(x) - H_\varepsilon(x) \big\|_2 &\leqslant \omega_H(\theta) + \sup_{u,v \in \mathcal{S}^n} \big\| H(u) - H(v) \big\|_2 \int_{y : d(x,y) > \theta} p(x,y;\varepsilon) \mathrm{d}y \\ &= \omega_H(\theta) + \omega_H(\pi) \int_{y : d(x,y) > \theta} p(x,y;\varepsilon) \mathrm{d}y. \end{aligned}$$

The heat kernel bounds of Nowak et al. (2019, Theorem 1) imply

$$\int_{y : d(x,y) > \theta} p(x,y;\varepsilon) \mathrm{d}y \leqslant c_n \, e^{-\theta^2/(5\varepsilon)},$$

where c_n is a constant depending only on n. Hence, we have that

$$\mathbb{E}\|H - H_\varepsilon\|_\infty \leqslant \mathbb{E}\big[\omega_H(\theta) \big] + c_n \, \mathbb{E}\big[\omega_H(\pi) \big] e^{-\theta^2/(5\varepsilon)}. \tag{5.11}$$

To bound $\mathbb{E}[\omega_H(\theta)]$ we apply Proposition 5.2, and use Markov's inequality to find that

$$\mathbb{P}\big(|H_i(x) - H_i(y)| \geqslant \lambda\big) \leqslant \frac{\mathbb{E}\big[(H_i(x) - H_i(y))^{2p}\big]}{\lambda^{2p}}.$$

Therefore (5.1) is satisfied with $\beta = \gamma = 2p$ and $c_0 = \hat{c}$, due to our assumption (5.10). To bound the covering numbers, standard volume arguments (see, for example, Vershynin (2018, Corollary 4.2.13)) imply that for all $\varepsilon \in (0,1)$, we have

$$\mathcal{N}(\mathcal{S}^n, \mathtt{d}, \varepsilon) \leqslant c_n\, \varepsilon^{-n},$$

where c_n is a constant depending only on n, thus (5.2) is satisfied with $\alpha = n$. Applying (5.4) of Proposition 5.2 we find that there exists a constant c, whose value from line to line may change, but depends that only on $\hat{c}, n$, and p, such that

$$\mathbb{E}\big[\omega_H(\theta)\big] \leqslant c\sqrt{d}\,\theta^{1-\frac{n}{p}}.$$

Substituting this inequality in (5.11) and setting $\theta = \sqrt{l\varepsilon\log(1/\varepsilon)}$ and $l \geqslant 5(1-\frac{n}{p})/2$, we conclude that

$$\mathbb{E}\|H - H_\varepsilon\|_\infty \leqslant c\sqrt{d}\,\varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)}. \qquad \square$$

6 PROOFS FOR WIDE RANDOM NEURAL NETWORK APPROXIMATIONS

We now apply the general results developed in the previous sections to prove Theorems 1.2 and 1.4 on the smooth and Wasserstein distance bounds for wide random neural network. We follow the strategy based on induction as previously described in Section 1.2. We first present the following result, obtained by applying Theorem 4.1 at a given, single layer of the network. One key element driving the result is the use of the classical Stein 'leave-one-out' approach, see (6.5).

Lemma 6.1. *Let $H : \mathcal{M} \to \mathbb{R}^m$ be a random field with continuous and i.i.d. coordinate processes $H_1, \dots, H_m$, and let $W : \mathbb{R}^m \to \mathbb{R}^n$ be an $n \times m$ random matrix that is independent of H and has centered independent entries having the same variance $\mathrm{Var}(W_{ij}) =: c_w/m$, also satisfying $\mathbb{E}[W_{ij}^4] \leqslant B(c_w/m)^2$, and $\sigma : \mathbb{R} \to \mathbb{R}$. Define $F : \mathcal{M} \to \mathbb{R}^n$ by*

$$F(x) = W\sigma\big(H(x)\big),$$

and assume $F \in L^2(\mathcal{M}; \mathbb{R}^n)$. Let $G \in \mathrm{C}(\mathcal{M}; \mathbb{R}^d)$ be a centered Gaussian random field with covariance function

$$C_{ij}(x,y) := \mathbb{E}\big[F_i(x)F_j(y)\big] = \delta_{ij}c_w\,\mathbb{E}\big[\sigma\big(H_1(x)\big)\sigma\big(H_1(y)\big)\big].$$

Then for any $\zeta \in \mathcal{F}$, we have

$$\big|\mathbb{E}[\zeta(F)] - \mathbb{E}[\zeta(G)]\big| \leqslant c_w^{3/2}B^{3/4}\,\mathbb{E}\big[\|\sigma(H_1)\|_\infty^3\big]\frac{n^{3/2}}{\sqrt{m}}. \tag{6.1}$$

Proof. We apply Theorem 4.1 with the Gaussian random field G and $d = n$. In particular, we obtain the bound (6.1) by substituting F for f in (4.4) and bounding the expectation of its right-hand side. Our first step is to derive a more useful representation for the second order term. We claim

$$\mathbb{E}\big[D^2\eta(f)[G,G]\big] = \mathbb{E}\big[D^2\eta(f)[W\sigma(H), W\sigma(H)]\big]. \tag{6.2}$$

More generally, if the covariance of a centered Gaussian random field $G \in \mathrm{C}(\mathcal{M}; \mathbb{R}^d)$ satisfies $C_{ij}(x,y) = \delta_{ij}\,\mathbb{E}[G_i(x)G_i(y)] = \mathbb{E}[R_i(x)R_j(y)]$ for some centered $L^2(\mathcal{M}; \mathbb{R}^d)$ random field R (not

necessarily assumed Gaussian), then for any bilinear form A with $\mathbb{E}\big[A[G,G]\big] < \infty$, we have $\mathbb{E}\big[A[G,G]\big] = \mathbb{E}\big[A[R,R]\big]$.

Equality (6.2) is a consequence of the Karhunen-Loeve expansion; see e.g., Adler and Taylor (2007, Chapter 3), that states that there is an orthonormal basis $(\varphi_k)_{k\geqslant 1}$ of $L^2(\mathcal{M};\mathbb{R})$ and independent one dimensional centered Gaussian random variables $(X_{ki})_{k\geqslant 1, 1\leqslant i\leqslant d}$ with $\mathrm{Var}(X_{ki}) = \lambda_{ki} > 0$ such that $G_i = \sum_{k\geqslant 1} X_{ki}\varphi_k$, and the convergence is in L^2. Since R is also L^2, we can expand $R_i = \sum_{k\geqslant 1} Y_{ki}\varphi_k$ with $Y_{ki} = \int R_i(x)\varphi_k(x)\mathrm{d}x$, where $\mathrm{d}x$ is the volume measure associated to $\mathcal{M}$, and the convergence is in L^2. Now, by linearity and that $\mathrm{Cov}(X_{ki}, X_{\ell j}) = \delta_{ij}\delta_{k\ell}\lambda_{ki}$, we find

$$\mathbb{E}\big[A[G,G]\big] = \sum_{i=1}^{d}\sum_{k\geqslant 1}\lambda_{ki}A[\mathbf{e}_i\varphi_{ki}, \mathbf{e}_i\varphi_{ki}],$$

where $\mathbf{e}_i$ is the d-dimensional vector with a one in the i^{th} position, and zero elsewhere.

To show that we obtain the same quantity with R replacing G, it is enough to show $\mathrm{Cov}(Y_{ki}, Y_{\ell j}) = \delta_{ij}\delta_{k\ell}\lambda_{ki}$. We use Mercer's theorem, which says that

$$\mathbb{E}[R_i(x)R_j(y)] = C_{ij}(x,y) = \delta_{ij}\sum_{m\geqslant 1}\lambda_{mi}\varphi_m(x)\varphi_m(y),$$

where the convergence in the sum is uniform, and we obtain

$$\begin{aligned}
\mathbb{E}[Y_{ki}Y_{\ell j}] &= \mathbb{E}\left[\iint R_i(x)R_j(y)\varphi_k(x)\varphi_\ell(y)\,\mathrm{d}x\,\mathrm{d}y\right] \\
&= \iint C_{ij}(x,y)\varphi_k(x)\varphi_\ell(y)\,\mathrm{d}x\,\mathrm{d}y = \delta_{ij}\sum_{m\geqslant 1}\lambda_{mi}\iint \varphi_m(x)\varphi_m(y)\varphi_k(x)\varphi_\ell(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \delta_{ij}\sum_{m\geqslant 1}\lambda_{mi}\delta_{mk}\delta_{ml} = \delta_{ij}\delta_{k\ell}\lambda_{ki},
\end{aligned}$$

as $\varphi_k, k\geqslant 1$ are orthonormal, thus proving claim (6.2).

Let the pair $(\widehat{W}, \widehat{H})$ be an independent copy of (W,H). Clearly, the right-hand side of (6.2) is the same for both pairs and hence

$$|\,\mathbb{E}\zeta(F) - \mathbb{E}\zeta(G)| = \big|\mathbb{E}\big[D^2\eta(W\sigma(H))[\widehat{W}\sigma(\widehat{H}), \widehat{W}\sigma(\widehat{H})] - D\eta(W\sigma(H))[W\sigma(H)]\big]\big|, \tag{6.3}$$

via (4.4) and independence. Hence, bounding the right-hand side of (6.3) yields a bound on the left-hand side of (4.4).

We first write

$$\widehat{W}\sigma(\widehat{H}) = \sum_{j=1}^{m}\widehat{V}_j \quad \text{where we set} \quad \widehat{V}_j \coloneqq \sum_{i=1}^{n}\widehat{W}_{ij}\sigma(\widehat{H}_j)\mathbf{e}_i,$$

and adopt parallel notation to define V_j. Because $\widehat{W}_{ij}$ are independent of each other and of W, centered, and assumed to have common variance c_w/m, for the first term in (6.3) we have

$$\mathbb{E}\big[D^2\eta(W\sigma(H))[\widehat{W}\sigma(\widehat{H}), \widehat{W}\sigma(\widehat{H})]\big] = \sum_{j=1}^{m}\mathbb{E}\Big\{D^2\eta(W\sigma(H))[\widehat{V}_j, \widehat{V}_j]\Big\}. \tag{6.4}$$

Working now on the second term of (6.3),

$$\big(W\sigma(H)\big)^j \coloneqq W\sigma(H) - V_j \quad \text{where} \quad V_j = \sum_{i=1}^{n}W_{ij}\sigma(H_j)\mathbf{e}_i, \tag{6.5}$$

and which is independent of $(W_{ij})_{i=1}^n$ and H_j. Using that independence to subtract a term with expectation zero in the second line below, followed by an application of a Taylor type argument, we have

$$\begin{aligned}
&\mathbb{E}\big[D\eta(W\sigma(H))[W\sigma(H)]\big] \\
&= \sum_{j=1}^m \mathbb{E}\Big\{ D\eta(W\sigma(H))[V_j] - D\eta\big((W\sigma(H))^j\big)[V_j] \Big\} \\
&= \sum_{j=1}^m \mathbb{E}\Big\{ D^2\eta\big((W\sigma(H))^j\big)[V_j, V_j] \Big\} \\
&\quad + \sum_{j=1}^m \int_0^1 \mathbb{E}\Big\{ D^2\eta\big(sW\sigma(H) + (1-s)(W\sigma(H))^j\big)[V_j, V_j] - D^2\eta\big((W\sigma(H))^j\big)[V_j, V_j] \Big\} \mathrm{Leb}(\mathrm{d}s) \\
&= \sum_{j=1}^m \mathbb{E}\Big\{ D^2\eta\big((W\sigma(H))^j\big)[\widehat{V}_j, \widehat{V}_j] \Big\} \qquad (6.6) \\
&\quad + \sum_{j=1}^m \int_0^1 \mathbb{E}\Big\{ D^2\eta\big(sW\sigma(H) + (1-s)(W\sigma(H))^j\big)[V_j, V_j] - D^2\eta\big((W\sigma(H))^j\big)[V_j, V_j] \Big\} \mathrm{Leb}(\mathrm{d}s). \qquad (6.7)
\end{aligned}$$

To bound (6.3), we first subtract this expression from (6.4) and, then bound the absolute value. In particular, we first bound the absolute difference between (6.4) and (6.6), and then the absolute value of (6.7). For the former, applying the second inequality of (4.2), which gives that the second derivative of η is Lipschitz, followed by Hölder's inequality, yields that this difference is bounded by

$$\begin{aligned}
&\sum_{j=1}^m \Big| \mathbb{E}\Big\{ D^2\eta\big((W\sigma(H))^j\big)[\widehat{V}_j, \widehat{V}_j] - \mathbb{E}D^2\eta\big(W\sigma(H)\big)[\widehat{V}_j, \widehat{V}_j] \Big\} \Big| \\
&\qquad\qquad \leqslant \frac{1}{3} \sum_{j=1}^m \mathbb{E}\Big[\big\| V_j \big\|_\infty \big\| \widehat{V}_j \big\|_\infty^2 \Big] \leqslant \frac{1}{3} \sum_{j=1}^m \mathbb{E}\Big[\big\| V_j \big\|_\infty^3 \Big]. \qquad (6.8)
\end{aligned}$$

Similarly, but more simply, the absolute value of (6.7) is bounded by one-half this same quantity.

To bound (6.8), we use the fact that H_j is independent from $W_{ij}, i = 1, \ldots, n$, and again apply Hölder's inequality, to find that

$$\begin{aligned}
\frac{1}{3} \sum_{j=1}^m \mathbb{E}\Big[\Big\| \sum_{i=1}^n W_{ij}\sigma(H_j)\mathbf{e}_i \Big\|_\infty^3 \Big] &\leqslant \frac{1}{3} \sum_{j=1}^m \mathbb{E}\big[\|\sigma(H_j)\|_\infty^3 \big] \, \mathbb{E}\Big[\Big\| \sum_{i=1}^n W_{ij}\mathbf{e}_i \Big\|^4 \Big]^{3/4} \\
&= \frac{1}{3} \sum_{j=1}^m \mathbb{E}\big[\|\sigma(H_j)\|_\infty^3 \big] \, \mathbb{E}\Big[\Big(\sum_{i=1}^n W_{ij}^2 \Big)^2 \Big]^{3/4} \\
&\leqslant \frac{m}{3} \, \mathbb{E}\big[\|\sigma(H_1)\|_\infty^3 \big] \Big(\frac{n^2 B c_w^2}{m^2} \Big)^{3/4} \\
&= \frac{1}{3} c_w^{3/2} B^{3/4} \, \mathbb{E}\big[\|\sigma(H_1)\|_\infty^3 \big] \frac{n^{3/2}}{\sqrt{m}},
\end{aligned}$$

where we have used that $\mathbb{E}\big[W_{ij}^4\big] \leqslant B(c_w/m)^2$. Hence, we obtain the desired inequality, (6.1). □

In Section 6.2, Lemma 6.1 is used to derive bounds on the difference between $G^{(\ell)}$ and $F^{(\ell)}$ in the smooth function metric for general $(\mathcal{M}, \mathtt{d})$. For informative bounds in the Wasserstein metric when $\mathcal{M} \equiv \mathcal{S}^n$, we apply Theorem 1.1. The following lemma gives some moment bounds that are used along with Lemma 5.3 to bound the terms $\mathbb{E}\|F - F_\varepsilon\|_\infty$ and $\mathbb{E}\|G - G_\varepsilon\|_\infty$ appearing in (1.2).

Lemma 6.2. *For fixed $p \in \mathbb{N}$, assume $H \in L^{2p}(\mathcal{M}; \mathbb{R}^m)$ is a random field with identically distributed coordinate processes, and let $W : \mathbb{R}^m \to \mathbb{R}$ be an $1 \times m$ random matrix that is independent of H and has centered independent entries satisfying $\mathbb{E}[W_{1j}^{2p}] \leqslant \tilde{c}/m^p$, for some $\tilde{c} > 0$. Letting $\sigma : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant Lip_σ, define $F : \mathcal{M} \to \mathbb{R}$ by*

$$F(x) = W\sigma\big(H(x)\big),$$

and finally, letting $A_m^{(2p)}$ be the set of $(j_1, \dots, j_{2p}) \in \{1, \dots, m\}^{2p}$ where the label of every coordinate appears at least twice, there is a constant c depending only on p and $\tilde{c}$ such that

$$\mathbb{E}\big[(F(x) - F(y))^{2p}\big] \leqslant \frac{\tilde{c}}{m^p} \sum_{(j_1,\dots,j_{2p}) \in A_m^{(2p)}} \prod_{\ell=1}^{2p} \mathbb{E}\Big[\Big(\sigma(H_{j_\ell}(x)) - \sigma(H_{j_\ell}(y))\Big)^{2p}\Big]^{1/(2p)} \tag{6.9}$$

$$\leqslant c\,\mathrm{Lip}_\sigma^{2p}\,\mathbb{E}\big[\big(H_1(x) - H_1(y)\big)^{2p}\big]. \tag{6.10}$$

Proof. For the first inequality, direct calculation gives

$$\begin{aligned}\mathbb{E}\big[(F(x) - F(y))^{2p}\big] &= \sum_{j_1,\dots,j_{2p}=1}^{n_1} \mathbb{E}\Big[\prod_{\ell=1}^{2p} W_{1,j_\ell}\Big]\,\mathbb{E}\Big[\prod_{\ell=1}^{2p}\Big(\sigma(H_{j_\ell}(x)) - \sigma(H_{j_\ell}(y))\Big)\Big] \\ &= \sum_{(j_1,\dots,j_{2p}) \in A_m^{(2p)}} \mathbb{E}\Big[\prod_{\ell=1}^{2p} W_{1,j_\ell}\Big]\,\mathbb{E}\Big[\prod_{\ell=1}^{2p}\Big(\sigma(H_{j_\ell}(x)) - \sigma(H_{j_\ell}(y))\Big)\Big],\end{aligned}$$

which follows since W_{1j} are independent and have mean zero. From this (6.9) easily follows by Hölder's inequality.

As H has identically distributed entries, we see that (6.9) satisfies,

$$\begin{aligned}&\frac{\tilde{c}}{m^p} \sum_{(j_1,\dots,j_{2p}) \in A_m^{(2p)}} \prod_{\ell=1}^{2p} \mathbb{E}\Big[\Big(\sigma(H_{j_\ell}(x)) - \sigma(H_{j_\ell}(y))\Big)^{2p}\Big]^{1/(2p)} \\ &\qquad = \frac{\tilde{c}}{m^p}\,\big|A_m^{(2p)}\big|\,\mathbb{E}\Big[\Big(\sigma(H_1(x)) - \sigma(H_1(y))\Big)^{2p}\Big], \\ &\qquad \leqslant c\,\mathbb{E}\Big[\Big(\sigma(H_1(x)) - \sigma(H_1(y))\Big)^{2p}\Big],\end{aligned}$$

where the last inequality follows because $\big|A_m^{(2p)}\big| = \mathrm{O}(m^p)$, with a constant depending only on p. The upper bound (6.10) now easily follows, since σ is Lipschitz. □

6.1 W_1 bounds for wide random neural networks: Proof of Theorem 1.2

Combining the previous results, we can now prove our main theorem for wide random neural networks.

Proof of Theorem 1.2. The proof proceeds by induction on $\ell = 2, \dots, L$ for the hypotheses that there is a constant c (which may change from line to line) depending only on $(c_w^{(m)}, c_b^{(m)}, B^{(m)})_{m=0}^{L}$, n, p and $\sigma(0)$ such that

$$\begin{aligned}&\mathtt{d}_{\mathcal{W}}\big(F^{(\ell)}, G^{(\ell)}\big)\\ &\leqslant c(1+\mathrm{Lip}_\sigma^3)^{\ell-1}\sum_{m=1}^{\ell-1}\left(n_{m+1}^{1/2}\left(\frac{n_{m+1}^4}{n_m}\right)^{(1-\frac{n}{p})/(8(1-\frac{n}{p})+6(n+\iota))}\log(n_m/n_{m+1}^4)\right)\prod_{j=m+1}^{\ell-1}\mathbb{E}\|W^{(j)}\|_{\mathrm{op}},\end{aligned} \tag{6.11}$$

and

$$\mathbb{E}\big[(G_i^{(\ell)}(x) - G_i^{(\ell)}(y))^{2p}\big] \leqslant c\,\mathrm{Lip}_\sigma^{2p(\ell-1)}\mathtt{d}(x,y)^{2p}, \;\; i = 1, \dots, n_\ell, \tag{6.12}$$

and finally

$$\mathbb{E}\big[\|\sigma(G_i^{(\ell)} - b_i^{(\ell)})\|_\infty^3\big] \leqslant c\,(1+\mathrm{Lip}_\sigma)^{3\ell}, \;\; i = 1, \dots, n_\ell. \tag{6.13}$$

We first note that the bias $b^{(\ell)}$ plays no role in the bound and can be set to zero. The reduction is obvious for (6.12) and (6.13), since we can write $G^{(\ell)} = \widetilde{G}^{(\ell)} + b^{(\ell)}$, with $\widetilde{G}^{(\ell)}$ a Gaussian process independent of $b^{(\ell)}$, having covariance $\widetilde{C}^{(\ell)}(x,y) = C^{(\ell)}(x,y) - \mathrm{I}_{n_2}c^{(\ell)}$. To see why we can also make this simplification for (6.11), assume that this inequality holds for $F^{(\ell)}$ and $G^{(\ell)}$ when the biases are zero. Define $\widetilde{F}^{(\ell)} = F^{(\ell)} + b^{(\ell)}$ and $\widetilde{G}^{(\ell)} = G^{(\ell)} + b^{(\ell)}$, where the summands are independent. For any Lipschitz $\zeta : \mathrm{C}(\mathcal{S}^n; \mathbb{R}^{n_2}) \to \mathbb{R}$ we have, by independence, that

$$\big|\mathbb{E}[\zeta(\widetilde{F}^{(\ell)})] - \mathbb{E}[\zeta(\widetilde{G}^{(\ell)})]\big| = \big|\mathbb{E}[\zeta(F^{(\ell)} + b^{(\ell)}) - \zeta(G^{(\ell)} + b^{(\ell)})]\big| = \big|\mathbb{E}[\widetilde{\zeta}(F^{(\ell)}) - \widetilde{\zeta}(G^{(\ell)})]\big|$$

where

$$\widetilde{\zeta}(f) = \mathbb{E}[\zeta(f + b^{(2)})],$$

which is 1-Lipschitz, since

$$\big|\widetilde{\zeta}(f) - \widetilde{\zeta}(g)\big| = \big|\mathbb{E}\big[\zeta(f + b^{(2)}) - \zeta(g + b^{(2)})\big]\big| \leqslant \|f - g\|_\infty.$$

Hence Wasserstein bounds in the case where the biases are non-zero are upper bounded by those in the zero bias case. Note that eliminating the biases $b^{(\ell)}$ in this manner requires them to be Gaussian, as otherwise the process $G^{(\ell)}$ may not be Gaussian.

We now begin the proof of the base case, $\ell = 2$. We first show (6.12), as well as some other related moment bounds used to show (6.11). We start by applying (6.10) from Lemma 6.2 with $W = W_{1,\cdot}^{(1)}, H = F^{(1)}$ and $m = n_1$, to find

$$\mathbb{E}\Big[\big(F_1^{(2)}(x) - F_1^{(2)}(y)\big)^{2p}\Big] \leqslant c\,\mathrm{Lip}_\sigma^{2p}\,\mathbb{E}\Big[\big(F_1^{(1)}(x) - F_1^{(1)}(y)\big)^{2p}\Big].$$

Applying (6.9) from Lemma 6.2 with $W = W_{1,\cdot}^{(0)}, H(x) = x$ and $m = n_0$, and σ there equal to the identity, we obtain

$$\begin{aligned}\mathbb{E}\Big[\big(F_1^{(1)}(x) - F_1^{(1)}(y)\big)^{2p}\Big] &\leqslant c\sum_{(j_1,\dots,j_{2p})\in A_{n_0}^{(2p)}}\prod_{\ell=1}^{2p}|x_{j_\ell} - y_{j_\ell}| \leqslant c\sum_{j_1,\dots,j_{2p}=1}^{n_0}\prod_{\ell=1}^{2p}|x_{j_\ell} - y_{j_\ell}|\\ &= c\Big(\sum_{j=1}^{n_0}|x_j - y_j|\Big)^{2p} \leqslant c\|x-y\|_2^{2p} \leqslant c\,\mathtt{d}(x,y)^{2p},\end{aligned}$$

where the last inequality holds as $\|x-y\|_2 \leqslant \mathtt{d}(x,y)$. Thus, we have shown

$$\mathbb{E}\Big[\big(F_1^{(2)}(x) - F_1^{(2)}(y)\big)^{2p}\Big] \leqslant c\,\mathrm{Lip}_\sigma^{2p}\mathtt{d}(x,y)^{2p}. \tag{6.14}$$

Letting the variance of $F_1^{(2)}(x) - F_2^{(2)}(y)$ be denoted τ^2, as the first and second moments of $G_1^{(2)}$ match those of $F_1^{(2)}$, we have in particular that $G_1^{(2)}(x) - G_2^{(2)}(y) \sim \mathcal{N}(0,\tau^2)$, and with $c_p = (2p-1)\times(2p-3)\ldots\times 3\times 1$, using Jensen's inequality and (6.14) we obtain

$$\begin{aligned} E\left[(G_1^{(2)}(x) - G_2^{(2)}(y))^{2p}\right] &= c_p\tau^{2p} \\ &= c_p\left(E[(F_1^{(2)}(x) - F_2^{(2)}(y))^2]\right)^p \leqslant c_p E[(F_1^{(2)}(x) - F_2^{(2)}(y))^{2p}] \leqslant c\,\mathrm{Lip}_\sigma^{2p}\mathtt{d}(x,y)^{2p}. \end{aligned} \tag{6.15}$$

The same inequalities holds for all indices $i = 1,\ldots,n_2$ since the coordinates all have the same distribution. Thus, we have established (6.12) for $\ell = 2$.

Now turning to (6.11), we bound $\|F^{(2)} - F_\varepsilon^{(2)}\|_\infty, \|G^{(2)} - G_\varepsilon^{(2)}\|_\infty$, and $\mathtt{d}_{\mathcal{F}}(F^{(2)}, G^{(2)})$ and then invoke Theorem 1.1. Using (6.14) and (6.15) in Lemma 5.3 applied to $F^{(2)}/\mathrm{Lip}_\sigma$ and $G^{(2)}/\mathrm{Lip}_\sigma$ implies that

$$\max\big\{\mathbb{E}\|F^{(2)} - F_\varepsilon^{(2)}\|_\infty, \mathbb{E}\|G^{(2)} - G_\varepsilon^{(2)}\|_\infty\big\} \leqslant c\,\mathrm{Lip}_\sigma\sqrt{n_2}\,\varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)}. \tag{6.16}$$

The right-hand side of inequality (2.12) of Theorem 2.9 with $k=2$ and $d=n_2$ gives an upper bound on the amount by which $\zeta_{\varepsilon,\delta}$ needs to be scaled in order to satisfy the second derivative condition in (1.1) and be an element of $\mathcal{F}$. Noting that $G^{(1)}$ is continuous with i.i.d. coordinate processes, we can apply Lemma 6.1 with $F = F^{(2)}$, $H = G^{(1)}$, $n = n_2$ and $m = n_1$ to find

$$\big|\mathbb{E}[\zeta_{\varepsilon,\delta}(F^{(2)})] - \mathbb{E}[\zeta_{\varepsilon,\delta}(G^{(2)})]\big| \leqslant c\,\delta^{-2}\varepsilon^{-2(n+\iota)}\big(c_w^{(1)}\big)^{3/2}\big(B^{(1)}\big)^{3/4}\,\mathbb{E}\big[\|\sigma(G_1^{(1)})\|_\infty^3\big]\,\frac{n_2^{5/2}}{\sqrt{n_1}}. \tag{6.17}$$

To bound $\mathbb{E}\big[\|\sigma(G_1^{(1)})\|_\infty^3\big]$, since σ is Lipschitz, for a fixed $y\in\mathcal{S}^n$ and any $x\in\mathcal{S}^n$, we have

$$\begin{aligned} \big|\sigma(G_1^{(1)}(x))\big| &\leqslant \big|\sigma(G_1^{(1)}(x)) - \sigma(G_1^{(1)}(y))\big| + |\sigma(G_1^{(1)}(y)) - \sigma(0)| + |\sigma(0)| \\ &\leqslant \mathrm{Lip}_\sigma\big(\omega_{G_1^{(1)}}(\pi) + |G_1^{(1)}(y)|\big) + |\sigma(0)|, \end{aligned}$$

where $\omega_{G_1^{(1)}}(\theta)$ denotes the modulus of continuity of $G_1^{(1)}$ at level θ; see Definition 5.1. Taking the supremum over x implies

$$\big\|\sigma(G_1^{(1)})\big\|_\infty \leqslant (\mathrm{Lip}_\sigma + 1)\big(\omega_{G_1^{(1)}}(\pi) + |G_1^{(1)}(y)| + |\sigma(0)|\big). \tag{6.18}$$

Because $G^{(1)}(y) = W^{(0)}y$, it is easy to see that

$$\mathbb{E}\Big[\big\|\sigma(G_1^{(1)})\big\|_\infty^3\Big] \leqslant (\mathrm{Lip}_\sigma + 1)^3. \tag{6.19}$$

Substituting this upper bound into (6.17) and combining with (6.16) in Theorem 1.1 implies

$$\mathtt{d}_{\mathcal{W}}(F^{(2)}, G^{(2)}) \leqslant c(\mathrm{Lip}_\sigma + 1)^3\sqrt{n_2}\Big(\varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)} + \delta + \delta^{-2}\varepsilon^{-2(n+\iota)}\frac{n_2^2}{\sqrt{n_1}}\Big).$$

Choosing

$$\delta = \varepsilon^{-\frac{2}{3}(n+\iota)}\left(\frac{n_2^4}{n_1}\right)^{1/6} \quad \text{and} \quad \varepsilon = \left(\frac{n_2^4}{n_1}\right)^{1/(3(1-\frac{n}{p})+4(n+\iota))}$$

we have shown that

$$\mathsf{d}_{\mathcal{W}}(F^{(2)}, G^{(2)}) \leqslant c(\mathrm{Lip}_\sigma + 1)^3\sqrt{n_2}\left(\frac{n_2^4}{n_1}\right)^{(1-\frac{n}{p})/(6(1-\frac{n}{p})+8(n+\iota))}\sqrt{\log(n_1/n_2^4)}.$$

For (6.13), in exactly the same way as (6.18), we have for any $y \in \mathcal{S}^n$,

$$\big|\sigma(G_1^{(2)}(x))\big| \leqslant (\mathrm{Lip}_\sigma + 1)\big(\omega_{G_1^{(2)}}(\pi) + |G_1^{(2)}(y)| + |\sigma(0)|\big). \tag{6.20}$$

But (6.12) and Proposition 5.2 together imply (scaling by Lip_σ) that $\mathbb{E}[\omega_{G_1^{(2)}}(\pi)^3] \leqslant c(\mathrm{Lip}_\sigma + 1)^3$. Because $G_1^{(2)}$ is Gaussian, we have that

$$\mathbb{E}\big[|G_1^{(2)}(y)|^3\big] = 2\sqrt{2/\pi}\,\mathrm{Var}(G_1^{(2)}(y))^{3/2},$$

and, by definition and using (6.19),

$$\mathrm{Var}(G_1^{(2)}(y)) = c_w^{(1)}\,\mathbb{E}\big[\sigma\big(G_1^{(1)}(y)\big)^2\big] + c_b^{(1)} \leqslant c(\mathrm{Lip}_\sigma + 1)^2.$$

Thus

$$\mathbb{E}\big[\|\sigma(G_1^{(2)})\|^3\big] \leqslant c(\mathrm{Lip}_\sigma + 1)^6,$$

and the base case is established.

For the induction step, assume (6.11), (6.12), and (6.13) for some $\ell \geqslant 2$; we show these three conditions are satisfied when ℓ is replaced by $\ell + 1$. For (6.12), we have from the definition of the covariance $C^{(\ell+1)}$ of $G^{(\ell+1)}$ that

$$\begin{aligned}\mathbb{E}\Big[\big(G_1^{(\ell+1)}(x) - G_1^{(\ell+1)}(y)\big)^2\Big] &= c\,\mathbb{E}\Big[\Big(\sigma\big(G_1^{(\ell)}(x)\big) - \sigma\big(G_1^{(\ell)}(y)\big)\Big)^2\Big] \\ &\leqslant c\mathrm{Lip}_\sigma^2\,\mathbb{E}\Big[\big(G_1^{(\ell)}(x) - G_1^{(\ell)}(y)\big)^2\Big] \\ &\leqslant c\,\mathrm{Lip}_\sigma^{2\ell}\mathsf{d}(x,y)^2,\end{aligned}$$

where the first inequality uses that σ is Lipschitz, and the second step the induction hypothesis. As $G^{(\ell+1)}$ is Gaussian, we now also have that

$$\mathbb{E}\Big[\big(G_1^{(\ell+1)}(x) - G_1^{(\ell+1)}(y)\big)^{2p}\Big] \leqslant c\,\mathrm{Lip}_\sigma^{2\ell p}\mathsf{d}(x,y)^{2p}, \tag{6.21}$$

thus advancing the induction hypothesis for (6.12).

Now turning to (6.11), we first define an intermediate random field

$$\widehat{F}^{(\ell+1)} := W^{(\ell)}\sigma\big(G^{(\ell)}\big), \tag{6.22}$$

where we take $G^{(\ell)}$ to be independent of $W^{(\ell)}$. By the triangle inequality, we have

$$\mathsf{d}_{\mathcal{W}}\big(F^{(\ell+1)}, G^{(\ell+1)}\big) \leqslant \mathsf{d}_{\mathcal{W}}\big(F^{(\ell+1)}, \widehat{F}^{(\ell+1)}\big) + \mathsf{d}_{\mathcal{W}}\big(\widehat{F}^{(\ell+1)}, G^{(\ell+1)}\big). \tag{6.23}$$

By definition, for the first term

$$\mathsf{d}_{\mathcal{W}}\big(F^{(\ell+1)}, \widehat{F}^{(\ell+1)}\big) = \mathsf{d}_{\mathcal{W}}\big(W^{(\ell)}\sigma(F^{(\ell)}), W^{(\ell)}\sigma(G^{(\ell)})\big). \tag{6.24}$$

The function $\widetilde{\zeta}(f) = \mathbb{E}\big[\zeta(W^{(\ell)}\sigma(f))\big]$ satisfies

$$\big|\widetilde{\zeta}(f) - \widetilde{\zeta}(g)\big| \leqslant \mathbb{E}\big[\|W^{(\ell)}\|_{\mathrm{op}}\big]\,\mathrm{Lip}_\sigma \|f-g\|_\infty,$$

and so the independence of $W^{(\ell)}$ from $F^{(\ell)}$ and $G^{(\ell)}$ implies (6.24) is upper bounded as

$$\mathsf{d}_{\mathcal{W}}\big(F^{(\ell+1)}, \widehat{F}^{(\ell+1)}\big) \leqslant \mathbb{E}\big[\|W^{(\ell)}\|_{\mathrm{op}}\big]\,\mathrm{Lip}_\sigma \mathsf{d}_{\mathcal{W}}\big(F^{(\ell)}, G^{(\ell)}\big). \tag{6.25}$$

Now working on the second term of (6.23), we apply Theorem 1.1 and bound $\|\widehat{F}^{(\ell+1)} - \widehat{F}_\varepsilon^{(\ell+1)}\|_\infty$, $\|G^{(\ell+1)} - G_\varepsilon^{(\ell+1)}\|_\infty$, and $\mathsf{d}_{\mathcal{F}}(F^{(\ell+1)}, G^{(\ell+1)})$. By (6.10) of Lemma 6.2 with $W = W_{1,\cdot}^{(\ell)}$, $H = G^{(\ell)}$ and $m = n_\ell$, we have

$$\mathbb{E}\Big[\big(F_1^{(\ell+1)}(x) - F_1^{(\ell+1)}(y)\big)^{2p}\Big] \leqslant c\,\mathrm{Lip}_\sigma^{2p}\,\mathbb{E}\Big[\big(G_1^{(\ell)}(x) - G_1^{(\ell)}(y)\big)^{2p}\Big] \leqslant c\,\mathrm{Lip}_\sigma^{2p\ell}\mathsf{d}(x,y)^{2p},$$

where the last inequality holds via the induction hypothesis (6.12). In conjunction with inequality (6.21) for $G^{(\ell+1)}$, Lemma 5.3 (applied after scaling by Lip_σ^ℓ) now implies that

$$\max\big\{\mathbb{E}\|F^{(\ell+1)} - F_\varepsilon^{(\ell+1)}\|_\infty, \mathbb{E}\|G^{(\ell+1)} - G_\varepsilon^{(\ell+1)}\|_\infty\big\} \leqslant c\,\mathrm{Lip}_\sigma^\ell \sqrt{n_{\ell+1}}\,\varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)}. \tag{6.26}$$

Now, Lemma 6.1 with $F = \widehat{F}^{(\ell+1)}$ and $H = G^{(\ell)}$, noting that $G^{(\ell)}$ is continuous with i.i.d. coordinate processes, implies

$$\begin{aligned}\mathsf{d}_{\mathcal{F}}\big(\widehat{F}^{(\ell+1)}, G^{(\ell+1)}\big) &\leqslant \big(c_w^{(\ell)}\big)^{3/2}\big(B^{(\ell)}\big)^{3/4}\,\mathbb{E}\big[\|\sigma(G_1^{(\ell)})\|^3\big]\frac{n_{\ell+1}^{3/2}}{\sqrt{n_\ell}} \\ &\leqslant (1+\mathrm{Lip}_\sigma)^{3\ell}\big(c_w^{(\ell)}\big)^{3/2}\big(B^{(\ell)}\big)^{3/4}\frac{n_{\ell+1}^{3/2}}{\sqrt{n_\ell}},\end{aligned}$$

where we have used the induction hypothesis (6.13) in the final inequality. Applying this inequality along with (6.26) in Theorem 1.1 yields

$$\mathsf{d}_{\mathcal{W}}(\widehat{F}^{(\ell+1)}, G^{(\ell+1)}) \leqslant c(1+\mathrm{Lip}_\sigma)^{3\ell}\sqrt{n_{\ell+1}}\bigg(\delta^{-2}\varepsilon^{-2(n+\iota)}\frac{n_{\ell+1}^2}{\sqrt{n_\ell}} + \varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)} + \delta\bigg), \tag{6.27}$$

and choosing

$$\delta = \varepsilon^{-\frac{2}{3}(n+\iota)}\bigg(\frac{n_{\ell+1}^4}{n_\ell}\bigg)^{1/6} \quad \text{and} \quad \varepsilon = \bigg(\frac{n_{\ell+1}^4}{n_\ell}\bigg)^{1/(3(1-\frac{n}{p})+4(n+\iota))}$$

gives

$$\mathsf{d}_{\mathcal{W}}(\widehat{F}^{(\ell+1)}, G^{(\ell+1)}) \leqslant c(1+\mathrm{Lip}_\sigma)^{3\ell}\sqrt{n_{\ell+1}}\bigg(\frac{n_{\ell+1}^4}{n_\ell}\bigg)^{(1-\frac{n}{p})/(6(1-\frac{n}{p})+8(n+\iota))}\sqrt{\log(n_\ell/n_{\ell+1}^4)}.$$

Using this bound and (6.25) in (6.23), and applying the induction hypothesis (6.11) advances the induction for (6.11).

Finally, advancing the induction for (6.13), i.e., bounding $\mathbb{E}\big[\|\sigma(G_1^{(\ell+1)})\|^3\big] \leqslant c\,(\mathrm{Lip}_\sigma + 1)^{3(\ell+1)}$, follows in exactly the same way as for the base case, starting at (6.20). □

6.2 Improved W_1 bounds: Proof of Theorem 1.4

This subsection proves Theorem 1.4, showing the rate improvement under the additional assumption that σ has three bounded derivatives. The rate improvement illustrated in Remark 1.5 comes from the fact that for the induction steps, we work with the smooth metric $\mathsf{d}_{\mathcal{F}}$, and only smooth at the final layer, rather than at each layer of the induction in the $\mathsf{d}_{\mathcal{W}}$ metric; compare (6.27) and (6.28).

Theorem 6.3. *Assume that σ has three bounded derivatives, and let the weights satisfy the moment condition in (1.3). Recalling the definition of β_L from (1.5), for any $L \geqslant 2$, there exists a positive constant c, depending only on $(c_w^{(\ell)}, c_b^{(\ell)}, B^{(\ell)})_{\ell=0}^{L}, n, p$, and $\|\sigma^{(k)}\|_\infty$, $k = 1, 2, 3$, such that $\mathsf{d}_{\mathcal{F}}(F^{(L)}, G^{(L)}) \leqslant c\,\beta_L$.*

Proof. The proof follows by an induction similar to that in the proof of Theorem 1.2. For the base case $L = 2$, first note we can again set $b^{(2)} = 0$, since if $\zeta \in \mathcal{F}$, then straightforward considerations imply $\widetilde{\zeta}(f) := \mathbb{E}[\zeta(f + b^{(2)})] \in \mathcal{F}$. Thus, for $\widetilde{G}^{(2)} - b^{(2)}$ and $\widetilde{F}^{(2)} = F^{(2)} - b^{(2)}$ we have

$$\big|\mathbb{E}[\zeta(F^{(2)})] - \mathbb{E}\big[\zeta(G^{(2)})\big]\big| = \big|\mathbb{E}[\widetilde{\zeta}(\widetilde{F}^{(2)})] - \mathbb{E}\big[\widetilde{\zeta}(\widetilde{G}^{(2)})\big]\big|,$$

and so it is enough to bound the right-hand side for generic $\widetilde{\zeta} \in \mathcal{F}$. With this simplification, we can apply Lemma 6.1 with $m = n_1$ and $n = n_2$ to find

$$\mathsf{d}_{\mathcal{F}}\big(F^{(2)}, G^{(2)}\big) \leqslant (c_w^{(1)})^{3/2} (B^{(1)})^{3/4}\, \mathbb{E}\big[\|\sigma(G_1^{(1)})\|_\infty^3\big] \frac{n_2^{3/2}}{\sqrt{n_1}} \leqslant c \frac{n_2^{3/2}}{\sqrt{n_1}},$$

where the last inequality follows from (6.19), which states $\mathbb{E}\big[\|\sigma(G_1^{(1)})\|_\infty^3\big] \leqslant c$.

To advance the induction, assume the bound on $\mathsf{d}_{\mathcal{F}}(F^{(\ell)}, G^{(\ell)})$. In exactly the same way as above, we can assume $b^{(\ell)} = 0$. Now, recall that in (6.22), we defined the intermediate random field

$$\widehat{F}^{(\ell+1)} := W^{(\ell)} \sigma\big(G^{(\ell)}\big),$$

where $G^{(\ell)}$ is independent of $W^{(\ell)}$. The triangle inequality, as before, yields

$$\mathsf{d}_{\mathcal{F}}\big(F^{(\ell+1)}, G^{(\ell+1)})\big) \leqslant \mathsf{d}_{\mathcal{F}}\big(F^{(\ell+1)}, \widehat{F}^{(\ell+1)}\big) + \mathsf{d}_{\mathcal{F}}\big(\widehat{F}^{(\ell+1)}, G^{(\ell+1)}\big).$$

and we again define the function $\widetilde{\zeta}(f) = \mathbb{E}\big[\zeta(W^{(\ell)}\sigma(f))\big]$. We need to argue that up to a constant factor, $\widetilde{\zeta} \in \mathcal{F}$. Starting with the first derivative and denoting component-wise (Hadamard) multiplication by $\circ$, we first have

$$\begin{aligned}\big|\widetilde{\zeta}(f+g) - \mathbb{E}\big[\zeta(W^{(\ell)}\big(\sigma(f) + \sigma'(f)\circ g\big)\big]\big| &\leqslant \sup_h \|D\zeta(h)\|\, \mathbb{E}\big\|W^{(\ell)}\big(\sigma(f+g) - \sigma(f) - \sigma'(f)\circ g\big)\big\| \\ &\leqslant \sup_h \|D\zeta(h)\|\, \mathbb{E}\|W^{(\ell)}\|_{\mathrm{op}} \|\sigma''\|_\infty \|g\|_\infty^2.\end{aligned}$$

Combining the above display with a direct Taylor-like computation, we next have that

$$\begin{aligned}&\widetilde{\zeta}(f+g) - \widetilde{\zeta}(f) \\ &\quad= \mathbb{E}\int_{[0,1]^2} D^2\zeta\big(W^{(\ell)}(\sigma(f) + s\,t\,\sigma'(f)\circ g)\big)\big[W^{(\ell)}\big(\sigma'(f)\circ g\big), W^{(\ell)}\big(\sigma'(f)\circ g\big)\big]\mathrm{Leb}(\mathrm{d}s, \mathrm{d}t) \\ &\qquad+ \mathbb{E}\Big[D\zeta(W^{(\ell)}\sigma(f))\big[W^{(\ell)}\big(\sigma'(f)\circ g\big)\big]\Big] + \mathrm{O}\big(\|g\|_\infty^2\big) \\ &\quad= \mathbb{E}\Big[D\zeta(W^{(\ell)}\sigma(f))\big[W^{(\ell)}\big(\sigma'(f)\circ g\big)\big]\Big] + \mathrm{O}\big(\|g\|_\infty^2\big),\end{aligned}$$

so that

$$D\widetilde{\zeta}(f)[g] = \mathbb{E}\Big[D\zeta(W^{(\ell)}\sigma(f))\big[W^{(\ell)}\big(\sigma'(f)\circ g\big)\big]\Big].$$

Since $\sup_h \|D\zeta(h)\| \leqslant 1$, it follows that

$$\sup_f \|D\widetilde{\zeta}(f)\| \leqslant \|\sigma'\|_\infty \,\mathbb{E}\|W^{(\ell)}\|_{\mathrm{op}}.$$

Similar but more onerous computations show

$$\begin{aligned} D^2\widetilde{\zeta}(f)[g^{(1)},g^{(2)}] = \mathbb{E}\Big[D^2\zeta(W^{(\ell)}\sigma(f))\big[W^{(\ell)}\big(\sigma'(f)\circ g^{(1)}\big), W^{(\ell)}\big(\sigma'(f)\circ g^{(2)}\big)\big]\Big] \\ + \mathbb{E}\Big[D\zeta(W^{(\ell)}\sigma(f))\big[W^{(\ell)}\big(\sigma''(f)\circ g^{(1)}\circ g^{(2)}\big)\big]\Big], \end{aligned}$$

so that

$$\sup_f \|D^2\widetilde{\zeta}(f)\| \leqslant \|\sigma'\|_\infty^2 \,\mathbb{E}\Big[\|W^{(\ell)}\|_{\mathrm{op}}^2\Big] + \|\sigma''\|_\infty \,\mathbb{E}\|W^{(\ell)}\|_{\mathrm{op}} < \infty.$$

Finally, some straightforward but space-consuming manipulations, using in particular that

$$|D^2\zeta(h)[g^{(1)},g^{(2)}]| \leqslant 3\|g^{(1)}\|_\infty \|g^{(2)}\|_\infty \|D^2\zeta(h)\|,$$

from Barbour et al. (2023, Lemma 2.4), imply that

$$\frac{\big\|D^2\widetilde{\zeta}(h) - D^2\widetilde{\zeta}(f)\big\|}{\|f-h\|} \leqslant c\,\max\big\{1, \mathbb{E}\big[\|W^{(\ell)}\|_{\mathrm{op}}^3\big]\big\}.$$

Hence, using the independence of $W^{(\ell)}$ with $F^{(\ell)}$ and $G^{(\ell)}$ we have

$$\mathsf{d}_{\mathcal{F}}\big(F^{(\ell+1)}, \widehat{F}^{(\ell+1)}\big) \leqslant c\,\max\big\{1, \mathbb{E}\big[\|W^{(\ell)}\|_{\mathrm{op}}^3\big]\big\}\mathsf{d}_{\mathcal{F}}\big(F^{(\ell)}, G^{(\ell)}\big),$$

and the proof now follows as that for Theorem 1.2, mutatis mutandis. □

We are now ready to prove Theorem 1.4, using Theorem 1.1. Compared to the proof of Theorem 1.2, the specific choice of the smoothing and regularization terms, ε and δ are different, resulting in the required rate improvement.

Proof of Theorem 1.4. We apply Theorem 1.1, with $F = F^{(L)}$ and $W = G^{(L)}$, and hence with $d = n_L$. Applying Lemma 5.3, using induction with (6.10) and (6.14), we have that

$$\mathbb{E}\|F^{(L)} - F_\varepsilon^{(L)}\|_\infty \leqslant c\,\sqrt{n_L}\,\varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)},$$

and the same bound also holds for $\mathbb{E}\|G^{(L)} - G_\varepsilon^{(L)}\|_\infty$. From Theorem 6.3, we have

$$\mathsf{d}_{\mathcal{F}}(F^{(L)}, G^{(L)}) \leqslant c\,\beta_L.$$

Putting everything together, we have

$$\mathsf{d}_{\mathcal{W}}(F^{(L)}, G^{(L)}) \leqslant c\sqrt{n_L}\left(\sqrt{n_L}\,\beta_L\,\delta^{-2}\varepsilon^{-2(n+\iota)} + \varepsilon^{\frac{1}{2}(1-\frac{n}{p})}\sqrt{\log(1/\varepsilon)} + \delta\right). \tag{6.28}$$

Picking ε and δ as

$$\delta = \varepsilon^{-2(n+\iota)/3}(n_L\beta_L^2)^{1/6} \qquad \varepsilon = (n_L\beta_L^2)^{1/(3(1-\frac{n}{p})+4(n+\iota))},$$

we obtain the desired result. □

REFERENCES

E. Abbe, E. B. Adsera, and T. Misiakiewicz. The merged-staircase property: A necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022. (Cited on page 6.)

R. J. Adler and J. E. Taylor. *Random fields and geometry*, volume 80. Springer, 2007. (Cited on pages 13 and 26.)

B. Arras and C. Houdré. On Stein's method for multivariate self-decomposable laws. *Electron. J. Probab.*, 2019. (Cited on page 10.)

B. Arras and C. Houdré. On some operators associated with non-degenerate symmetric α-stable probability measures. *Potential Anal.*, pages 1–52, 2022. (Cited on page 10.)

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022. (Cited on page 6.)

Y. Bahri and B. Hanin. Les Houches Lectures on Deep Learning at Large & Infinite Width. *arXiv preprint arXiv:2309.01592*, 2023. (Cited on page 9.)

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014. (Cited on page 13.)

K. Balasubramanian, P. Ghosal, and Y. He. High-dimensional scaling limits and fluctuations of online least-squares SGD with smooth covariance. *arXiv preprint arXiv:2304.00707*, 2023. (Cited on page 2.)

A. D. Barbour. Stein's method for diffusion approximations. *Probab. Theory Relat. Fields*, 84(3): 297–322, 1990. (Cited on pages 5, 20, and 21.)

A. D. Barbour, N. Ross, and G. Zheng. Stein's method, Gaussian processes and Palm measures, with applications to queueing. *Ann. Appl. Probab.*, 33(5), 2023. (Cited on pages 5, 18, 19, 20, 21, and 34.)

A. D. Barbour, N. Ross, and G. Zheng. Stein's method, smoothing and functional approximation. *Electron. J. Probab.*, 29:Paper No. 20, 29 pp, 2024. (Cited on pages 2, 4, 5, 8, 15, 16, 17, and 21.)

A. Basteri and D. Trevisan. Quantitative Gaussian approximation of randomly initialized deep neural networks. *arXiv preprint arXiv:2203.07379*, 2022. (Cited on pages 3, 9, and 10.)

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012. (Cited on page 2.)

A. Bordino, S. Favaro, and S. Fortini. Infinitely wide limits for deep stable neural networks: sub-linear, linear and super-linear activation functions. *Transactions on Machine Learning Research*, 2023a. (Cited on page 10.)

A. Bordino, S. Favaro, and S. Fortini. Non-asymptotic approximations of Gaussian neural networks via second-order Poincaré inequalities. *arXiv preprint arXiv:2304.04010*, 2023b. (Cited on pages 3 and 9.)

S. Bourguin and S. Campese. Approximation of Hilbert-valued Gaussians on Dirichlet structures. *Electron. J. Probab.*, 25:30, 2020. (Cited on page 21.)

D. Burago, S. Ivanov, and Y. Kurylev. Spectral stability of metric-measure Laplacians. *Isr. J. Math.*, 232(1):125–158, 2019. (Cited on page 6.)

S. Chatterjee. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Relat. Fields*, 143(1-2):1–40, 2009. (Cited on page 9.)

L. H. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*. Springer, 2011. (Cited on page 2.)

P. Chen, I. Nourdin, L. Xu, and X. Yang. Multivariate stable approximation by Stein's Method. *J. Theor. Probab.*, pages 1–43, 2023. (Cited on page 10.)

Z. Chen, G. Rotskoff, J. Bruna, and E. Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. (Cited on pages 2 and 6.)

L. Coutin and L. Decreusefond. Stein's method for Brownian approximations. *Communications on Stochastic Analysis*, 7(3):1, 2013. (Cited on page 21.)

L. Coutin and L. Decreusefond. Stein's method for rough paths. *Potential Anal.*, 53(2):387–406, 2020. (Cited on page 21.)

F. Dai and Y. Xu. *Approximation theory and harmonic analysis on spheres and balls*, volume 23. Springer, 2013. (Cited on page 11.)

A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022. (Cited on page 6.)

A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. (Cited on pages 2, 3, and 9.)

R. Der and D. Lee. Beyond Gaussian processes: On the distributions of infinite networks. In *Advances in Neural Information Processing Systems*, volume 18, 2005. (Cited on page 10.)

C. Dobler and M. J. Kasprzak. Stein's method of exchangeable pairs in multivariate functional approximations. *Electron. J. Probab.*, 26:1–50, 2021. (Cited on page 21.)

R. M. Dudley. *Real analysis and probability*. CRC Press, 2018. (Cited on page 4.)

R. Eldan, D. Mikulincer, and T. Schramm. Non-asymptotic approximations of neural networks by Gaussian processes. In *Conference on Learning Theory*, pages 1754–1775. PMLR, 2021. (Cited on pages 3 and 9.)

S. Favaro, S. Fortini, and S. Peluchetti. Deep stable neural networks: Large-width asymptotics and convergence rates. *Bernoulli*, 29(3):2574–2597, 2023a. (Cited on page 10.)

S. Favaro, B. Hanin, D. Marinucci, I. Nourdin, and G. Peccati. Quantitative CLTs in deep neural networks. *arXiv preprint arXiv:2307.06092*, 2023b. (Cited on page 9.)

X. Fernique. Intégrabilité des vecteurs Gaussiens. *CR Acad. Sci. Paris Serie A*, 270:1698–1699, 1970. (Cited on page 15.)

V. Fortuin, A. Garriga-Alonso, S. W. Ober, F. Wenzel, G. Ratsch, R. E. Turner, M. van der Wilk, and L. Aitchison. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022. (Cited on page 10.)

H. L. Gan and N. Ross. Stein's method for the Poisson-Dirichlet distribution and the Ewens Sampling Formula, with applications to Wright-Fisher models. *Ann. Appl. Probab.*, 31(2):625 – 667, 2021. (Cited on page 21.)

E. Golikov and G. Yang. Non-Gaussian tensor programs. In *Advances in Neural Information Processing Systems*, volume 35, 2022. (Cited on page 6.)

D. Grieser. Uniform bounds for eigenfunctions of the Laplacian on manifolds with boundary. *Comm. Partial Differential Equations*, 27(7-8):1283–1299, 2002. (Cited on page 5.)

A. Grigor'yan. *Heat kernel and analysis on manifolds*, volume 47 of *AMS/IP Studies in Advanced Mathematics*. American Mathematical Society, Providence, RI; International Press, Boston, MA, 2009. (Cited on page 5.)

B. Hanin. Correlation functions in random fully connected neural networks at finite width. *arXiv preprint arXiv:2204.01058v1*, 2022. (Cited on page 10.)

B. Hanin. Random neural networks in the infinite width limit as Gaussian processes. *Ann. Appl. Probab.*, 33(6A):4798–4819, 2023. (Cited on pages 2, 3, 8, and 10.)

P. Jung, H. Lee, J. Lee, and H. Yang. α-stable convergence of heavy-/light-tailed infinitely wide neural networks. *Adv. in Appl. Probab.*, 55(4):1415–1441, 2023. (Cited on page 10.)

S. Kakutani. On equivalence of infinite product measures. *Ann. Math.*, pages 214–224, 1948. (Cited on page 13.)

M. J. Kasprzak. Stein's method for multivariate Brownian approximations of sums under dependence. *Stochastic Processes Appl.*, 130(8):4927–4967, 2020a. (Cited on page 21.)

M. J. Kasprzak. Functional approximations via Stein's method of exchangeable pairs. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(4):2540–2564, 2020b. (Cited on page 21.)

M. J. Kasprzak, A. B. Duncan, and S. J. Vollmer. Note on A. Barbour's paper on Stein's method for diffusion approximations. *Electron. Commun. Probab.*, 22, 2017. (Cited on page 20.)

A. Klukowski. Rate of convergence of polynomial networks to Gaussian processes. In *Conference on Learning Theory*, pages 701–722. PMLR, 2022. (Cited on pages 3 and 9.)

H. Lee, F. Ayed, P. Jung, J. Lee, H. Yang, and F. Caron. Deep neural networks with dependent weights: Gaussian process mixture limit, heavy tails, sparsity and compressibility. *Journal of Machine Learning Research*, 24(289):1–78, 2023. (Cited on page 10.)

J. Lee, J. Sohl Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. (Cited on pages 2, 3, and 9.)

M. Li, M. Nica, and D. Roy. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. (Cited on page 3.)

R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1996. (Cited on pages 2 and 10.)

I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus: From Stein's method to universality*, volume 192. Cambridge University Press, 2012. (Cited on page 2.)

A. Nowak, P. Sjögren, and T. Z. Szarek. Sharp estimates of the spherical heat kernel. *J. Math. Pures Appl.*, 129:23–33, 2019. (Cited on page 24.)

D. Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 1984. (Cited on page 22.)

M. Raič. A multivariate central limit theorem for Lipschitz and smooth test functions. *arXiv preprint arXiv:1812.08268*, 2018. (Cited on page 5.)

N. Ross. Fundamentals of Stein's method. *Probab. Surv.*, 8:210–293, 2011. (Cited on page 2.)

G. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Commun. Pure Appl. Math.*, 75(9):1889–1935, 2022. (Cited on pages 2 and 6.)

H.-H. Shih. On Stein's method for infinite-dimensional Gaussian approximation in abstract Wiener spaces. *J. Funct. Anal.*, 261(5):1236–1283, 2011. (Cited on page 21.)

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes Appl.*, 130(3):1820–1852, 2020. (Cited on pages 2 and 6.)

K.-T. Sturm. Diffusion processes and heat kernels on metric spaces. *Ann. Probab.*, 26(1):1–55, 1998. (Cited on page 6.)

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. (Cited on pages 7, 8, and 25.)

A. Vidotto. An improved second-order Poincaré inequality for functionals of Gaussian fields. *J. Theor. Probab.*, 33(1):396–427, 2020. (Cited on page 9.)

L. Xu. Approximation of stable law in Wasserstein-1 distance by Stein's method. *Ann. Appl. Probab*, 29(1):458–504, 2019. (Cited on page 10.)

G. Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 32, 2019. (Cited on page 2.)

S. Zelditch. *Eigenfunctions of the Laplacian on a Riemannian manifold*, volume 125. American Mathematical Soc., 2017. (Cited on page 5.)