Proceedings of Machine Learning Research vol 125:1–54, 2020 33rd Annual Conference on Learning Theory

A Corrective View of Neural Networks: Representation, Memorization and Learning

Guy Bresler GUY@MIT.EDU
Department of EECS, MIT
Cambridge, MA, USA.

Dheeraj Nagaraj DHEERAJ@MIT.EDU
Department of EECS, MIT
Cambridge, MA, USA.

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We develop a *corrective mechanism* for neural network approximation: the total available non-linear units are divided into multiple groups and the first group approximates the function under consideration, the second approximates the error in approximation produced by the first group and corrects it, the third group approximates the error produced by the first and second groups together and so on. This technique yields several new representation and learning results for neural networks:

1. Two-layer neural networks in the random features regime (RF) can memorize arbitrary labels for n arbitrary points in $\mathbb{R}^d$ with $\tilde{O}(\frac{n}{\theta^4})$ ReLUs, where θ is the minimum distance between two different points. This bound can be shown to be optimal in n up to logarithmic factors.

2. Two-layer neural networks with ReLUs and smoothed ReLUs can represent functions with an error of at most ϵ with $O(C(a,d)\epsilon^{-1/(a+1)})$ units for $a \in \mathbb{N} \cup \{0\}$ when the function has $\Theta(ad)$ bounded derivatives. In certain cases d can be replaced with effective dimension $q \ll d$. Our results indicate that neural networks with only a single nonlinear layer are surprisingly powerful with regards to representation, and show that in contrast to what is suggested in recent work, depth is not needed in order to represent highly smooth functions.

3. Gradient Descent on the recombination weights of a two-layer random features network with ReLUs and smoothed ReLUs can learn low degree polynomials up to squared error ϵ with subpoly$(1/\epsilon)$ units. Even though deep networks can approximate these polynomials with polylog$(1/\epsilon)$ units, existing *learning* bounds for this problem require poly$(1/\epsilon)$ units. To the best of our knowledge, our results give the first sub-polynomial learning guarantees for this problem.

1. Introduction

Neural networks have been shown to be very powerful in various classification and regression tasks Goodfellow et al. (2016). A lot of the properties of multi-layer networks remain unexplained rigorously, despite their success in practice. In this paper we focus on three core questions regarding the capabilities of neural networks: representation, memorization, and learning low degree polynomials.

Representation. Neural networks are universal approximators for continuous functions over compact sets and hence, when trained appropriately can solve a variety of machine learning problems Cybenko (1989); Hornik et al. (1989); Funahashi (1989); Lu et al. (2017); Hanin and Sellke (2017). A long line of work, starting with Barron (1993), provides bounds on the number of activation functions required for two-layer neural networks to achieve a given error when the function being approximated satisfies certain smoothness conditions Klusowski and Barron (2018); Ma et al. (2019); Liang and Srikant (2016); Safran and Shamir (2017); Yarotsky (2017); Li et al. (2019). The papers Barron (1993) and Klusowski and Barron (2018) use a law of large numbers based argument using random neural networks (see Section 1.1) to achieve a squared error of $1/N$ using N neurons, whereas other works including Liang and Srikant (2016); Safran and Shamir (2017); Yarotsky (2017); Li et al. (2019) carry out a Taylor series approximation for the target function by implementing additions and multiplications using deep networks. These assume more smoothness (higher number of bounded derivatives) of f and give faster than $1/N$ rates for the squared error.

Deep neural networks are practically observed to be better approximators than shallow two-layer networks. Depth separation results construct functions that are easily and efficiently approximated by deep networks but cannot be approximated by shallower networks unless their width is very large (see Safran and Shamir (2017); Daniely (2017a); Delalleau and Bengio (2011); Telgarsky (2016) and references therein). While the results in Liang and Srikant (2016); Safran and Shamir (2017); Yarotsky (2017); Li et al. (2019) consider deep architectures to achieve faster representation results for a class of smooth functions, it remained unclear whether or not the class of functions they consider can be similarly represented by shallow networks. Recent work Bresler and Nagaraj (2020) gives sharp representation results for arbitrary depth networks which show that deeper networks are better at representing less smooth functions.

In this work, we show similar representation results to those achieved in Yarotsky (2017) using deep networks, but for a two-layer neural network. Crucial to our approach is a careful choice of activation functions which are the same as ReLU activation functions outside of a small neighborhood of zero and they are smoother near zero. We note that the Sobolev space assumption for the target function in Yarotsky (2017) is essentially the same as our assumption of fast enough decay in their Fourier transform (see Section 3) due to the relationship between smoothness of a function and the decay of its Fourier transform. The experiments in Zheng et al. (2015) and Elfwing et al. (2018) suggest that considering smoothed activation functions in some layers along with ReLU in some others can in fact give measurably better results in various problems. Theoretical results in Li et al. (2019) show that smooth functions can be more efficiently represented using rectified power units (RePU), which are smoother than ReLU.

Despite the guarantees given by representation results, in practice finding the optimal parameters for a neural network for a given problem involves large-scale non-convex optimization, which is in general very hard. Therefore, stating representation results in conjunction with training guarantees is important, and as described next, we do so in the context of the memorization and learning low-degree polynomials.

Memorization. Neural networks have the property that they can memorize (or interpolate) random labels quite easily Zhang et al. (2016); Belkin et al. (2018). In practice, neural

networks are trained using SGD and a long line of papers aims to understand memorization in over-parametrized networks via the study of SGD/GD (see Du et al. (2019); Allen-Zhu et al. (2018); Jacot et al. (2018) and references therein). A recent line of work studies the problem of memorization of arbitrary labels on n arbitrary data points and provides polynomial guarantees (polynomial in n) for the number of non-linear units required (see Zou et al. (2018); Zou and Gu (2019); Oymak and Soltanolkotabi (2019); Song and Yang (2019); Ji and Telgarsky (2019); Panigrahi et al. (2019) and references therein). These polynomials often have high degree ($O(n^{30})$ in Allen-Zhu et al. (2018) and $O(n^6)$ as in Du et al. (2019)). Oymak and Soltanolkotabi (2019) and Song and Yang (2019) improve this to $O(n^2)$ under stronger assumptions on the data. Moreover, the bounds in Du et al. (2019), Oymak and Soltanolkotabi (2019) and Song and Yang (2019) contain data and possibly dimension dependent condition number factors. Panigrahi et al. (2019) obtains intelligible bounds for such condition number factors for various kinds of activation functions, but do not improve upon the $O(n^6)$ upper bound. Ji and Telgarsky (2019); Chen et al. (2019) show a polylogarithmic bound on the number of non-linear units required for memorization, but only under the condition of NTK separability.

We consider the problem of memorization of arbitrary labels via gradient descent for arbitrary d dimensional data points under the assumption that any two of these points are separated by a Euclidean distance of at least θ. Under the distance condition which we use here, the results of Ji and Telgarsky (2019) still require $O(n^{12}/\theta^4)$ non-linear units. Our results obtain a dependence of $\tilde{O}(n/\theta^4)$ for two-layer ReLU networks. This is optimal in n up to log factors. A similar bound is shown in Kawaguchi and Huang (2019), but with additional polynomial dependence on the dimension. Under additional distributional assumptions on the data, Daniely (2019) shows the optimal bound of $O(n/d)$ whenever n is polynomially large in d. Subsequent to the present paper's appearance on arXiv, Bubeck et al. (2020) used a similar iterative corrective procedure as proposed in this paper to address the question of memorizing n points with the smallest possible total *weight* rather than number of units. Our memorization results also achieve the optimal dependence for weight in terms of number of points n, with a better dependence on the error ϵ and with fewer assumptions on the data, but a worse dependence on the dimension d.

Learning Low Degree Polynomials. An important toy problem studied in the neural networks literature is that of learning degree q polynomials with d variables via SGD/GD when $q \ll d$. This problem was first considered in Andoni et al. (2014), and they showed that a two-layer neural network can be trained via Gradient Descent to achieve an error of at most ϵ whenever the number of non-linear units is $\Omega(d^{2q}/\epsilon^2)$ and Yehudai and Shamir (2019) gives a bound of $\Omega(d^{q^2}/\epsilon^4)$ using the random features model. All the currently known results for learning polynomials with SGD/GD require $\Omega\left(d^{2q}\text{poly}(1/\epsilon)\right)$ non-linear units.

There are several representation theorems for low-degree polynomials with deep networks where the depth depends on the error ϵ (see Liang and Srikant (2016); Safran and Shamir (2017); Yarotsky (2017)) by systematically implementing addition and multiplication. They require a total of $O(d^q\text{polylog}(1/\epsilon))$ non-linear units. However, there are no training guarantees for these deep networks via any algorithm. We show that a two-layer neural network with $O(\text{subpoly}(1/\epsilon))$ activation functions trained via GD/SGD suffices. In particular, the number of non-linear units we require is $O\left(C(a,q)d^{2q}\epsilon^{-\frac{1}{a+1}}\right)$ for arbitrary

$a \in \mathbb{N} \cup \{0\}$, which is subpolynomial in ϵ when we take $a \to \infty$ slowly enough as $\epsilon \to 0$. To the best of our knowledge, these are the first subpolynomial bounds for learning low-degree polynomials via neural networks trained with SGD.

1.1. The Corrective Mechanism

We now describe the main theoretical tool developed in this work. Let $a, N \in \mathbb{N}$. With aN non-linear units in total, under appropriate smoothness conditions on the function $f : \mathbb{R}^d \to \mathbb{R}$ being approximated, we describe a way to achieve a squared error of $O(1/N^a)$. The same basic methodology is used, with suitable modifications, to prove all of our results.

For any activation function σ, the construction given in Barron (1993) obtains $O(1/N)$ error guarantees for a two-layer network by picking $\Theta_1, \ldots, \Theta_N$ i.i.d. from an appropriate distribution such that $\mathbb{E}\sigma(x; \Theta_1) \approx f(x)$ for every x in some bounded domain. Then, the empirical sum $\hat{f}^{(1)}(x) := \frac{1}{N}\sum_{i=1}^{N} \sigma(x; \Theta_i)$ achieves an error of the form C_f^2/N as shown by a simple variance computation, where C_f is a norm on the Fourier transform of f. Since the Fourier transform is a linear operator, it turns out that the error (or remainder function) $f - \hat{f}^{(1)}(x)$ has a Fourier norm on the order of $C_f/\sqrt{N}$, which is much smaller than that of f. We let the next N activation functions approximate this error function with $\hat{f}^{(2)}$, so that $\hat{f}^{(1)} + \hat{f}^{(2)}$ achieves an error of at most $\frac{1}{N^2}$. We continue this argument inductively to obtain rates of $1/N^a$. We note that to carry out this argument, we need stronger conditions on f than the ones used in Barron (1993) (see Section 3). We next briefly describe some of the technical challenges and general proof strategy.

Overview of Proof Strategy. The main representation results are given in Theorems 8 and 9 in Section 3. We briefly describe our proof strategy:

1. The Fourier transform of the ReLU function is not well-behaved, due to its non-differentiability at 0. We construct an appropriate class of *smoothed* ReLU functions SReLU, which is the same as ReLU except in a small neighborhood around the origin, by convolving ReLU with a specific probability density. This is done in Section A.

2. Cosine functions are represented as a convolution of SReLU functions in Theorem 11.

3. We prove a two-layer approximation theorem for f under a Fourier norm condition using SReLU activation functions. This is done in Theorems 13 and 16.

4. In Theorem 7 we extend the error function $f^{\mathsf{rem}} := f - \hat{f}^{(1)}$ to all of $\mathbb{R}^d$ and show that its Fourier norm is smaller by a factor of $1/\sqrt{N}$ than that of f. Since activation functions used to construct $\hat{f}^{(1)}$ are *one-dimensional* and their Fourier transforms are *generalized functions*, we will use the "mollification" trick from Fourier analysis to extend them to be d dimensional functions with continuous Fourier transforms.

5. We use the next set of non-linear units to represent the error f^{rem} and continue recursively until the rate of $\frac{1}{N^a}$ is achieved. Since the remainder function becomes less smooth after each approximation step, we can only continue this procedure while the remainder is smooth enough to be effectively approximated by the class of activation functions considered. This depends on the smoothness of the original function f. (Roughly, an increased number of bounded derivatives of f allows taking larger a.)

The guarantees we obtain above contain dimension dependent factors which can be quite large. By considering functions with *low-dimensional structure* – that is, d dimensional functions whose effective dimension is $q \ll d$ as described below, the dimension dependent factor can be improved to depend only on q and not on d.

1.2. Functions with Low-Dimensional Structure

Let $d \in \mathbb{N}$ and $d \geq q$. We build a function $f : \mathbb{R}^d \to \mathbb{R}$ from real valued functions $f_i : \mathbb{R}^q \to \mathbb{R}$ for $i = 1, \ldots, m$ as follows. Let $B_i \subset \mathbb{R}^d$ be finite sets such that $|B_i| = q$ and for all $u, v \in B_i$, $\langle u, v \rangle = \delta_{u,v}$. We fix an ordering for the elements of each set B_i. For ease of notation, for every $x \in \mathbb{R}^d$, define $\langle x, B_i \rangle \in \mathbb{R}^q$ to be the vector whose elements are $(\langle x, v \rangle)_{v \in B_i}$. Define $f : \mathbb{R}^d \to \mathbb{R}$ as

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(\langle x, B_i \rangle) \,. \tag{1}$$

This is a rich class of functions that is dense over the set of $C_c(\mathbb{R}^d)$ equipped with the L^2 norm. This can be seen in various ways, including via universal approximation theorems for neural networks. Such low dimensional structure is often assumed to avoid overfitting in statistics and machine learning – for instance, linear regression in which case $m = q = 1$.

Low-Degree Polynomials. Low-degree polynomials are a special case of functions in the form of (1). For each $V : [d] \to \{0\} \cup [d]$ such that $\sum_{j \in [d]} V(j) \leq q$ denote by $p_V : \mathbb{R}^d \to \mathbb{R}$ the corresponding monomial given by $p_V(x) = \prod_{j \in V} x_j^{V(j)}$. We note that each p_V can depend on at most q coordinates, and a standard dot and dash argument shows that the number of distinct V are $\binom{q+d}{q}$. We consider the class of polynomials of $x \in \mathbb{R}^d$ with degree at most q, where $q \ll d$, which are of the form

$$f(x) = \sum_{V} J_V p_V(x) \tag{2}$$

for arbitrary $J_V \in \mathbb{R}$. Our results in Theorem 10 show how to approximate $f(x)$ for $x \in [0,1]^d$ under some given probability measure over this set.

1.3. Preliminaries and Notation

In this paper d always denotes the dimension of some space like $\mathbb{R}^d$, which we take as the space of features of our data. We also consider $\mathbb{R}^q$ where $q \ll d$ and functions over them, especially when considering functions over $\mathbb{R}^d$ with a q dimensional structure as defined just above. $B_q^2(r)$ for $r > 0$ denotes the Euclidean ball $\{x \in \mathbb{R}^q : \|x\|_2 \leq r\}$. In this paper, we consider approximating a function f over some bounded set $B_d^2(r)$ or $B_q^2(r)$. Therefore, we are free to extend f outside this. The standard ℓ^2 Euclidean norm is denoted by $\|\cdot\|$.

We let capitals denote Fourier transforms. For example the Fourier transform of $g : \mathbb{R}^q \to \mathbb{R}$, $g \in L^1(\mathbb{R}^q)$ is denoted by $G(\omega) = \int_{\mathbb{R}^q} g(x) e^{i\langle \omega, x \rangle} dx$. Following the discussion in Barron (1993), we scale G to $\frac{G}{(2\pi)^q}$ to get the 'Fourier distribution' of g. Whenever $G \in L^1(\mathbb{R}^q)$, the Fourier inversion formula implies that for all $x \in \mathbb{R}^q$,

$$g(x) = \int_{\mathbb{R}^q} \tfrac{G(\omega)}{(2\pi)^q} e^{-i\langle \omega, x \rangle} d\omega \,. \tag{3}$$

Following Barron (1993), we also consider complex signed measures (instead of functions over $\mathbb{R}^q$) as "Fourier distributions" corresponding to g as long as Equation (3) holds for every x. In this case the formal integration against $\frac{G(\omega)}{(2\pi)^d}d\omega$ is understood to be integration with respect to this signed measure. This broadens the class of functions g that fall within the scope of our results. We denote the Schwartz space over $\mathbb{R}^q$ by $\mathcal{S}(\mathbb{R}^q)$. This space is closed under Fourier and inverse Fourier transforms. Finally, for real x let $\mathsf{ReLU}(x) = \max(0, x)$.

1.4. Random Features Model and Training

The random features model was first studied in Rahimi and Recht (2008b,a, 2009) as an alternative to kernel methods. The representation results in Barron (1993); Klusowski and Barron (2018); Sun et al. (2018); Bailey et al. (2019); Ji et al. (2019) and in this work use random features. In order to approximate a target function $f : \mathbb{R}^d \to \mathbb{R}$ we consider functions of the form $\hat{f}(x; \mathbf{v}) = \sum_{j=1}^{N} v_j \sigma(\langle \omega_j, x\rangle - T_j)$. Here we have denoted $(v_j) \in \mathbb{R}$ in the RHS collectively by $\mathbf{v}$ in the LHS, and $\omega_j \in \mathbb{R}^d$ and $T_j \in \mathbb{R}$ are random variables. We optimize over $\mathbf{v}$, keeping ω_j's and T_j's fixed to find the best approximator for f. More specifically, we want to solve the following loss minimization problem for some probability distribution ζ over $\mathbb{R}^d$:

$$\mathbf{v}^* = \arg \inf_{\mathbf{v}\in\mathbb{R}^N} \int \big(f(x) - \hat{f}(x; \mathbf{v})\big)^2 \zeta(dx)\,. \tag{4}$$

The problem above reduces to a least squares linear regression problem which can be easily and efficiently solved via gradient descent since this is an instance of a smooth convex optimization problem. By Theorem 3.3 in Bubeck et al. (2015), constant step-size gradient descent (GD) has an excess squared error $O(1/T)$ compared to the optimal parameter $\mathbf{v}^*$ after T steps. In this paper, whenever we prove a learning result, we first show that with high probability over the randomness in ω_j, T_j, there exists a $\mathbf{v}_0$ such that the loss in approximating f via $\hat{f}(\,\cdot\,; \mathbf{v}_0)$ is at most $\epsilon/2$. Then, running GD for the objective in Equation (4) for $T = \Omega(1/\epsilon)$ steps, we obtain $\mathbf{v}_T$ such that $\int \big(f(x) - \hat{f}(x; \mathbf{v}_T)\big)^2 \zeta(dx) \le \epsilon$.

Since this paper mainly concerns the complexity in terms of the number of activation functions, we omit the details about time complexity of GD in our results, but it is understood throughout to be $O(1/\epsilon)$. The random features model is considered a good model for networks with a large number of activation functions since during training with SGD, the weights ω_j and T_j do not change appreciably compared to the initial random value. Such a consideration has been used in the literature to obtain learning guarantees via SGD for large neural networks Andoni et al. (2014); Daniely (2017b); Du et al. (2019).

1.5. Organization

The paper is organized as follows. In Section 2, we illustrate the corrective mechanism by developing our results on memorization by two-layer ReLU networks via SGD to conclude Theorem 1. We then proceed to state our main results on function representation and learning polynomials in Section 3. We give the construction of the smoothed ReLU activation functions in Section A and state an integral representation for cosine functions in terms of these activation functions. The proof of the main technical result of the paper, Theorem 7, is in Section B. Sections C through F contain many of the proofs.

2. Memorization

We first present our results on memorization, as they are the least technical yet suffice to illustrate the corrective mechanism. Suppose we are given n labeled examples $(x_1, y_1), \ldots, (x_n, y_n)$ where each data point $x_i \in \mathbb{R}^d$ has label $y_i \in [0, 1]$. In memorization (also known as interpolation), the goal is to construct a neural network which can be trained via SGD and which outputs $\hat{f}(x_i) = \hat{y}_i \approx y_i$ when the input is x_i, for every $i \in [n]$. The basic question is: how many neurons are needed?

Theorem 1 *Suppose $x_1, \ldots, x_n \in \mathbb{R}^d$ are such that $\|x_j\| \leq 1$ and $\min_{k \neq l} \|x_l - x_k\| \geq \theta$. For each $i = 1, \ldots, n$ let $y_i \in [0, 1]$ be an arbitrary label for x_i. Let (ω_j, T_j) for $j = 1, \ldots, N$ be drawn i.i.d. from the distribution $\mathcal{N}(0, \sigma_0^2 I_d) \times \mathsf{Unif}[-2, 2]$, where $\sigma_0 = 1/\sqrt{C_0 \times \log n \times \log \max(1/\theta, 2)}$ for some large enough constant C_0. Let C be a sufficiently large universal constant and let $\epsilon, \delta \in (0, 1)$ be arbitrary. If $N \geq Cn \frac{\log^4(\max(1/\theta, 2)) \log^4 n}{\theta^4} \log \frac{n}{\delta\epsilon}$, then with probability at least $1 - \delta$ there exist $a_1, \ldots, a_N \in \mathbb{R}$ such that the function $\hat{f}_N^{\mathsf{ReLU}} := \sum_{j=1}^N a_j \mathsf{ReLU}(\langle x, \omega_j \rangle - T_j)$ satisfies*

$$\sum_{k=1}^n \left(f(x_k) - \hat{f}_N^{\mathsf{ReLU}}(x_k)\right)^2 \leq \epsilon \,.$$

Moreover, if we consider only $a_1, \ldots, a_N$ as the free parameters and keep the weights (ω_j, T_j) fixed, SGD/GD obtains the optimum because the objective is a convex function.

Remark 2 *In the initial version of this paper, there was an extra factor of d^2 in the guarantees given above. Based on reviewer comments, we have removed this dependence using a more refined analysis.*

In the remainder of this section we will prove Theorem 1. We will first show a Fourier-analytic representation. However, instead of using the regular Fourier transform, only in this section, we use the discrete Fourier Transform. For a function $f : \{x_1, \ldots, x_n\} \to \mathbb{R}$, define $F : \mathbb{R}^d \to \mathbb{R}$

$$F(\xi) := \sum_{j=1}^n f(x_j) e^{i\langle \xi, x_j \rangle} \,.$$

The proof now proceeds in five steps.

Step 1: Approximation via Fourier transform. Let $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ for $\sigma > 0$ to be specified momentarily and consider $\tilde{f} : \{x_1, \ldots, x_n\} \to \mathbb{R}$ defined as

$$\tilde{f}(x_k) := \mathbb{E} F(\xi) e^{-i\langle \xi, x_k \rangle} = f(x_k) + \sum_{j \neq k} f(x_j) \mathbb{E} e^{i\langle \xi, x_j - x_k \rangle} = f(x_k) + \sum_{j \neq k} f(x_j) e^{-\frac{\sigma^2 d_{jk}^2}{2}} \,,$$

where $d_{jk} = \|x_j - x_k\|_2$ and we have used the fact that the Gaussian $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ has characteristic function $\mathbb{E}[e^{-i\langle t, \xi \rangle}] = \exp(-\frac{1}{2}\sigma^2 \|t\|^2)$. Note that when σ is large enough compared to $1/\theta$, we have $\tilde{f}(x_k) \approx f(x_k)$, so in what follows we will aim to approximate $\tilde{f}$. We will take $\sigma = \theta^{-1}\sqrt{2s \log n}$ for some $s > 1$ to be fixed later.

We now record some properties of the random variable $F(\xi)$. Let $\|f\|_p$ denote the standard Euclidean ℓ^p norm when f is viewed as a n-dimensional vector $(f(x_1), \ldots, f(x_n))$. The proof of the following lemma is given in Section E.

Lemma 3 *Let $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ where $\sigma = \sqrt{2s \log n}/\theta$. We have:*

1. $|F(\xi)| \leq \|f\|_1$ *almost surely,*
2. $\mathbb{E}|F(\xi)|^2 \leq \|f\|_2^2 + \|f\|_1^2/n^s$, *and*
3. $|f(x_k) - \tilde{f}(x_k)| \leq \|f\|_1/n^s$.

Step 2: Replacing sinusoids by ReLU. We first state a lemma which allows us to represent sinusoids in terms of ReLU and Step functions. The proof is given in Section C.

Lemma 4 *Let $T \sim \mathsf{Unif}[-2, 2]$. There exist $C_c^\infty(\mathbb{R})$ functions $\eta(\,\cdot\,; \alpha, \psi)$, (where α and ψ are the parameters which define η) such that $\sup_{T \in \mathbb{R}} |\eta(T; \alpha, \psi)| \leq 1$ and for every $t \in [-1, 1]$ and for some absolute constant C, we have*

$$\cos(\alpha t + \psi) = \mathbb{E} C(1 + \alpha^2)\eta(T; \alpha, \psi)\mathsf{ReLU}(t - T)$$

Consider the event $\mathcal{A} = \{|\langle \xi, x_k \rangle| > \frac{2s \log n}{\theta} \text{ for some } k \in [n]\}$. By Gaussian concentration, we have $\mathbb{P}(\mathcal{A}) \leq 2/n^{s-1}$. Write $F(\xi) = |F(\xi)|e^{-i\phi(\xi)}$ for some $\phi : \mathbb{R}^d \to \mathbb{R}$. In Lemma 4 we take $\alpha = \frac{2s \log n}{\theta}$, $t = \langle \xi, x_k \rangle / \alpha$, and $\psi = \phi(\xi)$ to conclude that if $T \sim \mathsf{Unif}[-2, 2]$ and independent of ξ, then on the event $\mathcal{A}^c$

$$\cos\big(\langle \xi, x_k \rangle + \phi(\xi)\big) = \mathbb{E}_T C(1 + \tfrac{4s^2 \log^2 n}{\theta^2})\eta(T; \alpha, \psi)\mathsf{ReLU}\Big(\theta \tfrac{\langle \xi, x_k \rangle}{2s \log n} - T\Big)$$

Here $\mathbb{E}_T$ denotes the expectation only over the random variable T, C is a universal constant and η is as given by Lemma 4. We have used the fact that $\frac{\theta \langle \xi, x_k \rangle}{2s \log n} \in [-1, 1]$ since the event $\mathcal{A}^c$ holds. Now, by definition of $\tilde{f}$, we have

$$\tilde{f}(x_k) = \mathbb{E} F(\xi) e^{-i\langle \xi, x_k \rangle} = \mathbb{E}|F(\xi)|e^{-i\phi(\xi) - i\langle \xi, x_k \rangle} = \mathbb{E}|F(\xi)| \cos\big(\langle \xi, x_k \rangle + \phi(\xi)\big)\,.$$

The last two equations lead to the following lemma, with details given in Section E.

Lemma 5 *For some absolute constant C_1, we have*

$$\left|\tilde{f}(x_k) - C\mathbb{E}|F(\xi)|\big(1 + \tfrac{4s^2 \log^2 n}{\theta^2}\big)\eta(T; \alpha, \psi)\mathsf{ReLU}\Big(\theta \tfrac{\langle \xi, x_k \rangle}{2s \log n} - T\Big)\right| \leq C_1 \tfrac{s^{3/2}\|f\|_1 \log^{3/2} n}{\theta^2 n^{s/2}}\,. \quad (5)$$

Step 3: Empirical estimate. Let $N_0 \in \mathbb{N}$. We draw (ξ_l, T_l) for $l \in \{1, \ldots, N_0\}$ i.i.d. from the distribution $\mathcal{N}(0, \sigma^2 I_d) \times \mathsf{Unif}[-2, 2]$. We construct the following estimator for $\tilde{f}$, which is in turn an estimator for f:

$$\hat{f}_l(x) := C|F(\xi_l)|\big(1 + \tfrac{4s^2 \log^2 n}{\theta^2}\big)\eta(T_l; \alpha, \phi(\xi_l))\mathsf{ReLU}\Big(\theta \tfrac{\langle \xi_l, x_k \rangle}{2s \log n} - T_l\Big)\,.$$

From Equation (5), we conclude that $\mathbb{E}\hat{f}_l(x_k) = \tilde{f}(x_k) + O\Big(\frac{s^{3/2}\|f\|_1 \log^{3/2} n}{\theta^2 n^{s/2}}\Big)$ and we construct the empirical estimate

$$\hat{f}(x) := \frac{1}{N_0} \sum_{l=1}^{N_0} \hat{f}_l(x)\,. \quad (6)$$

Lemma 6 *For some universal constant C and $L := C\frac{s^4 \log^4 n}{\theta^4}$,*

$$\mathbb{E}\big(f(x_j) - \hat{f}(x_j)\big)^2 \leq \left[\frac{L}{N_0} + \frac{Cs^3 \log^3 n}{\theta^4 n^{s-1}}\right] \|f\|_2^2 .$$

In particular, letting $s = C_1 + C_2 \log(\max(1/\theta, 2))$ for some constants C_1, C_2 and $N_0 = 2neL$ yields

$$\mathbb{E}\big(f(x_j) - \hat{f}(x_j)\big)^2 \leq \frac{\|f\|_2^2}{en} . \tag{7}$$

The proof, given in Section E, follows from an application of Gaussian concentration.

Step 4: Iterative correction. We define $f^0 : \{x_1, \ldots, x_j\} \to \mathbb{R}$ by $f^0(x_j) := y_j$ where $y_j \in [0, 1]$ are the desired labels for x_j. In the procedure above, we replace f with f^0 and obtain the estimator $\hat{f}^0$ as per Equation (6). We now define the remainder function $f^{\mathsf{rem},1} : \{x_1, \ldots, x_n\} \to \mathbb{R}$ as the error obtained by the approximation: $f^{\mathsf{rem},1}(x_k) := f^0(x_k) - \hat{f}^0(x_k)$. Summing the bound in Equation 7 over $j \in [n]$ yields

$$\mathbb{E}\|f^{\mathsf{rem},1}\|_2^2 \leq \frac{\|f^0\|_2^2}{e} . \tag{8}$$

We define higher order remainders $f^{\mathsf{rem},l}$ for $l \geq 2$ inductively as follows. Suppose we have $f^{\mathsf{rem},l-1}$. We replace f in the procedure above with $f^{\mathsf{rem},l-1}$ to obtain the estimator $\hat{f}^{\mathsf{rem},l-1}$ as given in Equation (6), independent of all the previous estimators. We define the remainder $f^{\mathsf{rem},l} = f^{\mathsf{rem},l-1} - \hat{f}^{\mathsf{rem},l-1}$. Repeating the argument leading to Equation (8), with the given choice of s and N_0 we conclude that: $\mathbb{E}\|f^{\mathsf{rem},l}\|_2^2 \leq e^{-l}\|f^0\|_2^2$. Take $N = lN_0$. Unrolling the recursion above, we note that $f^{\mathsf{rem},l}(x)$ is $f^0(x) - \hat{f}^l(x)$, where $\hat{f}^l(x)$ is of the form

$$\hat{f}^{(l)}(x) = \sum_{j=1}^{N} a_j \mathsf{ReLU}\left(\frac{\theta\langle \xi_j, x\rangle}{2s \log n} - T_j\right) . \tag{9}$$

This is the output of a two-layer network with lN ReLU units. We recall that (ξ_j, T_j) are i.i.d. $\mathcal{N}(0, \sigma^2 I_d) \times \mathsf{Unif}[-2, 2]$ which agrees with the choice of weights in Theorem 1. The remainder $f^{\mathsf{rem},l}(x)$ can be seen as the error of approximating f^0 using $N_0 l := N$ random activation functions as given in Equation (9). By assumption, the labels $f^0(x_j) \in [0, 1]$, so $\|f^0\|_2^2 \leq n$. This gives us an error bound on the L^2 loss $\mathcal{E}_N(f^0) := \sum_{j=1}^n (f^0(x_j) - \hat{f}^{(l)}(x_j))^2$:

$$\mathbb{E}\mathcal{E}_N(f^0) \leq e^{-l} n .$$

Step 5: Markov's inequality. Denoting by $E_N(f^0) = e^{-l} n$ the RHS of the bound just above, Markov's inequality implies that for any $\delta \in (0, 1)$

$$\mathbb{P}\Big(\mathcal{E}_N(f^0) \geq \tfrac{E_N(f^0)}{\delta}\Big) \leq \delta . \tag{10}$$

Now the choice $l \geq \log n + \log\frac{1}{\delta} + \log\frac{1}{\epsilon}$ gives $\frac{E_N(f^0)}{\delta} \leq e^{-l}\frac{n}{\delta} \leq \frac{n}{\delta} e^{-\log\frac{n}{\epsilon\delta}} \leq \epsilon$ and plugging into Equation (10) shows that when s, l and N are chosen as above, we have $\mathbb{P}(\mathcal{E}_N(f^0) \geq \epsilon) \leq \delta$ as claimed in Theorem 1.

3. Representation via the Corrective Mechanism

We now turn to the representation problem. Given a function $g : \mathbb{R}^q \to \mathbb{R}$, the goal is to construct a neural network whose output $\hat{g}$ is close to g. The arguments resemble those given in the previous section on memorization, but the details are more technically involved.

The approximation guarantees of our theorems depend on certain *Fourier norms*. These can be thought of as measures of the complexity of the function g to be approximated. Let $g : \mathbb{R}^q \to \mathbb{R}$ be the function we are trying to approximate over the domain $B_q^2(r)$ and let $\frac{G(\omega)}{(2\pi)^q}$ be the 'Fourier distribution' of $g : \mathbb{R}^q \to \mathbb{R}$ as defined in Equation (3). We take its magnitude-phase decomposition to be: $G(\omega) = |G(\omega)|e^{-i\psi(\omega)}$. For each integer $s \geq 0$ we define the *Fourier norm*

$$C_g^{(s)} := \frac{1}{(2\pi)^q} \int_{\mathbb{R}^q} |G(\omega)| \cdot \|\omega\|^s d\omega \,.$$

We will assume that $C_g^{(s)} < \infty$ for $s = 0, 1, \ldots, L$ for some $L \in \mathbb{N}$.

Because having small Fourier norm can be thought of as a smoothness property, *smoothed* ReLU functions can be efficiently used for the task of approximating such functions. In Section A we define a sequence of smoothed ReLU functions SReLU_k for integers $k \geq 0$, of increasing smoothness. These are obtained from the ReLU by convolving with an appropriate function. The use of smoothed ReLU functions is crucial in order that the *remainder* following approximation is itself sufficiently smooth, which then allows the approximation procedure to be iterated. We start with the basic approximation theorem, which has an approximation guarantee as well as a smoothness guarantee on the remainder.

Theorem 7 *Let $k \geq \max(1, \frac{q-3}{4})$. Let $g : \mathbb{R}^q \to \mathbb{R}$ be such that $C_g^{(2k+2)}, C_g^{(0)} < \infty$. Then, given any probability measure ζ over $B_q^2(r)$ there exists a two-layer SReLU_k network, with N non-linear units, whose output is $\hat{g}(x)$ such that the following hold simultaneously:*

1. $$\int (g(x) - \hat{g}(x))^2 \zeta(dx) \leq \frac{C(r,k)\big(C_g^{(0)} + C_g^{(2k+2)}\big)^2}{N} \,.$$

2. *There exists a function $g^{\mathsf{rem}} : \mathbb{R}^q \to \mathbb{R}$ such that:*

   (a) *For every $x \in B_q^2(r)$, $g^{\mathsf{rem}}(x) = g(x) - \hat{g}(x)$.*

   (b) *Its Fourier transform $G^{\mathsf{rem}} \in L^1(\mathbb{R}^q) \cap C(\mathbb{R}^q)$.*

   (c) *For every $s < \frac{3-q}{2} + 2k$, $C_{g^{\mathsf{rem}}}^{(s)} \leq C_1(s, r, q, k)(C_g^{(0)} + C_g^{2k+2})/\sqrt{N}$.*

We will use this theorem to give a faster approximation rate of $\frac{1}{N^{a+1}}$ for g, where $a \in \mathbb{N} \cup \{0\}$. We then extend this to functions of the from given in Equation (1). The main conclusion of the following theorem is that the approximating network achieves an error of at most ϵ with $N = O(C(a)\epsilon^{-\frac{1}{a+1}})$ activation functions. If the theorem below holds for every $a \in \mathbb{N} \cup \{0\}$, we note that if we take $a \to \infty$ slowly enough as $\epsilon \to 0$, we get subpolynomial dependence on ϵ.

Theorem 8 *Fix $q \in \mathbb{N}$ and for each $b \in \mathbb{N} \cup \{0\}$ let*

$$k_b = \begin{cases} b\left\lceil \frac{1+q}{4} \right\rceil & \text{if } q \not\equiv 3 \pmod 4 \\ b\left(\frac{1+q}{4} + 1\right) & \text{if } q \equiv 3 \pmod 4 . \end{cases} \tag{11}$$

Suppose $g : \mathbb{R}^q \to \mathbb{R}$ has bounded Fourier norms $C_g^{(0)} < \infty$ and $C_g^{(2k_a+2)} < \infty$ for some $a \in \{0\} \cup \mathbb{N}$. Then, for any probability measure ζ over $B_q^2(r)$, there exists a two-layer neural network with random weights and N activation functions consisting of a mixture of SReLU_{k_b} units for $b \in \{0, 1, \ldots, a\}$ with output $\hat{g} : \mathbb{R}^q \to \mathbb{R}$ such that

1. *For every $x \in B_q^2(r)$, $\mathbb{E}\hat{g}(x) = g(x)$.*
2. $$\mathbb{E} \int \left(g(x) - \hat{g}(x)\right)^2 \zeta(dx) \leq C_0(q, r, a) \frac{\left(C_g^{(0)} + C_g^{(2k_a+2)}\right)^2}{N^{a+1}} .$$

The expectation here is with respect to the randomness in the weights of the neural network. Moreover, writing $\hat{g}$ in the form $\hat{g}(x) = \sum_{j=1}^N \kappa_j \mathsf{SReLU}_{k(j)}(\langle \omega_j, x \rangle - T_j)$, the κ_j and ω_j satisfy $\sum_{j=1}^N |\kappa_j| \leq C_1(q, r, a)(C_g^0 + C_g^{2k_a+2})$ and $\|\omega_j\| \leq \frac{1}{r}$ almost surely.

Proof The main idea of the proof is to use Theorem 7 repeatedly. We will first use $\sim N/(a+1)$ SReLU_{k_a} units to approximate g by $\hat{g}^{(0)}$. This gives a squared error of the order $O(1/N)$. We then consider the error term $g - \hat{g}^{(0)}$ and approximate this error term using another $\sim N/(a+1)$ $\mathsf{SReLU}_{k_{a-1}}$ units and try to offset the first error to obtain a squared error guarantee of $1/N^2$, and repeat this procedure until we obtain the stated guarantees. We reduce the smoothness parameter k in every iteration as error terms become progressively less smooth. A complete proof is provided in Section F. ■

Now we prove the version of Theorem 8 for functions of the form in Equation (1). The main advantage of Theorem 9 is that the bounds do not depend on the dimension d, only on the effective dimension $q \ll d$.

Theorem 9 *Consider the low-dimensional function defined in Equation (1). Assume that*

$$\sup_i \left(C_{f_i}^{(0)} + C_{f_i}^{(2k_a+2)}\right)^2 \leq M .$$

Then, for any probability measure ζ over $B_d^2(r)$, there exists a one non-linear layer neural network with ReLU and SReLU_k units for $k \leq k_a$ with N neurons with output $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ such that

$$\int \left(f(x) - \hat{f}(x)\right)^2 \zeta(dx) \leq C_0(q, r, a, l) \frac{M m^a}{N^{a+1}} .$$

Proof We use N/m neurons to approximate each of the component functions f_i just like in Theorem 8, and then average the outputs. The full proof is in Section F. ■

We will now develop our results on learning low-degree polynomials. The results are based on Theorem 32 which is similar to Theorem 9, but with a stronger bounded sup norm type assumption on the Fourier transform instead. This has the advantage that we can sample the weights independent of the target function g and k to construct our network. The proofs are developed in Section D, which is roughly similar to Section B.

Let the probability measure μ_l over $\mathbb{R}$ be defined by $\mu_l(dt) \propto \frac{dt}{1+t^{2l}}$ for $l \in \mathbb{N}$. Given $a, m, N \in \mathbb{N} \cup \{0\}$ such that $\frac{N}{(a+1)m} \in \mathbb{N}$ and the orthonormal sets B_i be as used in Equation (1), consider the following sampling procedure:

1. Partition $[N] \subseteq \mathbb{N}$ into m disjoint sets, each with $N/(m(a+1))$ elements.
2. For $i \in [m]$, $b \in \{0, \ldots, a\}$, $j \in [\frac{N}{m(a+1)}]$, we draw $\omega^0_{i,j,b} \sim \mathsf{Unif}\left(\mathbb{S}^{\mathrm{span}(B_i)}\right)$ and $T_{i,j,b} \sim \mu_l$ independently for some $l \geq \max(q+3, 3a+3)$.

We now specialize to the low degree polynomials defined in Equation (2). Define the following orthonormal set associated with each V in the summation:

$$B_V = \{e_j : V(j) \neq 0\} \cup \bar{B}_V \,,$$

where e_j are the standard basis vectors in $\mathbb{R}^d$ and $\bar{B}_V \subseteq \{e_1, \ldots, e_d\}$ is chosen such that $|B_V| = q$. Consider the sampling procedure given above with the bases B_V. Since the bases B_V are known explicitly, this sampling can be done without knowledge of the polynomial. We have the following theorem about learning low-degree polynomials, proved in Section F.

Theorem 10 *Let $m = \binom{q+d}{q}$, $r = \sqrt{q}$ and let J be the m-dimensional vector whose entries are J_V. Let $a \in \mathbb{N} \cup \{0\}$, $\delta \in (0,1)$ and $\epsilon, R_c > 0$ be arbitrary. Let N be chosen such that and $N/(a+1)m \in \mathbb{N}$. Let ζ be any probability measure over $[0,1]^d$. Generate the weights $(\omega^0_{i,j,b}, T_{i,j,b})$ according to the sampling procedure described above. Construct the two-layer neural network with N activation functions*

$$\hat{f}(x; \mathbf{v}) = \sum_{i=1}^{m} \sum_{b=0}^{a} \sum_{j=1}^{\frac{N}{m(a+1)}} v_{i,j,b} \mathsf{SReLU}_{k_b^S}\left(\frac{\langle \omega^0_{i,j,b}, x\rangle}{r} - T_{i,j,b}\right) . \tag{12}$$

Here we have denoted the vector comprising of $v_{i,j,k}$ by $\mathbf{v}$. Let $\mathbf{v}^* \in \arg\inf_{\mathbf{v} \in B^2_N(R_c)} \int (f(x) - \hat{f}(x;\mathbf{v}))^2 \zeta(dx)$. Let $b \in \mathbb{N} \cup \{0\}$, $b \leq a$. There exists a constant $C(a,q,l)$ such that if

$$N \geq C(a,q,l) \max\left(\frac{\delta^{-1/(b+1)} \|J\|^2_{2(b+1)} m^{2-1/(b+1)}}{R_c^2}, \left(\frac{\|J\|_2^2}{\epsilon\delta}\right)^{\frac{1}{a+1}} m\right) ,$$

then with probability at least $1-\delta$,

$$\int \left(f(x) - \hat{f}(x; \mathbf{v}^*)\right)^2 d\zeta \leq \epsilon \,.$$

Moreover, we can obtain the coefficients $v^*_{i,j,b}$ using GD over the outer layer only since this is a convex optimization problem.

4. Acknowledgments

This work was supported in part by MIT-IBM Watson AI Lab and NSF CAREER award CCF-1940205.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

Bolton Bailey, Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Approximation power of random neural networks. *arXiv preprint arXiv:1906.07709*, 2019.

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

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.

Guy Bresler and Dheeraj Nagaraj. Sharp representation theorems for relu networks with precise dependence on depth. *arXiv preprint arXiv:2006.04048*, 2020.

Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and weights size for memorization with two-layers neural networks. *arXiv preprint arXiv:2006.02855*, 2020.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

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

Amit Daniely. Depth separation for neural networks. *arXiv preprint arXiv:1702.08489*, 2017a.

Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017b.

Amit Daniely. Neural networks learning and memorization with (almost) no over-parameterization. *arXiv preprint arXiv:1911.09873*, 2019.

Olivier Delalleau and Yoshua Bengio. Shallow vs. deep sum-product networks. In *Advances in Neural Information Processing Systems*, pages 666–674, 2011.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

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

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. *arXiv preprint arXiv:1910.06956*, 2019.

Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 92–99. IEEE, 2019.

Jason M Klusowski and Andrew R Barron. Approximation by combinations of relu and squared relu ridge functions with l^1 and l^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

Bo Li, Shanshan Tang, and Haijun Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *arXiv preprint arXiv:1903.05858*, 2019.

Shiyu Liang and Rayadurgam Srikant. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239, 2017.

Chao Ma, Lei Wu, et al. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

Abhishek Panigrahi, Abhishek Shetty, and Navin Goyal. Effect of activation functions on the training of overparametrized neural nets. *arXiv preprint arXiv:1908.05660*, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008a.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, 2008b.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2979–2987. JMLR. org, 2017.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

Yitong Sun, Anna Gilbert, and Ambuj Tewari. On the approximation properties of random relu features. *arXiv preprint arXiv:1810.04374*, 2018.

Matus Telgarsky. benefits of depth in neural networks. In *Conference on Learning Theory*, pages 1517–1539, 2016.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Hao Zheng, Zhanlei Yang, Wenju Liu, Jizhong Liang, and Yanpeng Li. Improving deep neural networks using softplus units. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4. IEEE, 2015.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

Appendix A. Smoothed ReLU functions and Integral Representations

In this section we introduce the necessary technical results and constructions for function approximation by smoothed ReLU units SReLU_k, as used in Theorems 7 and 8. In Theorem 11 we show that cosine functions can be represented in terms of SReLU_k functions similar to Step 2 in Section 2. This will later be used along with the Fourier inversion formula to represent the target function g in terms of the activation functions SReLU_k. We note that this idea is taken from Barron (1993) and Klusowski and Barron (2018). All of the results stated here are proved in Section C.

A.1. Smoothing the ReLU

Consider the triangle function

$$\lambda(t) = \begin{cases} 1 - |t| & \text{for } |t| \leq 1 \\ 0 & \text{otherwise}\,. \end{cases} \tag{13}$$

Clearly, λ is a symmetric, bounded and continuous probability density over $\mathbb{R}$. Denoting the Fourier transform of λ by Λ, one can verify the standard fact that $\Lambda(\xi) = \frac{\sin^2(\xi/2)}{(\xi/2)^2}$.

We also consider k-fold convolution of λ with itself: Let $\lambda_1 := \lambda$ and $\lambda_{l+1} := \lambda_1 * \lambda_l$ for $l \geq 1$. For each $k \geq 1$ the function λ_k has support $[-k, k]$, it is a symmetric, bounded and continuous probability density over $\mathbb{R}$, and its Fourier transform is $\Lambda_k(\xi) = \frac{\sin^{2k}(\xi/2)}{(\xi/2)^{2k}}$. For arbitrary $w_0 > 0$, we define $\lambda_{k,w_0}(t) := \frac{k}{w_0}\lambda_k(\frac{tk}{w_0})$, which can also be verified to be a symmetric, continuous probability density over $\mathbb{R}$ with support $[-w_0, w_0]$, and its Fourier transform is given by $\Lambda_{k,w_0}(\xi) = \Lambda_k(\frac{\xi w_0}{k})$.

We now "cosine regularize" λ_k so that its Fourier transform is non-zero everywhere. This transformation is for purely technical reasons and is useful in the proof of Theorem 11 stated below. Let $\alpha_0 > 0$ and $w_0 \leq \min(\frac{\pi}{2\alpha_0}, \frac{\pi k}{4\alpha_0})$ and define

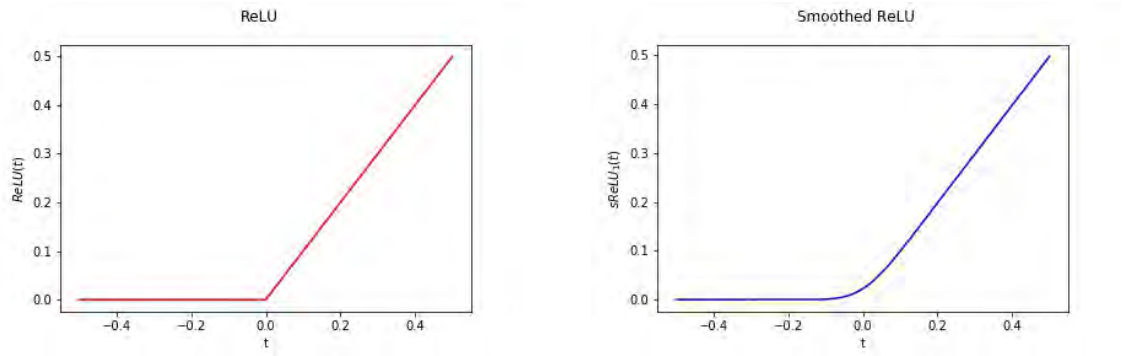
$$\lambda^{\alpha_0}_{k,w_0}(t) := \cos(\alpha_0 t)\lambda_{k,w_0}(t) \Big/ \int_{-\infty}^{\infty} \cos(\alpha_0 T)\lambda_{k,w_0}(T)dT\,. \tag{14}$$

The constraints given on α_0 and w_0 ensure that $\lambda^{\alpha_0}_{k,w_0}(t) \geq 0$ for every t. We will henceforth think of α_0 and w_0 as fixed (say $w_0 = 0.5$ and $\alpha_0 = \frac{\pi}{16}$). We define the *smoothed* ReLU functions

$$\mathsf{SReLU}_k := \mathsf{ReLU} * \lambda^{\alpha_0}_{k,w_0} \quad \text{for all } k \geq 1$$

and hide the dependence on w_0, α_0. Clearly, SReLU_k is an increasing, positive function and $\mathsf{SReLU}_k(t) = \mathsf{ReLU}(t)$ whenever $t \notin (-w_0, w_0)$. We follow the convention that for $k = 0$, $\mathsf{SReLU}_k = \mathsf{ReLU}$. In particular, $\mathsf{SReLU}_k(t) = 0$ whenever $t \leq -w_0$. We give an illustration of these functions in Figure 1. The higher the value of k, the smoother the function is at 0. In the sequel, whenever we say "smoothed by filter $\lambda^{\alpha_0}_{k,w_0}$", we mean convolution with the function $\lambda^{\alpha_0}_{k,w_0}$.

Theorem 11 (Cosine Representation Theorem) *Consider the probability measure μ_l over $\mathbb{R}$ given by $\mu_l(dT) = \frac{c_{\mu_l} dT}{1+T^{2l}}$ (here c_{μ_l} is the normalizing constant). Let $\alpha, \psi \in \mathbb{R}$ be*



(a) ReLU (b) SReLU_1

Figure 1: Illustrating ReLU and SReLU activation functions.

given. There exists a continuous function $\kappa : \mathbb{R} \to \mathbb{R}$ *depending on* $\alpha, \theta, l, k, w_0$ *and* α_0 *such that* $\|\kappa\|_\infty \leq C(k,l)\left(1 + |\alpha|^{2k+2}\right)$ *and for every* $t \in [-1, 1]$

$$\cos(\alpha t + \psi) = \int_{-\infty}^{\infty} \kappa(T)\mathsf{SReLU}_k(t - T)\mu_l(dT)\,.$$

Remark 12 *We note that the upperbound on* κ *gets worse as the smoothness parameter* k *gets larger. This is due to the fact that smoother activation functions find it harder to track fast oscillations in* $\cos(\alpha t + \psi)$ *as* α *gets larger.*

Appendix B. Proof of Theorem 7: Unbiased Estimator for the Function and its Fourier transform

Through the following steps, we describe the proof of Theorem 7, which was in turn used to prove Theorem 8.

Step 1: Representing g **in terms of** SReLU_k Consdier the setup in section 3 and assume $C_g^{(2k+2)}, C_g^{(0)} < \infty$. From Fourier inversion formula, using the fact that g is real-valued, it follows that

$$g(x) = \int_{\mathbb{R}^q} \cos(\langle \omega, x\rangle + \psi(\omega))\frac{|G(\omega)|}{(2\pi)^q} d\omega\,. \tag{15}$$

We combine Theorem 11 and Equation (15) to show the following integral representation for g. The proof is given in Section C.

Theorem 13 *Let* μ_l *be the probability measure defined by its density* $\mu_l(dt) \propto \frac{dt}{1+t^{2l}}$ *for a given* $l \in \mathbb{N}$. *Define the probability distribution* $\nu_{g,k}$ *by* $\nu_{g,k}(d\omega) = \frac{1+r^{2k+2}\|\omega\|^{2k+2}}{C_g^{(0)}+r^{2k+2}C_g^{(2k+2)}}\frac{|G(\omega)|}{(2\pi)^q}d\omega$. *For every* $x \in B_q^{(2)}(r)$

$$g(x) = \beta_{g,k}\iint \eta(T; r, \omega)\mathsf{SReLU}_k\left(\tfrac{\langle \omega, x\rangle}{r\|\omega\|} - T\right)\mu_l \times \nu_{g,k}(dT \times d\omega)\,, \tag{16}$$

where $|\eta(T; r, \omega)| \leq 1$ *almost surely with respect to measure* $\mu_l \times \nu_{g,k}$ *and* $\eta(T; r, \omega) = 0$ *whenever* $T > 1 + w_0$ *and* $\beta_{g,k} := \left(C_g^{(0)} + r^{2k+2}C_g^{(2k+2)}\right)C(k,l)$

Remark 14 *The case $\omega = 0$ might appear ambiguous in the integral representations above. But following our discussion preceding Theorem 13, we use the convention that $\frac{\langle \omega, x \rangle}{\|\omega\|} := 0$ whenever $\omega = 0$. We check that even the constant function can be represented as an integral in Theorem 11 by setting $\alpha = 0$.*

Remark 15 *The probability measure $\nu_{g,k}$ depends on the function g and can be complicated. Therefore, when training a neural network, it is not possible to sample from it since g is unknown. We only use the existence of this measure to prove representation theorems as found in the literature (see Barron (1993), Klusowski and Barron (2018)). To give the training results, we will impose more conditions on G and show that we can get similar representation theorems when a known, fixed measure ν_0 is used instead of $\nu_{g,k}$. This is done in Section D.*

We start by converting Theorem 6, the integral representation of g in terms of SReLU_k units, into a statement about existence of a good approximating network $\hat{g}$.

Step 2: Empirical estimate. Let $\mu_l \times \nu_{g,k}$, η and $\beta_{g,k}$ be as given by Theorem 13. For $j \in \{1, \dots, N\}$, draw (T_j, ω_j) to be i.i.d. from the distribution $\mu_l \times \nu_{g,k}$. Let θ_j^u for $j \in [N]$ be i.i.d. $\mathsf{Unif}[-1, 1]$ and independent of everything else. We define

$$\theta_j := \mathbb{1}\left(\theta_j^u < \eta(T_j; r, \omega_j)\right) - \mathbb{1}\left(\theta_j^u \geq \eta(T_j; r, \omega_j)\right)$$

and observe that $\theta_j \in \{-1, 1\}$ almost surely and $\mathbb{E}\left[\theta_j | T_j, \omega_j\right] = \eta(T_j; r, \omega_j)$. That is, it is an unbiased estimator for $\eta(T_j; r, \omega_j)$ and independent of other $\theta_{j'}$ for $j \neq j'$.

Now define the estimate $\hat{g}_j(x)$ based on a single SReLU_k unit

$$\hat{g}_j(x) := \beta_{g,k} \theta_j \mathsf{SReLU}_k\left(\frac{\langle \omega_j, x \rangle}{r\|\omega_j\|} - T_j\right) \mathbb{1}(T_j \leq 1 + w_0) \tag{17}$$

where we have made the dependence of $\hat{g}_j$ on T_j, ω_j implicit. We also define the empirical estimator

$$\hat{g}(x) = \frac{1}{N} \sum_{j=1}^{N} \left[\hat{g}_j(x)\right] .$$

Note that $\hat{g}$ is the output of a two-layer neural network with one SReLU_k layer and one linear layer.

Theorem 16 *Consider the probability measure μ_l with density $\mu_l(dt) \propto dt/(1 + t^{2l})$ and let $T_j \sim \mu_l$ for $l \geq 2$ (so that $\mathbb{E}|T_j|^2 < \infty$). Then*

1. *For every $x \in B_q^2(r)$, $g(x) = \mathbb{E}\hat{g}_j(x)$.*
2. *For every $x \in B_q^2(r)$, $\mathbb{E}(g(x) - \hat{g}(x))^2 \leq \frac{C\beta_{g,k}^2}{N}$.*
3. *There is a constant C depending on l such that for any probability distribution ζ over $B_q^2(r)$,*

$$\mathbb{E} \int \left(g(x) - \hat{g}(x)\right)^2 \zeta(dx) \leq \frac{C\beta_{g,k}^2}{N} .$$

Therefore there exists a configuration a choice of $(T_j, \omega_j, \theta_j)$ *such that*

$$\int \big(g(x) - \hat{g}(x)\big)^2 \zeta(dx) \leq \frac{C\beta_{g,k}^2}{N}\,.$$

Proof The first item follows from Theorem 13. For the second item, let $x \in B_q^2(r)$. Since $\hat{g}_j(x)$ is an unbiased estimator for $g(x)$ as shown in Item 1, we conclude that :

$$\mathbb{E}\big(g(x) - \hat{g}(x)\big)^2 = \frac{1}{N}\left[\mathbb{E}\big(\hat{g}_j(x)\big)^2 - \big(g(x)\big)^2\right] \leq \frac{1}{N}\mathbb{E}\big(\hat{g}_j(x)\big)^2$$

Now $|\hat{g}_j(x)| \leq \beta_{g,k}(1 + |T_j| + w_0)$. Squaring and taking expectations on both sides yields the result, using that $\mathbb{E}|T_j|^2 < \infty$ since $l \geq 2$. For Item 3, we use Fubini's theorem and Item 2 to conclude that

$$\mathbb{E}\int (g(x) - \hat{g}(x))^2 \zeta(dx) \leq \frac{C\beta_{g,k}^2}{N}\,.$$

The desired bound holds in expectation, so it must also hold for some configuration. ■

Note that in Theorem 16, the RHS of the error upper bounds depend on the Fourier norm $C_g^{(s)}$. As explained in Section 1.1, in order to apply the corrective mechanism we need to consider $g^{\mathsf{rem}}(x) = g(x) - \hat{g}(x)$ for $x \in B_q^2(r)$ and show that, roughly, the corresponding Fourier norm $C_{g^{\mathsf{rem}}}^{(s)} \leq C\frac{C_g^{(s)}}{\sqrt{N}}$. Since Fourier transform is a linear mapping, an unbiased estimator for g (i.e, $\hat{g}$) should be such that the Fourier transform of $\hat{g}$ (i.e, $\hat{G}(\xi)$) is an unbiased estimator for $G(\xi)$ for every $\xi \in \mathbb{R}^q$. There are several technical roadblocks to this argument:

1. $\hat{g}(x)$ is only an unbiased estimator when $x \in B_q^2(r)$.
2. $\hat{g}_j(x)$ is a 'one dimensional function' - that is it depends only on $\langle \omega_j, x\rangle$. This makes its Fourier transform contain tempered distributions like dirac delta and we cannot apply a variance computation to show that the Fourier transform contracts by $1/\sqrt{N}$.
3. $\hat{g}_j(x)$, even along the direction $\langle \omega_j, x\rangle$ is not well behaved since $\mathsf{SReLU}_k(\cdot)$ is not compactly supported. Therefore this is not an L^1 function and hence its Fourier transform isn't very well behaved.

We resolve the issues above by considering the fact that we only care about the values of g (and $\hat{g}$) in $B_q^2(r)$ and hence we are free to modify g (and $\hat{g}$) outside this domain. Along these lines, we modify g to $g(\,\cdot\,; R)$ and $\hat{g}_j$ to $\hat{g}_j(\,\cdot\,; R)$. Ultimately, we will show the existence of $g^{\mathsf{rem}} : \mathbb{R}^q \to \mathbb{R}$ such that $g^{\mathsf{rem}}(x) = g(x) - \hat{g}(x)$ whenever $x \in B_q^2(r)$ and such that its Fourier transform is 'well behaved enough' to carry out the corrective mechanism describe above and in Section 1.1. As a first step towards modification, we resolve item 3 first above by replacing SReLU_k by smoothed triangles $\mathsf{S}\Delta_k$ as defined below. This compactifies $\hat{g}_j$ along the direction ω_j.

Step 3: Replacing SReLU by smoothed triangles. In the notations used below, we hide the dependence on r, w_0, α_0 and l for the sake of clarity. Consider the statement of Theorem 13 for every $x \in B_q^2(r)$:

$$g(x) = \beta_{g,k} \iint \eta(T; r, \omega) \mathsf{SReLU}_k \left(\tfrac{\langle \omega, x \rangle}{r\|\omega\|} - T \right) \mu_l(dT) \nu_{g,k}(d\omega) \,. \tag{18}$$

For $t \in \mathbb{R}$, let

$$\mathsf{S}\Delta_k(t; T) := \mathsf{SReLU}_k(t - T) - 2\mathsf{SReLU}_k(t - 1 - w_0) + \mathsf{SReLU}_k(t - 2 - 2w_0 + T) \,,$$

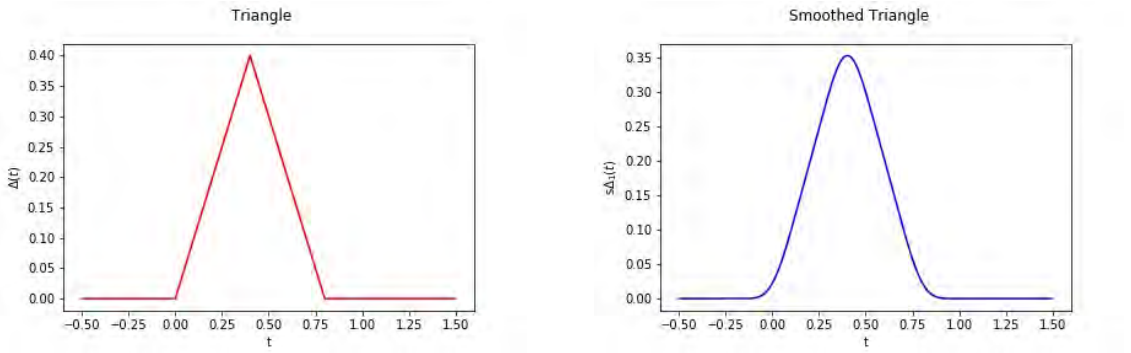
and

$$\Delta(t; T) := \mathsf{ReLU}(t - T) - 2\mathsf{ReLU}(t - 1 - w_0) + \mathsf{ReLU}(t - 2 - 2w_0 + T) \,.$$

Note that $\Delta = \mathsf{S}\Delta_0$. Clearly, when $T \leq 1 + w_0$ and $x \in B_q^2(r)$, we have

$$\mathsf{SReLU}_k\left(\tfrac{\langle \omega, x \rangle}{r\|\omega\|} - T \right) = \mathsf{S}\Delta_k\left(\tfrac{\langle \omega, x \rangle}{r\|\omega\|}, T \right) ,$$

and $\eta(T; r, \omega) = 0$ whenever $T > 1 + w_0$. Therefore, we can replace SReLU_k with $\mathsf{S}\Delta_k$ in Equation (18). When $T \leq 1 + w_0$, $\Delta(\cdot\,; T) : \mathbb{R} \to \mathbb{R}$ gives a triangle graph as can be easily verified and hence is compactly supported. Its Fourier transform is an L^1 function. $\mathsf{S}\Delta_k$ is obtained by convolving Δ with the filter $\lambda_{k,w_0}^{\alpha_0}$. We refer to Figure 2 for an illustration. Lemma 17 below follows from the preceding discussion.



(a) Δ (b) $\mathsf{S}\Delta_1$

Figure 2: Illustrating Δ and $\mathsf{S}\Delta_1$ activation functions.

Lemma 17 *For every $x \in B_q^2(r)$,*

$$g(x) = \beta_{g,k} \iint \eta(T; r, \omega) \mathsf{S}\Delta_k \left(\tfrac{\langle \omega, x \rangle}{r\|\omega\|}, T \right) \mu_l(dT) \nu_{g,k}(d\omega) \,.$$

Consider the technical issues listed before Step 3. We resolved item 3 in Step 3 above. In Step 4 below, will resolve item 2 by modifying $\hat{g}_j$ by first replacing SReLU_k with $\mathsf{S}\Delta_k$ as in Step 3 to 'compactify' it along the direction ω_j and then 'mollify' it along the perpendicular directions by multiplying it with a function which is 1 in $B_q^2(r)$ and vanishes outside a compact set to obtain $\hat{g}_j(\cdot\,; R)$. To resolve item 1, we define $g(x; R)$ to be the expectation of $\hat{g}_j(x; R)$ for every x. As a consequence we have show that for $\xi \in \mathbb{R}^q$, the fourier transform of $\hat{g}_j(\cdot\,; R)$, given by $\hat{G}_j(\xi; R)$ is an unbiased estimator for $G(\xi; R)$ which is the Forier transfrom of $g(\cdot\,; R)$.

Step 4: Truncation and modification Let $\gamma \in \mathcal{S}(\mathbb{R})$ be the function defined in Section 1.3 - such that $\gamma(t) \geq 0$ for every $t \in \mathbb{R}$, $\gamma(t) = 0$ when $|t| \geq 2$ and $\gamma(t) = 1$ for every $t \in [-1, 1]$. Let $R \geq r$ and $q > 1$. For every $x \in \mathbb{R}^d$ and $\omega \neq 0$, we define $\gamma_\omega^\perp(x) := \gamma\left(\frac{\|x\|^2 - \frac{1}{\|\omega\|^2}\langle\omega,x\rangle^2}{R^2}\right)$ when $q > 1$. We use the convention that when $\omega = 0$, $\frac{1}{\|\omega\|^2}\langle\omega, x\rangle^2 := 0$ as stated in Remark 14. When $q = 1$, we let $\gamma_\omega^\perp(x) := 1$ for every x. Let $l \geq 2$. Draw (T_j, ω_j) i.i.d. from the distribution $\mu_l \times \nu_{g,k}$ and let the random variable θ_j be as in Equation (17). Define $\hat{g}_j(\,\cdot\,; R) : \mathbb{R}^q \to \mathbb{R}$:

$$\hat{g}_j(x; R) := \begin{cases} 0 & \text{when } T_j > 1 + w_0 \\ \beta_{g,k}\theta_j \mathsf{S}\Delta_k\left(0, T_j\right)\gamma(\frac{\|x\|^2}{R^2}) & \text{othwerwise when } \omega_j = 0 \\ \beta_{g,k}\theta_j \mathsf{S}\Delta_k\left(\frac{\langle\omega_j,x\rangle}{r\|\omega_j\|}, T_j\right)\gamma_{\omega_j}^\perp(x) & \text{otherwise}\,. \end{cases} \tag{19}$$

We also define $g(\,\cdot\,; R) : \mathbb{R}^q \to \mathbb{R}$ by

$$g(x; R) = \mathbb{E}\hat{g}_j(x; R)\,. \tag{20}$$

The expectation on the RHS exists for every x whenever $l \geq 2$ because then $\mathbb{E}_{T\sim\mu_l}|T| < \infty$. We note that $g(x; R)$ and $\hat{g}_j(x; R)$ are both implicitly dependent on k, l, α_0, w_0. Let $\hat{G}_j(\xi; R)$ be the Fourier transform of $\hat{g}_j(x; R)$ and let $G(\xi; R)$ be the Fourier transform of $g(x; R)$. Even though we allowed the Fourier distribution $G/(2\pi)^d$ to be singular entities like δ measures, we will see that for our extension, we show below that the $G(\xi; R)$ is a $L^1(\mathbb{R}^q) \cap C(\mathbb{R}^q)$ function. This allows us to construct estimators for $G(\,\cdot\,; R)$. In the lemma below we construct an unbiased estimator for $g(\,\cdot\,; R)$, whose Fourier transform is an unbiased estimator for $G(\,\cdot\,; R)$.

Let $\Gamma_{q,R}$ be the Fourier transform of $\gamma(\|x\|^2/R^2)$. We conclude from spherical symmetry of the function $\gamma(\|x\|^2/R^2)$ that $\Gamma_{q,R}(\xi)$ is a function of $\|\xi\|$ only. When convenient, we will abuse notation and replace $\Gamma_{q,R}(\xi)$ by $\Gamma_{q,R}(\|\xi\|)$. We note some useful identities in Lemma 18 and give its proof in Section E.

Lemma 18 *Let μ_l be the probability measure defined in Theorem 11. Let $l \geq 2$ so that $\mathbb{E}_{T\sim\mu_l}T^2 < \infty$.*

1. *For every $x \in B_q^2(r)$,*

$$g(x; R) = g(x) \quad \text{and} \quad \hat{g}_j(x; R) = \hat{g}_j(x)\,,$$

where $\hat{g}_j(x)$ is as defined in (17).

2. *$\hat{g}_j(\,\cdot\,; R) \in L^1(\mathbb{R}^q)$ almost surely and $g(\,\cdot\,; R) \in L^1(\mathbb{R}^q)$*

3. *For every $\xi, \omega \in \mathbb{R}^q$ such that $\omega \neq 0$ we define $\xi_\omega := \frac{\langle\xi,\omega\rangle}{\|\omega\|} \in \mathbb{R}$ and $\xi_\omega^\perp := \xi - \frac{\omega\langle\xi,\omega\rangle}{\|\omega\|^2}$. For any fixed value of T_j and ω_j:*

$$\hat{G}_j(\xi; R) = \begin{cases} 0 \text{ if } T_j > 1 + w_0 \\ \beta_{g,k}\theta_j\Gamma_{q,R}(\|\xi\|)\mathsf{S}\Delta_k(0; T_j) \text{ when } T_j \leq 1 + w_0 \text{ and } \omega_j = 0 \\ \beta_{g,k}\theta_j\Gamma_{q-1,R}(\|\xi_{\omega_j}^\perp\|)\Lambda_{k,w_0}^{\alpha_0}(\xi_{\omega_j})\left[\frac{4e^{i(1+w_0)r\xi_{\omega_j}}}{\xi_{\omega_j}^2 r}\sin^2((1 + w_0 - T)\xi_{\omega_j} r/2)\right] \\ \text{otherwise} \end{cases} \tag{21}$$

Here we stick to the convention that RHS is $\frac{\langle \omega, x \rangle}{\|\omega\|} = 0$ *when* $\omega_j = 0$ *and when* $q = 1$, *we let* $\Gamma_{q-1,R}(\cdot) = 1$. *We recall that* $\Lambda^{\alpha_0}_{k,w_0}$ *is the Fourier transform of the filter* $\lambda^{\alpha_0}_{k,w_0}$.

4. $\hat{G}_j(\,\cdot\,; R) \in L^1(\mathbb{R}^q)$ *almost surely,* $G(\,\cdot\,; R) \in L^1(\mathbb{R}^q)$ *and for every* $\xi \in \mathbb{R}^q$,

$$G(\xi; R) = \mathbb{E}\hat{G}_j(\xi; R)\,.$$

Step 5: Controlling Fourier norm of remainder term. As per Theorem 16, $g(x)$ is approximated by $\frac{1}{N}\sum_{j=1}^N \hat{g}_j(x)$ up to a squared error of the order $\frac{1}{N}$ and $\frac{1}{N}\sum_{j=1}^N \hat{g}_j(x)$ is the output of a two-layer SReLU_k network with N non-linear activation functions. We will now consider the remainder term: $g(x) - \frac{1}{N}\sum_{j=1}^N \hat{g}_j(x)$. Since we are only interested in $x \in B^2_q(r)$, we can define the following version of the remainder term using the truncated functions $g(x; R)$ and $\hat{g}_j(x; R)$:

$$g^{\mathsf{rem}}(x) := g(x; R) - \frac{1}{N}\sum_{j=1}^N \hat{g}_j(x; R)\,.$$

We will now show that the expected 'Fourier norm' of $g^{\mathsf{rem}}(x)$ is smaller by an order of $\frac{1}{\sqrt{N}}$. We note that g^{rem} is a 'random function' such that $\mathbb{E}g^{\mathsf{rem}}(x) = 0$ for every x. Let G^{rem} be the Fourier transform of g^{rem}.

Lemma 19 *Recall the probability measure* μ_l *from Theorem 11. Let* $l = 3$ *so that* $\mathbb{E}_{T\sim\mu_l} T^4 < \infty$ *and let* $R = r$. *For* $s \in \{0\} \cup \mathbb{N}$, *consider*

$$C^{(s)}_{g^{\mathsf{rem}}} := \int_{\mathbb{R}^q} \|\xi\|^s \cdot |G^{\mathsf{rem}}(\xi)| d\xi\,.$$

Whenever $k \geq \max(1, \frac{q-3}{4})$ *and* $s < \frac{3-q}{2} + 2k$, *we have that*

$$\mathbb{E}C^{(s)}_{g^{\mathsf{rem}}} \leq \frac{C(C^{(0)}_g + C^{(2k+2)}_g)}{\sqrt{N}}\,,$$

where C *is a constant depending only on* s, r, q *and* k.

We give the proof in Section E. It is based on Item 4 in Lemma 18, which ensures that $|G^{\mathsf{rem}}|$ is of the order $\frac{1}{\sqrt{N}}$ in expectation. The technical part of the proof involves controlling the integral with respect to the Lebesgue measure using a polar decomposition.

We now combine the results above to complete the proof of Theorem 7. The proof applies Markov's inequality to the results in Theorem 16 and Lemma 19. Let $\hat{g}$ and g^{rem} be defined randomly as in the discussion above. By Markov's inequality:

1. There is a constant C' such that with probability at least $3/4$,

$$\int (g(x) - \hat{g}(x))^2 \zeta(dx) \leq \frac{C'\left(C^{(0)}_g + C^{(2k+2)}_g\right)^2}{N}\,.$$

2. There is a constant C_1' such that with probability at least $3/4$,

$$C_{g^{\text{rem}}}^{(s)} \leq \frac{C_1'(C_g^{(0)} + C_g^{2k+2})}{\sqrt{N}}\,.$$

By the union bound, with probability at least $1/2$ both the inequalities above hold, and hence these must hold for some configuration.

Appendix C. Integral Representations for Cosine Functions

The objective of this section is to prove Theorem 11.

The Lemmas 20 and 21 below establish important properties of the the filter $\lambda_{k,w_0}^{\alpha_0}$ and will be used extensively in the sequel. Their proofs are given in Section C.

Lemma 20 *$\lambda_{k,w_0}^{\alpha_0}(t)$ as defined in Equation (14) is a symmetric, continuous probability density over $\mathbb{R}$ which is supported over $[-w_0, w_0]$. Its Fourier transform $\Lambda_{k,w_0}^{\alpha_0}$ is such that $1 \geq \Lambda_{k,w_0}^{\alpha_0}(\xi) > 0$ for every ξ.*

Proof The first part of the Lemma follows directly from the definition. Let

$$C_{\alpha_0} := \int_{-\infty}^{\infty} \cos(\alpha_0 T)\lambda_{k,w_0}(T)dT > 0\,.$$

For the second part, we observe that

$$\begin{aligned}\Lambda_{k,w_0}^{\alpha_0}(\xi) &= \frac{1}{2C_{\alpha_0}}\left[\Lambda_k\left(\tfrac{(\xi+\alpha_0)w_0}{k}\right) + \Lambda_k\left(\tfrac{(\xi-\alpha_0)w_0}{k}\right)\right] \\ &= \frac{1}{2C_{\alpha_0}}\left[\frac{\sin^{2k}\left(\frac{(\xi+\alpha_0)w_0}{2k}\right)}{\left(\frac{(\xi+\alpha_0)w_0}{2k}\right)^{2k}} + \frac{\sin^{2k}\left(\frac{(\xi-\alpha_0)w_0}{2k}\right)}{\left(\frac{(\xi-\alpha_0)w_0}{2k}\right)^{2k}}\right]\,.\end{aligned}$$

We observe that this vanishes only when both $\sin^{2k}\left(\frac{(\xi+\alpha_0)w_0}{2k}\right)$ and $\sin^{2k}\left(\frac{(\xi-\alpha_0)w_0}{2k}\right)$ vanish. This can happen only if $\alpha_0 = \frac{l\pi k}{w_0}$ for some $l \in \mathbb{Z}$. Since by assumption we have $0 < \alpha_0 < \frac{\pi k}{2w_0}$, this condition cannot hold, which implies the result. ■

Lemma 21 *Let α_0 and w_0 be fixed. Then, there exist constants $C_0, C_1 > 0$ depending only on α_0 and w_0 and C_2 depending only on α_0, w_0 and k such that for every $\xi \in \mathbb{R}$,*

$$\frac{C_0}{C_1 + \max((\frac{\xi}{\alpha_0} - 1)^{2k}, (\frac{\xi}{\alpha_0} + 1)^{2k})} \leq \Lambda_{k,w_0}^{\alpha_0}(\xi) \leq \frac{C_2}{1+\xi^{2k}}\,. \tag{22}$$

For every $i \in \mathbb{N}$, denoting the i times differentiation operator by $D^{(i)}$,

$$\left|D^{(i)}\left[\frac{1}{\Lambda_{k,w_0}^{\alpha_0}(\xi)}\right]\right| \leq C(i,k,w_0,\alpha_0)\left(1+\left|\xi\right|^{2k}\right)\,.$$

For every $\xi \in \mathbb{R}$ and $i \in \mathbb{N}$ there is a constant $C_1(i,k,w_0,\alpha_0)$ such that

$$\left|D^{(i)}\Lambda_{k,w_0}^{\alpha_0}(\xi)\right| \leq \frac{C_1(i,k,w_0,\alpha_0)}{1+\xi^{2k}}\,.$$

Proof Let $\theta \leq \frac{\pi}{4}$. Define $\eta(x) := \frac{\sin^{2k}(x+\theta)}{(x+\theta)^{2k}} + \frac{\sin^{2k}(x-\theta)}{(x-\theta)^{2k}}$. We will use the following claim.

Claim 1 *Let $\theta \in [0, \frac{\pi}{4}]$. Then for every $x \in \mathbb{R}$, either $\sin^{2k}(x+\theta) \geq \sin^{2k}(\theta)$ or $\sin^{2k}(x-\theta) \geq \sin^{2k}(\theta)$.*

Proof of claim: It is sufficient to show this for $x \in [0, \pi)$ because of periodicity. If $x \leq \pi - 2\theta$ then, $\theta \leq x + \theta \leq \pi - \theta$. Therefore, $\sin^{2k}(x+\theta) \geq \sin^{2k}(\theta)$. If $x > \pi - 2\theta$ then $\pi - \theta > x - \theta > \pi - 3\theta \geq \theta$. Therefore, $\sin^{2k}(x-\theta) \geq \sin^{2k}(\theta)$. ■

Clearly,

$$\begin{aligned}
\eta(x) &\geq \frac{\sin^{2k}(x+\theta)}{\sin^{2k}(x+\theta)+(x+\theta)^{2k}} + \frac{\sin^{2k}(x-\theta)}{\sin^{2k}(x-\theta)+(x-\theta)^{2k}} \\
&\geq \min\left(\frac{\sin^{2k}(\theta)}{\sin^{2k}(\theta)+(x-\theta)^{2k}}, \frac{\sin^{2k}(\theta)}{\sin^{2k}(\theta)+(x+\theta)^{2k}}\right) \\
&= \frac{\sin^{2k}(\theta)}{\sin^{2k}(\theta)+\max((x-\theta)^{2k}, (x+\theta)^{2k})}\,. \qquad (23)
\end{aligned}$$

In the second step we have used Claim 1. We note that when $\theta = \frac{\alpha_0 w_0}{2k}$, $\Lambda^{\alpha_0}_{k,w_0}(\xi) = \frac{c_0}{2}\eta(\frac{\xi w_0}{2k})$ where $c_0 = \frac{1}{\int_{-\infty}^{\infty}\cos(\alpha_0 T)\lambda_{k,w_0}(T)dT} \geq 1$. From equation (23), we conclude that

$$\begin{aligned}
\Lambda^{\alpha_0}_{k,w_0}(\xi) &\geq \frac{c_0}{2}\frac{\sin^{2k}(\frac{\alpha_0 w_0}{2k})/(\frac{\alpha_0 w_0}{2k})^{2k}}{\sin^{2k}(\frac{\alpha_0 w_0}{2k})/(\frac{\alpha_0 w_0}{2k})^{2k} + \max((\frac{\xi}{\alpha_0}-1)^{2k}, (\frac{\xi}{\alpha_0}+1)^{2k})} \\
&\geq \frac{1}{2}\frac{\sin^{2k}(\frac{\alpha_0 w_0}{2k})/(\frac{\alpha_0 w_0}{2k})^{2k}}{\sin^{2k}(\frac{\alpha_0 w_0}{2k})/(\frac{\alpha_0 w_0}{2k})^{2k} + \max((\frac{\xi}{\alpha_0}-1)^{2k}, (\frac{\xi}{\alpha_0}+1)^{2k})}\,. \qquad (24)
\end{aligned}$$

In the second step we have used the fact that $c_0 \geq 1$. Now, using Taylor's theorem, we conclude that when $0 \leq x \leq \frac{\pi}{2}$, $\frac{\sin x}{x} \geq 1 - \frac{x^2}{6}$. Therefore,

$$\lim_{k\to\infty} \sin^{2k}(\tfrac{\alpha_0 w_0}{2k})/(\tfrac{\alpha_0 w_0}{2k})^{2k} = 1\,.$$

Using this, we conclude that we can bound $\sin^{2k}(\frac{\alpha_0 w_0}{2k})/(\frac{\alpha_0 w_0}{2k})^{2k}$ away from 0, uniformly for all k. Using this in the Equation (24), we conclude the first part of the lemma. Now, we will consider the derivatives. We first show the following claim:

Claim 2 *Let $f \in C^\infty(\mathbb{R})$ such that $f(x) \neq 0$ for every $x \in \mathbb{R}$. Then, for any $i \geq 1$ $D^{(i)}(\frac{1}{f})$ is a linear combination of the functions of the form $\frac{1}{f^{r+1}}\prod_{l=1}^{r} D^{(n_l)}(f)$, where $1 \leq r \leq i$, $n_l \in \mathbb{N}$, and $\sum_{l=1}^{r} n_l = i$. The coefficients in the linear combination do not depend on f.*

Proof of claim: We show this using induction with base case $D^{(1)}\frac{1}{f} = -\frac{1}{f^2}D^{(1)}f$, which satisfies the hypothesis. Suppose the hypothesis is true for $D^{(i)}\frac{1}{f}$. Then $D^{(i+1)}\frac{1}{f}$ is a linear combination of functions of the form $D^{(1)}\left(\frac{1}{f^{r+1}}\prod_{l=1}^{r} D^{(n_l)}(f)\right)$, where $1 \leq r \leq i$, $r_l \in \mathbb{N}$, and $\sum_{l=1}^{r} n_l = i$. Now,

$$\begin{aligned}
D^{(1)}\left(\frac{1}{f^{r+1}}\prod_{l=1}^{r} D^{(n_l)}(f)\right) &= -\frac{r+1}{f^{r+2}}D^{(1)}(f)\prod_{l=1}^{r} D^{(n_l)}(f) \\
&\quad + \frac{1}{f^{r+1}}\sum_{l_0=1}^{r} D^{(n_{l_0}+1)}(f)\prod_{l\neq l_0} D^{(n_l)}(f)\,.
\end{aligned}$$

This is a linear combination with the required property for $i+1$. Therefore, we conclude the claim. ■

We now show another estimate necessary for the proof:

Claim 3 *For every $i \in \mathbb{N}$ and some constant $C(i,k,w_0,\alpha_0) > 0$ depending only on i,k,w_0,α_0,*

$$|D^{(i)}\Lambda^{\alpha_0}_{k,w_0}(\xi)| \leq \frac{C(i,k,w_0,\alpha_0)}{(1+|\xi|^{2k})}$$

Proof of claim: Let $g(\xi) = \frac{\sin^{2k}(\xi)}{\xi^{2k}}$. Since $\Lambda^{\alpha_0}_{k,w_0}$ is a linear combination of the scaled and shifted version of g, the same bounds hold for $\Lambda^{\alpha_0}_{k,w_0}$ up to constants depending on k,w_0,α_0 and i. Clearly, $g \in C^\infty(\mathbb{R})$. Therefore, $|D^{(i)}(g)(\xi)| \leq C(i)$ whenever $|\xi| \leq 1$. Now assume that $|\xi| \geq 1$. It is easy to show that $D^{(i)}(g)$ is a linear combination of the functions of the form $\frac{g_r(\xi)}{\xi^{2k+r}}$, where $g_r(\xi)$ is a bounded trigonometric function, and $r \in \{0,1,\ldots,i\}$. Therefore, $|D^{(i)}(g)(\xi)| \leq \frac{C'(i)}{|\xi|^{2k}} \leq \frac{2C'(i)}{1+|\xi|^{2k}}$ whenever $|\xi| \geq 1$. Combining this with the case $|\xi| \leq 1$, we conclude the result. ■

From Claim 2, it is sufficient to upper bound terms of the form $|\frac{1}{f^{r+1}}\prod_{l=1}^{r} D^{(n_l)}(f)|$, where $1 \leq r \leq i$, $n_l \in \mathbb{N}$, and $\sum_{l=1}^{r} n_l = i$ for $f = \Lambda^{\alpha_0}_{k,w_0}$. From the bound in Equation (22) on $\Lambda^{\alpha_0}_{k,w_0}$ and bounds on the derivatives in Claim 3, we have

$$\left|\frac{1}{f^{r+1}}\prod_{l=1}^{r} D^{(n_l)}(f)\right|(\xi) \leq C(k,i,w_0,\alpha_0)(1+|\xi|^{2k})\,.$$

From this we conclude the upper bound on the derivatives. The proof of upper bound on $\Lambda^{\alpha_0}_{k,w_0}$ is similar to the proof of Claim 3 and the bounds on $D^{(i)}\Lambda^{\alpha_0}_{k,w_0}$ follows from Claim 3. This completes the proof of Lemma 21. ■

Let $C^\infty_c(\mathbb{R})$ denote the set of infinitely differentiable, compactly supported real valued functions. Let p be any symmetric continuous probability density supported over $[-w_0,w_0]$. Define

$$\mathsf{SReLU}(t) = \int_{-\infty}^{\infty} \mathsf{ReLU}(t-T)p(T)dT\,. \tag{25}$$

We also define the convolution operator $\mathcal{P} : C^0(\mathbb{R}) \to C^0(\mathbb{R})$ by

$$\mathcal{P}g(t) := \int_{-\infty}^{\infty} g(t-T)p(T)dT\,,$$

and let $\mathcal{I}$ denote the identity operator over $C^0(\mathbb{R})$.

Lemma 22 *Let $h \in C^\infty_c(\mathbb{R})$ function such that $\mathsf{supp}(h) \subseteq [a,b]$ for some $a,b \in \mathbb{R}$. Then*

1. *For any $t \in [a,b]$,*

$$h(t) = \int_{-\infty}^{\infty} h''(T)\mathsf{ReLU}(t-T)dT\,.$$

2. Let SReLU *be as defined in Equation* (25)*. For every* $n \in \mathbb{N}$*,*

$$h(t) = \int_{-\infty}^{\infty} h''(T) \left[(\mathcal{I} - \mathcal{P})^{n+1}\right] \mathsf{ReLU}(t - T)dT + \sum_{i=0}^{n} \int_{-\infty}^{\infty} h''(T) \left[(\mathcal{I} - \mathcal{P})^{i} \mathsf{SReLU}\right](t - T)dT\,.$$

Proof

1. Since h is infinitely differentiable and supported over $[a, b]$, $\mathsf{supp}(h'') \subseteq [a, b]$. Therefore, the integral in question reduces to:

$$\int_a^b h''(T)\mathsf{ReLU}(t - T)dT = \int_a^t h''(T)(t - T)dT\,.$$

The proof follows from integration by parts and using the fact that $h'(a) = h(a) = 0$.

2. Since h'' is compactly supported, it is sufficient to show that

$$\sum_{i=0}^{n} \left[(\mathcal{I} - \mathcal{P})^{i} \mathsf{SReLU}\right] + \left[(\mathcal{I} - \mathcal{P})^{n+1}\right] \mathsf{ReLU} = \mathsf{ReLU}\,.$$

Since $\mathsf{SReLU} = \mathcal{P}\left(\mathsf{ReLU}\right)$, this reduces to showing that

$$\sum_{i=0}^{n} \left[(\mathcal{I} - \mathcal{P})^{i}\right] \mathcal{P} + (\mathcal{I} - \mathcal{P})^{n+1} = \mathcal{I}\,,$$

which can be verified via a straightforward induction argument. ■

Lemma 23 *Let* h *be as defined in Lemma 22. Let* P*, the Fourier transform of density* p *be such that* $P(\xi) \in \mathbb{R}$ *for every* ξ *and* $1 \geq P(\xi) > 0$ *for almost all* ξ *(w.r.t lebesgue measure over* $\mathbb{R}$*). Then for every* $t \in [a, b]$ *the following limit holds uniformly.*

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} h''(T) \left[(\mathcal{I} - \mathcal{P})^{n+1} \mathsf{ReLU}\right](t - T)dT = 0\,.$$

And for every $t \in [a, b]$ *the following holds uniformly:*

$$h(t) = \lim_{n\to\infty} \sum_{i=0}^{n} \int_{-\infty}^{\infty} \left[(\mathcal{I} - \mathcal{P})^{i} h''\right](T)\mathsf{SReLU}(t - T)dT\,.$$

Proof Fix $t \in [a, b]$. By a simple application of Fubini's theorem, the fact that h'' has compact support and that $p(\cdot)$ is compactly supported, it is easy to show the following "self-adjointness" of the operator $\mathcal{P}$. For any continuous $f : \mathbb{R} \to \mathbb{R}$:

$$\int_{-\infty}^{\infty} h''(T) \left[\mathcal{P}f\right](t - T)dT = \int_{-\infty}^{\infty} \left[\mathcal{P}h''\right](T) f(t - T)dT\,. \tag{26}$$

From Equation (26) it follows that

$$\int_{-\infty}^{\infty} h''(T)\left[(\mathcal{I}-\mathcal{P})^{n+1}\mathsf{ReLU}\right](t-T)dT = \int_{-\infty}^{\infty}\left[(\mathcal{I}-\mathcal{P})^{n}\right]h''(T)\left[(\mathcal{I}-\mathcal{P})\mathsf{ReLU}\right](t-T)dT\,.$$

From the definition of the ReLU and the fact that p is symmetric and of compact support, it is clear that $[(\mathcal{I}-\mathcal{P})]\,\mathsf{ReLU}$ is a continuous function with compact support. $\|\left[(\mathcal{I}-\mathcal{P})\right]\mathsf{ReLU}\|_2 < \infty$ where $\|\cdot\|_2$ is the standard L^2 norm of functions w.r.t Lebesgue measure. Hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\int_{-\infty}^{\infty} h''(T)\left[(\mathcal{I}-\mathcal{P})^{n+1}\mathsf{ReLU}\right](t-T)dT\right| \\
=&\left|\int_{-\infty}^{\infty}\left[(\mathcal{I}-\mathcal{P})^{n}h''\right](T)\left[(\mathcal{I}-\mathcal{P})\mathsf{ReLU}(t-T)\right]dT\right| \\
\leq& \|(\mathcal{I}-\mathcal{P})\mathsf{ReLU}\|_2\|(\mathcal{I}-\mathcal{P})^n h''\|_2 \\
\leq& C\|(\mathcal{I}-\mathcal{P})^n h''\|_2\,, \qquad (27)
\end{aligned}$$

where C is independent of n. To prove the lemma, it is sufficient to show that $\lim_{n\to\infty}\|(\mathcal{I}-\mathcal{P})^n h''\|_2 = 0$. We do this using Parseval's theorem. Let $H^{(2)}$ be the Fourier transform of h''. We note that $H^{(2)} \in L^2$ since $h \in \mathcal{S}(\mathbb{R})$. By the duality of convolution-multiplication with respect to Fourier transform, we conclude that the Fourier transform of $(\mathcal{I}-\mathcal{P})^n h''$ is $(1-P)^n H^{(2)}$. By Plancherel's theorem,

$$\|(\mathcal{I}-\mathcal{P})^n h''\|_2 = \frac{1}{\sqrt{2\pi}}\|(1-P)^n H^{(2)}\|_2\,. \qquad (28)$$

Since $0 < P(\xi) \leq 1$ almost everywhere, we conclude that $\lim_{n\to\infty}(1-P)^n H^{(2)} = 0$ almost everywhere. Since $|(1-P)^n H^{(2)}| \leq |H^{(2)}|$ almost everywhere and $H^{(2)} \in L^2$, we conclude by dominated convergence theorem that

$$\lim_{n\to\infty}\|(\mathcal{I}-\mathcal{P})^n h''\|_2 = \frac{1}{\sqrt{2\pi}}\lim_{n\to\infty}\|(1-P)^n H^{(2)}\| = 0\,.$$

Equation (27) along with item 2 of Lemma 22, this implies that for every $t \in [a,b]$, the following uniform convergence holds:

$$h(t) = \lim_{n\to\infty}\sum_{i=0}^{n}\int_{-\infty}^{\infty} h''(T)\left[(\mathcal{I}-\mathcal{P})^{i}\mathsf{SReLU}\right](t-T)dT\,.$$

Using Equation (26) along with the equation above, we get

$$h(t) = \lim_{n\to\infty}\sum_{i=0}^{n}\int_{-\infty}^{\infty}\left[(\mathcal{I}-\mathcal{P})^{i}h''\right](T)\mathsf{SReLU}(t-T)dT\,.$$

■

In Lemma 24 below, we will show that when we choose the operator $\mathcal{P}$ carefully, the sum $h_n^{(2)} := \sum_{i=0}^{n}(\mathcal{I}-\mathcal{P})^i h''$ converges a.e. and in L^2 to a Schwartz function $\bar{h}:\mathbb{R}\to\mathbb{R}$. The proof is based on standard techniques from Fourier analysis. Let $D^{(n)}$ denote the n-fold differentiation operator over $\mathbb{R}$ and we take $D^{(0)}$ to be the identity operator.

Lemma 24 *Let the filter p and its Fourier transform P be such that*

1. *They obey all the conditions in Lemma 23*
2. $\frac{1}{P} \in C^\infty(\mathbb{R})$
3. $\|D^{(i)}(P)\|_\infty \leq C_i$ *for some constant* C_i.
4. *For every* $n \in \mathbb{N} \cup \{0\}$ *there exists a constant* $C_n > 0$ *such that* $|D^n \frac{1}{P(\xi)}| \leq C_n(1 + \xi^{2m(n)})$ *for some* $m(n) \in \mathbb{N}$

Let $\bar{h}$ be the inverse Fourier transform of $\frac{H^{(2)}}{P}$, where $H^{(2)}$ is the Fourier transform of h''. Then:

1. $\bar{h} \in \mathcal{S}(\mathbb{R})$
2. $(1 + |T|^3)h_n^{(2)}(T) \to (1 + |T|^3)\bar{h}(T)$ *as* $n \to \infty$ *uniformly for all* $T \in \mathbb{R}$
3. *For every* $t \in [a, b]$, h *admits the integral representation*

$$h(t) = \int_{-\infty}^{\infty} \bar{h}(T)\mathsf{SReLU}(t - T)dT\,.$$

Furthermore, the filter $p = \lambda_{k,w_0}^{\alpha_0}$ (defined in Equation (14)) satisfies the above conditions.

Proof Since $h'' \in \mathcal{S}(\mathbb{R})$, we conclude that $H^{(2)} \in \mathcal{S}(\mathbb{R})$ because Fourier transform maps Schwartz functions to Schwartz functions. It is easy to show from definitions that $\frac{H^{(2)}}{P} \in \mathcal{S}(\mathbb{R})$. By definition $\bar{h} := \mathcal{F}^{-1}\left(\frac{H^{(2)}}{P}\right)$ (where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform). Therefore, $\bar{h} \in \mathcal{S}(\mathbb{R})$. We will first show that $h_n^{(2)}(T) \to \bar{h}(T)$ uniformly for every $T \in \mathbb{R}$. By definition of $h_n^{(2)} \in \mathcal{S}(\mathbb{R})$, it is clear that $h_n^{(2)} \in C_c^\infty(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$ and hence its Fourier transform $H_n^{(2)} \in \mathcal{S}(\mathbb{R})$. Since $H_n^{(2)}(\xi) = \sum_{i=1}^n (1 - P(\xi))^i H(\xi)$. Since $0 < P(\xi) \leq 1$ for every $\xi \in \mathbb{R}$ by hypothesis, we conclude that $H_n^{(2)}(\xi) \to \frac{H}{P}(\xi)$ and $|H_n^{(2)}(\xi)| \leq \left|\frac{H}{P}(\xi)\right|$ for every $\xi \in \mathbb{R}$. Therefore, $\left|H_n^{(2)}(\xi) - \frac{H}{P}(\xi)\right| \leq 2\left|\frac{H}{P}(\xi)\right| \in L^1(\mathbb{R})$. From the Fourier inversion formula, the following holds for every $T \in \mathbb{R}$:

$$\begin{aligned}|h_n^{(2)}(T) - \bar{h}(T)| &= \frac{1}{2\pi}\left|\int_{\mathbb{R}} e^{-i\xi T}\left(\frac{H^{(2)}}{P}(\xi) - H_n^{(2)}(\xi)\right)d\xi\right| \\ &\leq \frac{1}{2\pi}\int_{\mathbb{R}}\left|\frac{H^{(2)}}{P}(\xi) - H_n^{(2)}(\xi)\right|d\xi\,.\end{aligned}$$

By the dominated convergence theorem, the integral in the last step converges to 0 as $n \to \infty$ and we conclude that $h_n^{(2)}(T) \to \bar{h}(T)$ uniformly for every T. To show the uniform convergence of $T^3 h_n^{(2)}(T) \to T^3\bar{h}(T)$, we use the duality between multiplication by a polynomial and differentiation under Fourier transform. The Fourier transform of $T^3 h_n^{(2)}(T)$ is

$iD^{(3)}H_n^{(2)}(\xi)$ and that of $T^3\bar{h}(T)$ is $iD^{(3)}\frac{H^{(2)}}{P}$. We proceed just like above. We need to show that $D^{(3)}H_n^{(2)}(\xi) \to D^{(3)}\frac{H^{(2)}}{P}$ for every ξ and that $D^{(3)}H_n^{(2)}(\xi)$ is dominated by a L^1 function uniformly for every n. It is clear that $H_n^{(2)}(\xi) - \frac{H^{(2)}}{P}(\xi) = -\frac{(1-P(\xi))^{n+1}}{P(\xi)}H^{(2)}(\xi)$. Differentiating both sides thrice and applying the product rule, we conclude that $D^{(i)}H_n^{(2)}(\xi) \to D^{(i)}\frac{H^{(2)}}{P}(\xi)$ for every ξ and for every $i \leq 3$. Consider $D^{(3)}\left[\frac{(1-P(\xi))^{n+1}}{P(\xi)}H^{(2)}(\xi)\right]$, we get a finite linear combination of the functions of the form

$$n^r \frac{(1-P)^{n+1-l}}{P^{c_0}} D^{(a)}(H^{(2)}) \prod_{s=1}^{3} D^{(b_s)}(P) \tag{29}$$

for some $c_0, r, l, a, b_s, k \in \mathbb{N} \cup \{0\}$, all of them independent of n and such that $l, r, b_s, a \leq 3$ and $c_0 \leq 4$. To show domination above from a L^1 function, it is sufficient to show that each of terms of the form described in Equation (29). Now, by assumption, $\|D^{(b_s)}(P)\|_\infty \leq C$ for some constant C. $\frac{1}{P^{c_0}(\xi)} \leq C(1+|\xi|^{2m(0)})^4$ (where $m(0)$ is as given in the conditions of the lemma and $c_0 \leq 4$ as given above) and $D^{(a)}H^{(2)} \in \mathcal{S}(\mathbb{R})$. It is therefore sufficient to show that $n^r(1-P)^{n+l-1}$ is dominated by a fixed polynomial in $|\xi|$ for every n large enough. Indeed, for $n \geq 3$, we have

$$\begin{aligned}
n^r(1-P(\xi))^{n-l+1} &\leq n^r(1-P(\xi))^{n-2} \\
&\leq n^r e^{-P(\xi)(n-2)} \\
&\leq e^2 n^r e^{-P(\xi)n} \\
&\leq e^2 \sup_{x \geq 0} x^r e^{-P(\xi)x} \\
&= \frac{e^2 r^r e^{-2}}{P(\xi)^r} \\
&\leq C(1+|\xi|^{2m(0)})^3 .
\end{aligned}$$

Here we have used the fact that $r \leq 3$. Therefore, the remainder term for each n is uniformly dominated by a product of a polynomial of ξ and a Schwartz function. Therefore, we conclude that the sequence $H_n^{(2)}$ is dominated by a L^1 function and from the discussion above conclude that $(1+|T|^3)h_n^{(2)}(T) \to (1+|T|^3)h(T)$ uniformly for every $T \in \mathbb{R}$. To show the final result, we apply Lemma 23 for $t \in [a,b]$ to obtain

$$h(t) = \int_{-\infty}^{\infty} h_n^{(2)}(T)\mathsf{SReLU}(t-T)dT + o_n(1) ,$$

where $o_n(1)$ tends to 0 uniformly for all $t \in [a,b]$. Plugging in this expression for $h(t)$ yields

$$\left| h(t) - \int_{-\infty}^{\infty} \bar{h}(T)\mathsf{SReLU}(t-T)dT \right| = \left| \int_{-\infty}^{\infty} (h_n^{(2)}(T) - \bar{h}(T))\mathsf{SReLU}(t-T)dT \right| + o_n(1)$$

which we upper bound by

$$\begin{aligned}
&\leq \int_{-\infty}^{\infty} \left|h_n^{(2)}(T) - \bar{h}(T)\right| \mathsf{SReLU}(t-T)dT + o_n(1) \\
&= \int_{-\infty}^{\infty} (1+|T|^3)\left|h_n^{(2)}(T) - \bar{h}(T)\right| \frac{\mathsf{SReLU}(t-T)}{1+|T|^3}dT + o_n(1) \\
&\leq \|(1+|\eta|^3)\left|h_n^{(2)}(\eta) - \bar{h}(\eta)\right|\|_\infty \int_{-\infty}^{\infty} \frac{\mathsf{SReLU}(t-T)}{1+|T|^3}dT + o_n(1)
\end{aligned}$$

Now using the fact that $|\mathsf{SReLU}(s)| = \int_{-w_0}^{w_0} \mathsf{ReLU}(s-\tau)p(\tau)d\tau \leq |s| + w_0$ for every $s \in \mathbb{R}$, the above is bounded as

$$\begin{aligned}
&\leq \|(1+|\eta|^3)\left|h_n^{(2)}(\eta) - \bar{h}(\eta)\right|\|_\infty \int_{-\infty}^{\infty} \frac{b+|T|+w_0}{1+|T|^3}dT + o_n(1) \\
&= \|(1+|\eta|^3)\left|h_n^{(2)}(\eta) - \bar{h}(\eta)\right|\|_\infty C + o_n(1) \\
&\to 0\,.
\end{aligned}$$

It is simple to verify that $\lambda_{k,w_0}^{\alpha_0}$ satisfies all the conditions of the lemma using the results from Lemma 21. ■

We will now specialize to the filter defined in Section A and set $p := \lambda_{k,w_0}^{\alpha_0}$ as defined in Equation (14) for some $k \in \mathbb{N} \cup \{0\}$. We denote the activation function obtained as SReLU_k, in keeping with the notation defined in Section A. A well known result from analysis shows the existence of a "bump function" $\gamma \in C_c^\infty(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$ such that $\gamma(t) = 1$ when $|t| \leq 1$, $\gamma(t) = 0$ when $|t| \geq 2$ and $\gamma(t) \geq 0$ for every $t \in \mathbb{R}$. Let Γ be the Fourier transform of γ. Henceforth, we let $h(t) = \gamma(t)\cos(\alpha t + \psi)$ for some $\alpha, \psi \in \mathbb{R}$. Clearly $h \in C_c^\infty(\mathbb{R})$. It is clear that for $t \in [-1,1]$, $h(t) = \cos(\alpha t + \psi)$. Therefore, from Lemma 24, we conclude that there exists $\bar{h} \in \mathcal{S}(\mathbb{R})$ such that for every $t \in [-1,1]$,

$$\cos(\alpha t + \psi) = \int_{\mathbb{R}} \bar{h}(T)\mathsf{SReLU}_k(t-T)dT\,. \tag{30}$$

In the following discussion, we will estimate about how 'large' $\bar{h}$ is in terms of α. Let H denote the Fourier transform of h. A simple calculation shows that:

1.
$$H(\xi) = \frac{1}{2}\left[e^{i\psi}\Gamma(\xi+\alpha) + e^{-i\psi}\Gamma(\xi-\alpha)\right] \tag{31}$$

2.
$$H^{(2)}(\xi) = -\frac{\xi^2}{2}\left[e^{i\psi}\Gamma(\xi+\alpha) + e^{-i\psi}\Gamma(\xi-\alpha)\right] \tag{32}$$

Lemma 25 *Let $h(t) = \gamma(t)\cos(\alpha t + \psi)$ and $\bar{h}$ be the corresponding limiting function given by Lemma 24. Then for all $T \in \mathbb{R}$ and $l \in \mathbb{N}$, we have*

$$|(1+T^{2l})\bar{h}(T)| \leq C(k, \alpha_0, w_0, l)(1+|\alpha|^{2k+2})\,.$$

Proof Let $\bar{H}$ be the Fourier transform of $\bar{h}$. By the inversion formula we have that for every T

$$|\bar{h}(T)| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\bar{H}(\xi)| d\xi \,. \tag{33}$$

By Lemma 24, it is clear that $\bar{H}(\xi) = \frac{H^{(2)}(\xi)}{\Lambda^{\alpha_0}_{k,w_0}(\xi)}$. Using Lemma 21, there exists a constant $C(k, \alpha_0, \omega_0)$ such that:

$$\begin{aligned}
|\bar{H}(\xi)| =& \left| \frac{H^{(2)}(\xi)}{\Lambda^{\alpha_0}_{k,w_0}(\xi)} \right| \\
&\leq C(k, \alpha_0, w_0)(1 + |\xi|^{2k})\xi^2 \Big(|\Gamma(\xi - \alpha)| + |\Gamma(\xi + \alpha)| \Big) \\
&\leq C(k, \alpha_0, w_0)(1 + |\xi|^{2k+2}) \Big(|\Gamma(\xi - \alpha)| + |\Gamma(\xi + \alpha)| \Big) \\
&\leq C(k, \alpha_0, w_0)(1 + |\xi|^{2k+2}) \left(\frac{1}{1 + |\xi - \alpha|^{2k+4}} + \frac{1}{1 + |\xi + \alpha|^{2k+4}} \right) .
\end{aligned} \tag{34}$$

We have absorbed universal constants and constants depending only on k into $C(k, \alpha_0, w_0)$ throughout. In the second step we have used the fact that $|\xi|^2 \leq 1 + |\xi|^{2k+2}$ for every $\xi \in \mathbb{R}$ and used the expressions for $H^{(2)}(\xi)$ given in Equation (32). In the last step, we have used the fact that since $\Gamma \in \mathcal{S}(\mathbb{R})$, there exists a constant C_k such that $|\Gamma(\xi)| \leq \frac{C_k}{1+|\xi|^{2k+4}}$ for every $\xi \in \mathbb{R}$. Using Equations (33) and (34) along with an elementary application of Jensen's inequality to the function $x \to |x|^{2k+2}$, we have

$$|\bar{h}(T)| \leq C(k, \alpha_0, w_0) \left(1 + |\alpha|^{2k+2} \right) . \tag{35}$$

To bound $|T^{2l}\bar{h}(T)|$, we consider the derivatives of its Fourier transform. Clearly, the Fourier transform of $T^{2l}\bar{h}(T)$ is $(-1)^l D^{(2l)}\bar{H}(\xi)$. Therefore, for all T, we have from the inversion formula that

$$|T^{2l}\bar{h}(T)| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |D^{(2l)}\bar{H}(\xi)| d\xi \,.$$

Now, $D^{(2l)}\bar{H}(\xi) = D^{(2l)} \left(-\frac{\xi^2}{2\Lambda^{\alpha_0}_{k,w_0}(\xi)} \left[e^{i\psi}\Gamma(\xi + \alpha) + e^{-i\psi}\Gamma(\xi - \alpha) \right] \right)$. Using the product rule here results in a sum of the form

$$\begin{aligned}
&D^{(2l)}\bar{H}(\xi) \\
&= -\frac{1}{2} \sum_{\substack{a,b,c \in \mathbb{Z}^+ \\ a+b+c=2l}} N_{a,b,c} \left(D^{(a)}\xi^2 \right) \left(D^{(b)} \frac{1}{\Lambda^{\alpha_0}_{k,w_0}(\xi)} \right) \left[e^{i\psi} D^{(c)}\Gamma(\xi + \alpha) + e^{-i\psi} D^{(c)}\Gamma(\xi - \alpha) \right]
\end{aligned}$$

for some positive integers $N_{a,b,c}$. We consider each term separately.

Using Lemma 21, we conclude for every a, b in the summation,

$$\left| D^{(a)}\xi^2 D^{(b)} \frac{1}{\Lambda^{\alpha_0}_{k,w_0}(\xi)} \right| \leq C(l, k, \alpha_0, w_0)(1 + |\xi|^{2k+2}) \,.$$

Now, $D^{(c)}\Gamma \in \mathcal{S}(\mathbb{R})$ for every c. Therefore we can find a constant C_k such that $\left|D^{(c)}\Gamma(\xi)\right| \leq \frac{C_k}{1+|\xi|^{2k+4}}$. Therefore, using similar integration as the previous case, we conclude that:

$$|T^{2l}\bar{h}(T)| \leq C(l,k,\alpha_0,w_0)(1+|\alpha|^{2k+2}) \tag{36}$$

Combining equations (36) and (35) we conclude the result. ■

We will now give the proof of Theorem 11 and Lemma 4:

Proof of Theorem 11: From Lemma 24 and Equation (30) we conclude that for every $t \in [-1,1]$:

$$\cos(\alpha t+\psi) = \int_{-\infty}^{\infty} \bar{h}(T)\mathsf{SReLU}_k(t-T)dT\,. \tag{37}$$

For some $\bar{h} \in \mathcal{S}(\mathbb{R})$. From Lemma 25 we conclude that

$$\|(1+T^{2l})\bar{h}(T)\|_\infty \leq C(k,w_0,\alpha_0,l)(1+|\alpha|^{2k+2})\,.$$

Taking $\kappa(T) := \frac{(1+T^{2l})}{c_\mu}\bar{h}(T)$ in Equation (37), we conclude the result. ■

Proof of Lemma 4: The proof follows from an application of Lemma 22 with $h(t) = \gamma(t)\cos(\alpha t+\psi)$. ■

Proof of Theorem 13: From Equation (15) and the definition of $\nu_{g,k}$,

$$g(x) = \int \frac{C_g^{(0)} + r^{2k+2}C_g^{(2k+2)}}{1+r^{2k+2}\|\omega\|^{2k+2}} \cos\left(r\|\omega\|\tfrac{\langle\omega,x\rangle}{r\|\omega\|} + \psi(\omega)\right)\nu_{g,k}(d\omega)\,.$$

We follow the convention that $\frac{\langle\omega,x\rangle}{r\|\omega\|} = 0$ when $\omega = 0$ without loss of meaning in the equation above. When $x \in B_q^2(r)$, Cauchy-Schwarz inequality implies that $\frac{\langle\omega,x\rangle}{r\|\omega\|} \in [-1,1]$. In Theorem 11, we take $\alpha = r\|\omega\|$ and $\psi = \psi(\omega)$ to conclude that there exists a continuous function $\kappa(T;r,\omega)$ such that for every $x \in B_q^2(r)$

$$g(x) = \left(C_g^{(0)} + r^{2k+2}C_g^{(2k+2)}\right)\iint \frac{\kappa(T;r,\omega)}{1+r^{2k+2}\|\omega\|^{2k+2}}\mathsf{SReLU}_k\left(\tfrac{\langle\omega,x\rangle}{r\|\omega\|} - T\right)\mu_l(dT)\nu_{g,k}(d\omega)\,,$$

where $\left|\frac{\kappa(T;r,\omega)}{1+r^{2k+2}\|\omega\|^{2k+2}}\right| \leq C(k,l)$ a.s. In order to make the notation more compact we define

$$\eta(T;r,\omega) := \frac{1}{C(k,l)}\frac{\kappa(T;r,\omega)\mathbb{1}(T\leq 1+w_0)}{1+r^{2k+2}\|\omega\|^{2k+2}}$$

and $\beta_{g,k} := \left(C_g^{(0)} + r^{2k+2}C_g^{(2k+2)}\right)C(k,l)$ (we hide the dependence on l).

The theorem follows from the discussion above when, in the definition of η, the extra factor of $\mathbb{1}(T \leq 1+w_0)$ is removed. However, we note that when $x \in B_q^2(r)$, $\frac{\langle\omega,x\rangle}{r\|\omega\|} \leq 1$ and it follows that when $T > 1+w_0$,

$$\mathsf{SReLU}_k\left(\tfrac{\langle\omega,x\rangle}{r\|\omega\|} - T\right) = 0\,.$$

Therefore, we can include the factor of $\mathbb{1}(T \leq 1+w_0)$ without altering the equality. ■

Appendix D. Neural Network Approximation with Function Independent Sampling

We consider a similar setup as in Section 3. Let $g : \mathbb{R}^q \to \mathbb{R}$ be such that $g \in L^1(\mathbb{R}^q)$ and its Fourier transform $G \in L^1(\mathbb{R}^q) \cap C(\mathbb{R}^q)$. We define the following norms for G:

$$S_g^{(l)} = \sup_{\omega \in \mathbb{R}^q} \|\omega\|^l (1 + \|\omega\|^{q+1}) \frac{|G(\omega)|}{(2\pi)^q} \,. \tag{38}$$

We assume that $S_g^{(l)} < \infty$ for $l = 0, 1, \ldots, L$ for some L to be chosen later. We consider the spherically symmetric probability measure ν_0 over $\mathbb{R}^q$ defined by its Randon-Nikodym derivative: $\nu_0(d\omega) = C_q \frac{d\omega}{1+\|\omega\|^{q+1}}$, where C_q is the normalizing constant.

Remark 26 *We note that G has to be a function and not a generalized function/measure (like dirac delta) for the norms $S_g^{(l)}$ to make sense. Unlike $\nu_{g,k}$, ν_0 depends neither on g nor on k. We intend to draw the weights $\omega_j \sim \nu_0$. Clearly $\omega_j \neq 0$ almost surely. We therefore skip the corner cases for $\omega_j = 0$ as considered in Section B.*

We let μ_l be as defined in Theorem 11. We again consider Equation (15). Assume $S_g^{2k+2}, S_g^0 < \infty$. Suppose $x \in B_q^2(r)$

$$\begin{aligned}
g(x) &= \int_{\mathbb{R}^q} \cos(\langle \omega, x \rangle + \psi(\omega)) \frac{|G(\omega)|}{(2\pi)^q} d\omega \\
&= \int \cos(\langle \omega, x \rangle + \psi(\omega)) \frac{|G(\omega)|}{(2\pi)^q} \frac{(1 + \|\omega\|^{q+1})}{C_q} \nu_0(d\omega) \\
&= \int \frac{|G(\omega)|(1 + \|\omega\|^{q+1})(1 + r^{2k+2}\|\omega\|^{2k+2})}{C_q (2\pi)^q / C(k,l)} \eta(T; r, \omega) \mathsf{SReLU}_k \left(\frac{\langle \omega, x \rangle}{r\|\omega\|} - T \right) \mu_l(dT) \nu_0(d\omega)
\end{aligned}$$

Here we have used Theorem 11 in the third step where $|\eta| \leq 1$ almost surely. For the sake of clarity, we will abuse notation and redefine

$$\eta(T; r, \omega) \leftarrow \frac{|G(\omega)|(1 + \|\omega\|^{q+1})(1 + r^{2k+2}\|\omega\|^{2k+2})}{(S_g^{(0)} + r^{2k+2} S_g^{2k+2})(2\pi)^q} \eta(T; r, \omega) \,.$$

By similar considerations as in Theorem 13, we can replace $\eta(T; r, \omega)$ with $\eta(T; r, \omega)\mathbb{1}(T \leq 1 + w_0)$. Clearly $|\eta| \leq 1$ almost surely even under this redefinition. We will take $\beta_{g,k}^S := \frac{C(k,l)}{C_q}(S_g^0 + r^{2k+2} S_g^{2k+2})$. We conclude that for every $x \in B_2^q(r)$

$$g(x) = \beta_{g,k}^S \int \eta(T; r, \omega) \mathsf{SReLU}_k \left(\frac{\langle \omega, x \rangle}{r\|\omega\|} - T \right) \mu_l(dT) \nu_0(d\omega) \,. \tag{39}$$

For $j \in \{1, \ldots, N\}$, draw (T_j, ω_j) to be i.i.d. from the distribution $\mu_l \times \nu_0$. Let θ_j^u for $j \in [N]$ be i.i.d. $\mathsf{Unif}[-1, 1]$ and independent of everything else. We define

$$\theta_j := \mathbb{1}\left(\theta_j^u < \eta(T_j; r, \omega_j)\right) - \mathbb{1}\left(\theta_j^u \geq \eta(T_j; r, \omega_j)\right) \,.$$

Clearly, $\theta_j \in \{-1, 1\}$ almost surely and $\mathbb{E}\left[\theta_j | T_j, \omega_j\right] = \eta(T_j; r, \omega_j)$. That is, it is an unbiased estimator for $\eta(T_j; r, \omega_j)$ and independent of other $\theta_{j'}$ for $j \neq j'$. Define the estimator

$$\hat{g}_j(x) := \begin{cases} 0 & \text{when } T > 1 + w_0 \\ \beta^S_{g,k} \theta_j \mathsf{SReLU}_k \left(\frac{\langle \omega_j, x \rangle}{r \|\omega_j\|} - T_j \right) & \text{otherwise}\,. \end{cases} \tag{40}$$

Recall the definition of $\mathsf{S}\Delta_k$ in the discussion preceding Lemma 17. We give a similar lemma below. The proof is the same as the proof of Lemma 17.

Lemma 27 *For every $x \in B^2_q(r)$,*

$$g(x) = \beta^S_{g,k} \iint \eta(T; r, \omega) \mathsf{S}\Delta_k \left(\frac{\langle \omega, x \rangle}{r\|\omega\|}, T \right) \mu_l(dT) \nu_0(d\omega)\,.$$

Recall $\gamma \in \mathcal{S}(\mathbb{R})$, $\gamma^{\perp}_{\omega}$, R and $\Gamma_{q,R}$ as used in Section B. We define $g(x; R)$ and $\hat{g}_j(x; R)$ similarly. Draw (T_j, ω_j) i.i.d. from the distribution $\mu_l \times \nu_0$. Let

$$\hat{g}_j(x; R) := \begin{cases} 0 & \text{when } T_j > 1 + w_0 \\ \beta^S_{g,k} \theta_j \mathsf{S}\Delta_k \left(\frac{\langle \omega_j, x \rangle}{r\|\omega_j\|}, T_j \right) \gamma^{\perp}_{\omega_j}(x) & \text{otherwise}\,. \end{cases} \tag{41}$$

Define for $x \in \mathbb{R}^q$

$$g(x; R) = \mathbb{E}\hat{g}_j(x; R)\,.$$

The definition makes sense when $l \geq 2$ in μ_l since $\mathbb{E}|T_j| < \infty$. We note that $g(x; R)$ is implicitly dependent on k, l, α_0, w_0. Let $\hat{G}_j(\xi; R)$ be the Fourier transform of $\hat{g}_j(x; R)$ and let $G(\xi; R)$ be the Fourier transform of $g(x; R)$. In the Lemma below we show that through the truncation modification above, we can construct an unbiased estimator for both g such that the estimator's derivatives are unbiased estimators for the respective derivatives of g. We give a result similar to Lemma 18 below. The discussion diverges from that in Section B henceforth. Let $\mathbf{b} = (b_1, \ldots, b_q)$ such that $b_1, \ldots, b_q \in \mathbb{N} \cup \{0\}$. By ∂^{b} we denote the differential operator where we differentiate partially with respect to i-th co-ordinate b_i times. We define $|\mathbf{b}| = \sum_{i=1}^q b_i$.

Lemma 28 *Let $R \geq r$ and $l \geq 2$ (where l determines the measure μ_l) so that $\mathbb{E}_{T \sim \mu_l} |T|^2 < \infty$. We also assume that $\beta_{g,k} < \infty$.*

1. *For every $x \in B^2_q(r)$,*

$$g(x; R) = g(x)$$
$$\hat{g}_j(x; R) = \hat{g}_j(x)\,.$$

Where $\hat{g}_j(x)$ is as defined in Equation (40).

2. *For every $\xi, \omega_j \in \mathbb{R}^q$ such that $\omega_j \neq 0$ we define $\xi_\omega := \frac{\langle \xi, \omega \rangle}{\|\omega\|} \in \mathbb{R}$ and $\xi^{\perp}_{\omega} := \xi - \frac{\omega \langle \xi, \omega \rangle}{\|\omega\|^2}$. We have, for any fixed value of T_j and ω_j:*

$$\hat{G}_j(\xi; R) = \begin{cases} 0 \text{ if } T_j > 1 + w_0 \\ \beta^S_{g,k} \theta_j \Gamma_{q-1,R}(\|\xi^{\perp}_{\omega_j}\|) \Lambda^{\alpha_0}_{k,w_0}(\xi_{\omega_j}) \left[\frac{4 e^{i(1+w_0) r \xi_{\omega_j}} \sin^2\left(\frac{(1+w_0-T)\xi_{\omega_j} r}{2}\right)}{\xi^2_{\omega_j} r} \right] & o/w \end{cases} \tag{42}$$

When $q = 1$, we let $\Gamma_{q-1,R}(\cdot) = 1$ identically. We recall that $\Lambda^{\alpha_0}_{k,w_0}$ is the Fourier transform of the filter $\lambda^{\alpha_0}_{k,w_0}$.

3. *The functions $\hat{g}_j(x;R) \in C^{2k}(\mathbb{R}^q)$ a.s. and for every $\mathbf{b} \in (\mathbb{N} \cup \{0\})^q$ such that $|\mathbf{b}| \leq 2k$, almost surely the following holds:*

$$\partial^{\mathbf{b}} \hat{g}_j(\,\cdot\,;R) \in L^1(\mathbb{R}^q) \text{ a.s.}$$

For some constant B_k and for every $x \in \mathbb{R}^q$, we have:

$$|\partial^{\mathbf{b}} \hat{g}_j(x;R)| \leq \beta^S_{g,k} B_k (1+|T_j|) \mathbb{1}\left(T_j \leq -\left|\tfrac{\langle \omega_j, x\rangle}{r\|\omega_j\|}\right| + 2 + 3w_0\right) \mathbb{1}(\|x^{\perp}_{\omega_j}\| \leq 2R)$$

Where B_k is a constant which depends on q, r, k and R but not on g, T_j or ω_j.

4. *$g(x;R) \in C^{2k}(\mathbb{R}^q)$. For every $\mathbf{b} \in (\mathbb{N} \cup \{0\})^q$ such that $|\mathbf{b}| \leq 2k$. Then $\partial^{\mathbf{b}} g(\,\cdot\,;R) \in L^1(\mathbb{R}^q)$ and for every $x \in \mathbb{R}^q$,*

$$\partial^{\mathbf{b}} g(x;R) = \mathbb{E}\partial^{\mathbf{b}} \hat{g}_j(x;R)\,. \tag{43}$$

Some parts of the proof are similar to the proof of Lemma 18. Items 3 and 4 use the duality between multiplication by polynomials and differentiation under Fourier transform.

We define the remainder function similarly as in Section B.

$$g^{\mathsf{rem}}(x) := g(x;R) - \frac{1}{N}\sum_{j=1}^{N} \hat{g}_j(x;R)\,. \tag{44}$$

Clearly $g^{\mathsf{rem}}(x) = g(x) - \frac{1}{N}\sum_{j=1}^{N} \hat{g}_j(x)$ whenever $x \in B^2_q(r)$. Let G^{rem} be its Fourier transform. Lemma 28 implies that $g^{\mathsf{rem}}(x)$ is continuous and L^1. Therefore, it is clear that G^{rem} is continuous. The following lemma is the sup type norm variant of Lemma 19.

Lemma 29 *Let $l \geq 2+q$ so that $\mathbb{E}_{T\sim\mu_l} T^2 < \infty$. Assume $\beta^S_{g,k} < \infty$. For $s \in \{0\} \cup \mathbb{N}$, consider*

$$S^{(s)}_{g^{\mathsf{rem}}} := \sup_{\xi \in \mathbb{R}^q} (1 + \|\xi\|^{q+1}) \|\xi\|^s \frac{|G^{\mathsf{rem}}(\xi)|}{(2\pi)^q}\,.$$

Assume $2k \geq q+1$ and $s \leq 2k-q-1$. We have:

1. $$\mathbb{E} S^{(s)}_{g^{\mathsf{rem}}} \leq \frac{C(S^{(0)}_g + S^{(2k+2)}_g)}{\sqrt{N}}$$

   Where C is a constant depending only on l, s, r, q and k.

2. *$S^{(s)}_{g^{\mathsf{rem}}} \leq C(S^{(0)}_g + S^{2k+2}_g)\left(\frac{1}{N}\sum_{j=1}^{N} 1 + |T_j|^2\right)$ almost surely.*

Remark 30 *Instead of the $s \leq 2k-q-1$ above, a more delicate proof would only require $s < 2k - (q+1)/2$. We will prove the weaker version for the sake of clarity.*

Proof We first consider the expectation bound in item 1. We begin by giving a bound on $\mathbb{E}\int_{\mathbb{R}^q}|\partial^{\mathbf{b}} g^{\mathsf{rem}}(x;R)|dx$ when $|b|\le 2k$:

$$\begin{aligned}\mathbb{E}\int_{\mathbb{R}^q}|\partial^{\mathbf{b}} g^{\mathsf{rem}}(x;R)|dx &\le \int_{\mathbb{R}^q}\sqrt{\mathbb{E}|\partial^{\mathbf{b}} g^{\mathsf{rem}}(x;R)|^2}dx \\ &\le \frac{1}{\sqrt{N}}\int_{\mathbb{R}^q}\sqrt{\mathbb{E}|\partial^{\mathbf{b}}\hat{g}_1(x;R)|^2}dx\,. \end{aligned} \tag{45}$$

Here we have used the fact that $\hat{g}_j(x;R)$ are i.i.d. unbiased estimators for $g(x;R)$. Using the bound in item 3 of Lemma 28, we conclude that

$$\mathbb{E}|\partial^{\mathbf{b}}\hat{g}_j(x;R)|^2 \le (\beta^S_{g,k}B_k)^2\mathbb{E}\left[(1+|T_j|^2)\mathbb{1}\left(T_j\le -\left|\tfrac{\langle\omega_j,x\rangle}{r\|\omega_j\|}\right|+2+3w_0\right)\mathbb{1}(\|x^\perp_{\omega_j}\|\le 2R)\right]. \tag{46}$$

It is clear from integrating tails that

$$\mathbb{E}\left[(1+|T_j|^2)\mathbb{1}\left(T_j\le -\left|\tfrac{\langle\omega_j,x\rangle}{r\|\omega_j\|}\right|+2+3w_0\right)\middle|\omega_j\right] \le \frac{C(l)}{1+\left|\frac{\langle\omega_j,x\rangle}{r\|\omega_j\|}\right|^{(2l-3)}}\,.$$

Using this in Equation (46) and absorbing the constant $C(l,w_0)$ into B_k gives

$$\mathbb{E}|\partial^{\mathbf{b}}\hat{g}_j(x;R)|^2 \le (\beta^S_{g,k}B_k)^2\mathbb{E}\left[\frac{\mathbb{1}(\|x^\perp_{\omega_j}\|\le 2R)}{1+\left|\frac{\langle\omega_j,x\rangle}{r\|\omega_j\|}\right|^{(2l-5)}}\right]. \tag{47}$$

Let

$$\tau(\omega_j,x) := \frac{\mathbb{1}(\|x^\perp_{\omega_j}\|\le 2R)}{1+\left|\frac{\langle\omega_j,x\rangle}{r\|\omega_j\|}\right|^{(2l-3)}}\,.$$

Clearly, $|\tau(\omega_j,x)|\le 1$ almost surely for every x and $\tau(\omega_j,x)$ is non-zero only when $\|x^\perp_{\omega_j}\|\le 2R$. Consider the following conditions on x:

1. $\|x\|\ge 3R$.
2. $\|x^\perp_{\omega_j}\|\le 2R$

It is clear that under these conditions, we have the following:

$$\begin{aligned}\frac{5\|x\|^2}{9} &= \|x\|^2(1-4/9)\le\|x\|^2(1-\tfrac{4R^2}{\|x\|^2}) \\ &= \|x\|^2-4R^2\le\left|\tfrac{\langle x,\omega_j\rangle}{\|\omega_j\|}\right|^2.\end{aligned}$$

Therefore, for some universal constant $c>0$,

$$\tau(\omega_j,x)\le \mathbb{1}(\|x\|\le 3R)+\frac{\mathbb{1}(\|x\|>3R)}{1+\left(\frac{c\|x\|}{r}\right)^{2l-3}}\,. \tag{48}$$

Plugging Equation (47) into Equation (45) gives

$$\mathbb{E}\int_{\mathbb{R}^q}|\partial^{\mathbf{b}}g^{\mathsf{rem}}(x;R)|dx \le \frac{\beta_g^S B_k}{\sqrt{N}}\int_{\mathbb{R}^q}\sqrt{\mathbb{E}\tau(\omega_j,x)}dx\,,$$

and now using Equation (48), we obtain

$$\begin{aligned}\mathbb{E}\int_{\mathbb{R}^q}|\partial^{\mathbf{b}}g^{\mathsf{rem}}(x;R)|dx &\le \frac{\beta_g^S B_k}{\sqrt{N}}\int_{\mathbb{R}^q}\sqrt{\mathbb{1}(\|x\|\le 3R)+\frac{\mathbb{1}(\|x\|>3R)}{1+\left(\frac{c\|x\|}{r}\right)^{2l-5}}}dx\\ &= \frac{\beta_g^S B_k}{\sqrt{N}}\int_{\rho=0}^{\infty}C_q\rho^{q-1}\sqrt{\mathbb{1}(\rho\le 3R)+\frac{\mathbb{1}(\rho>3R)}{1+\left(\frac{c\rho}{r}\right)^{2l-3}}}d\rho\,.\end{aligned}$$

The integral on the right is smaller than some constant $C(q,l,R,r)$ if $l\ge q+2$. Absorbing this constant into B_k too we have that

$$\mathbb{E}\int_{\mathbb{R}^q}|\partial^{\mathbf{b}}g^{\mathsf{rem}}(x;R)|dx \le \frac{\beta_g^S B_k}{\sqrt{N}}\,. \tag{49}$$

By item 4 of Lemma 28, $\partial^{\mathbf{b}}g^{\mathsf{rem}}$ is a continuous L^1 function for $|\mathbf{b}|\le 2k$. We conclude by the Fourier duality of multiplication and differentiation that

$$G^{\mathsf{rem}}(\xi)\prod_{j=1}^{q}\xi_j^{b_j} = (i)^{|b|}\int_{\mathbb{R}^q}\partial^{\mathbf{b}}g^{\mathsf{rem}}(x)e^{i\langle\xi,x\rangle}dx\,. \tag{50}$$

Consider any integer $k\ge u\ge 0$. Now, from Equation (50),

$$\|\xi\|^{2u}G^{\mathsf{rem}}(\xi) = \sum_{\mathbf{b}:|\mathbf{b}|\le 2u}C_{\mathbf{b}}i^{|\mathbf{b}|}\int_{\mathbb{R}^q}\partial^{\mathbf{b}}g^{\mathsf{rem}}(x)e^{i\langle\xi,x\rangle}dx$$

for some constants $C_{\mathbf{b}}$ depending only on u and $\mathbf{b}$. Therefore, we have

$$\begin{aligned}\mathbb{E}\sup_{\xi\in\mathbb{R}^q}\|\xi\|^{2u}|G^{\mathsf{rem}}(\xi)| &\le \sum_{\mathbf{b}:|\mathbf{b}|\le 2u}|C_{\mathbf{b}}|\mathbb{E}\int_{\mathbb{R}^q}|\partial^{\mathbf{b}}g^{\mathsf{rem}}(x)|dx\\ &\le \frac{\beta_{g,k}^S B_k}{\sqrt{N}}\,.\end{aligned} \tag{51}$$

In the second step we have used Equation (49). We have absorbed the constants $|C_{\mathbf{b}}|$ into B_k. It is clear that taking B_k large enough, we can make it depend only on k and not on u. Suppose $2k\ge q+1$. We let $0\le s\le 2k-q-1$. For any $t\ge 0$ we have that $t^s(1+t^{q+1})\le 2(1+t^{2k})$. This follows from the fact that if $c_1\ge c_0>0$, we have $t^{c_0}\le t^{c_1}+1$.

$$\begin{aligned}\mathbb{E}S_{g^{\mathsf{rem}}}^{(s)} &= \mathbb{E}\sup_{\xi}\|\xi\|^s(1+\|\xi\|^{q+1})|G^{\mathsf{rem}}(\xi)|\\ &\le \mathbb{E}\sup_{\xi}2(1+\|\xi\|^{2k})|G^{\mathsf{rem}}(\xi)|\\ &\le \frac{\beta_{g,k}^S B_k}{\sqrt{N}}\,.\end{aligned}$$

Here we have absorbed more constants into B_k. From this we conclude the statement of the lemma in item 1. We now consider the almost sure bound in item 2. Clearly,

$$|G^{\mathsf{rem}}(\xi)| \leq |G(\xi;R)| + \frac{1}{N}\sum_{j=1}^{N}|\hat{G}_j(\xi;R)|\,.$$

We will first bound $\sup_{\xi\in\mathbb{R}^q}|\hat{G}_j(\xi;R)|\|\xi\|^s(1+\|\xi\|^{q+1})$. Integrating the bound in item 3 of Lemma 28, we conclude that the following holds almost surely whenever $|\mathbf{b}|\leq 2k$:

$$\int|\partial^{\mathbf{b}}\hat{g}_j(x,R)|dx \leq B_k\beta_{g,k}^S(1+|T_j|^2)\,.$$

Using similar considerations as in Equation (51), we conclude that whenever $0\leq u\leq k$, almost surely:

$$\sup_{\xi\in\mathbb{R}^q}\|\xi\|^{2u}|\hat{G}_j(\xi;R)| \leq B_k\beta_{g,k}^S(1+|T_j|^2)\,.$$

Since $G(\xi;R)=\mathbb{E}\hat{G}_j(\xi;R)$, taking an expectation of the equation above yields that

$$\sup_{\xi\in\mathbb{R}^q}\|\xi\|^{2u}|G(\xi;R)| \leq B_k\beta_{g,k}\,.$$

Combining the results above proves item 2. ■

For $b\in\mathbb{N}\cup\{0\}$, define

$$k_b^S := b\lceil\tfrac{q+3}{2}\rceil\,. \tag{52}$$

Henceforth, we fix $R=r$ for the sake of clarity. We proceed with the corrective mechanism similar to the one in Theorem 8. Suppose for some $a\in\mathbb{N}\cup\{0\}$ we have $S_g^{(0)}+S_g^{(2k_a^S+2)}<\infty$. Suppose $a=0$. Then, it is clear that there exists a ReLU network with 1 non-linear layer and N non-linear units which achieves a squared error of the order $\frac{1}{N}$. Now consider $a\geq 1$. Define $g^{\mathsf{rem},0}$ to be the remainder for g as defined in equation (44) with $k=k_a^S$ and N replaced with $N/(a+1)$. Now, by Lemma 29,

$$\mathbb{E}\left(S^{(0)}_{g^{\mathsf{rem},0}}+S^{(2k_{a-1}^S+2)}_{g^{\mathsf{rem},0}}\right)\leq B_a\frac{S_g^0+S_g^{2k_a^S+2}}{\sqrt{N}}\,.$$

We recursively obtain $g^{\mathsf{rem},j}$ by replacing g in Equation (44) with $g^{\mathsf{rem},j-1}$, the estimators $\hat{g}_j$ by outputs of $\mathsf{SReLU}_{k_{a-j}^S}$ units which estimate $g^{\mathsf{rem},j-1}$ and with N replaced with $N/(a+1)$. Continuing this way, we deduce that

$$\mathbb{E}\left(S^{(0)}_{g^{\mathsf{rem},a-1}}+S^{(2)}_{g^{\mathsf{rem},a-1}}\right)\leq B_a\frac{S_g^0+S_g^{2k_a^S+2}}{N^{a/2}}\,.$$

Now, $g^{\mathsf{rem},a-1}$ can be estimated by a $N/(a+1)$ unit ReLU network with squared error of the order $\frac{1}{N^{a+1}}$. We note that $g^{\mathsf{rem},a-1}(x)$ is equal to $g(x)$ minus the output of smoothed ReLUs. This implies the following theorem.

Theorem 31

There exists a random neural network with one non-linear layer and N non-linear activations of type ReLU *and* $\mathsf{SReLU}_{k_b^S}$ *for $b \leq a$ such that for any probability distribution ζ on $B_q^2(r)$, we have*

$$\mathbb{E}\int (g(x) - \hat{g}(x))^2 \zeta(dx) \leq B_a \frac{(S_g^0 + S_g^{2k_a^S+2})^2}{N^{a+1}}\,.$$

Here the non-linear activation functions SReLU_k *are of the form* $\mathsf{SReLU}_k(\frac{\langle \omega_j, x}{r\|\omega_j\|} - T_j\rangle)$ *for $j \in [N]$ such that $(\frac{\omega_j}{\|\omega_j\|}, T_j)$ are drawn i.i.d. from probability measure* $\mathsf{Unif}(\mathbb{S}^{q-1}) \times \mu_l$ *where μ_l is the probability measure defined in Theorem 11 with $l \geq q+2$.*

Consider functions of the form defined in Equation (1). Let ν_i be the uniform distribution over the sphere embedded in $X_i := \mathrm{span}(B_i)$. Clearly, X_i is isomorphic to $\mathbb{R}^q$. Let $N/(a+1)m$ be an integer. We can find a random neural network, according to Theorem 31 with N/m neurons such that $\mathbb{E}\hat{f}_i(x) = f_i(x)$ and

$$\mathbb{E}\int (f_i(\langle B_i, x\rangle) - \hat{f}_i(x))^2 \zeta(dx) \leq B_a \left(S_{f_i}^0 + S_{f_i}^{2k_a^S+2}\right)^2 \frac{m^{a+1}}{N^{a+1}}\,.$$

To consider functions of the form given by Equation (1) to obtain Theorem 32 we need to modify Theorem 31 a bit since $x \in \mathbb{R}^d$ instead of $\mathbb{R}^q$ in this case. It is clear that this can be mitigated if we choose the weights according ω_j such that $\omega_j \sim \mathsf{Unif}(\mathbb{S}_i)$ where $\mathbb{S}_i$ is the sphere embedded in $\mathrm{span}(B_i)$.

Theorem 32 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of the form given by Equation (1). We assume that $\left(S_{f_i}^0 + S_{f_i}^{2k_a^S+2}\right)^2 =: M_i$ for some $M_i < \infty$ and define $L = \sum_{i=1}^m M_i$. Let the probability measure μ_l over $\mathbb{R}$ be defined by $\mu_l(dt) \propto \frac{dt}{1+t^{2l}}$ for $l \in \mathbb{N}$. Consider the following sampling procedure:*

1. *Partition $[N] \subseteq \mathbb{N}$ into m disjoint sets, each with $N/(m(a+1))$ elements.*
2. *For $i \in [m]$, $b \in \{0, \ldots, a\}$, $j \in [\frac{N}{m(a+1)}]$, we draw $\omega_{i,j,b}^0 \sim \mathsf{Unif}\left(\mathbb{S}^{\mathrm{span}(B_i)}\right)$ and $T_{i,j,b} \sim \mu_l$ independently for some $l \geq \max(q+3, 3a+3)$.*

Let ζ be any probability distribution over $\mathbb{R}^d$ such that $\langle \zeta(dx), B_i\rangle$ is supported over $B_q^2(r)$. There exist random $\kappa_1, \ldots, \kappa_N \in \mathbb{R}$, depending only on $\omega_{i,j,b}, T_{i,j,b}$ such that for

$$\hat{f}(x) = \sum_{i=1}^{m}\sum_{b=0}^{a}\sum_{j=1}^{\frac{N}{m(a+1)}} \kappa_{i,j,b}\mathsf{SReLU}_{k_b^S}\left(\frac{\langle \omega_{i,j,b}^0, x\rangle}{r} - T_{i,j,b}\right)\,, \tag{53}$$

where $\kappa_{i,j,b} = \kappa_{(i-1)\frac{N}{m} + \frac{bN}{m(a+1)} + j}$, we have:

1. $$\mathbb{E}\int (f - \hat{f})^2 \zeta(dx) \leq B(l,q,r,a) L \frac{m^{a-1}}{N^{a+1}}\,,$$

2. *Whenever* $\delta \in (0,1)$ *and* $\epsilon > 0$ *are given, then with probability at least* $1-\delta$,

$$\int (f - \hat{f})^2 \zeta(dx) \leq \epsilon \,,$$

whenever $N = \Omega\left((\frac{L}{\epsilon\delta})^{\frac{1}{a+1}} m^{\frac{a-1}{a+1}} \right)$. *Here* $\Omega(\,\cdot\,)$ *hides factors depending on* l, q, r *and* a

3. *With probability at least* $1-\delta$, *for any* $b \in \mathbb{N}$ *such that* $b \leq a$,

$$\sum_{j=1}^{N} \kappa_j^2 \leq \frac{C(l,q,r,a)\delta^{-1/(b+1)}}{N} \left(\frac{1}{m} \sum_{i=1}^{m} M_i^{(b+1)} \right)^{1/(b+1)} .$$

Proof As in the proof of Theorem 9, we dedicate N/m activation functions to approximate each of the functions f_i with neural network output $\hat{f}_i$ using the procedure in the proof of Theorem 31. We then approximate $f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(\langle B_i, x\rangle)$ by $\frac{1}{m}\sum_{i=1}^{m} \hat{f}_i(x)$. We choose κ_j as described in the discussion preceding the statement of Theorem 31.

1. The proof is similar to the proof of Theorem 9.

2. The proof follows from a direct application of Markov's inequality on item 1.

3. Consider the random variable $K := \sum_{j=1}^{N} \kappa_j^2$. Consider

$$\begin{aligned} \mathbb{E}K^{b+1} &= N^{b+1} \left(\frac{1}{N} \sum_{j=1}^{N} \kappa_j^2 \right)^{b+1} \\ &\leq N^b \mathbb{E} \sum_{j=1}^{N} \kappa_j^{2b+2} \end{aligned} \tag{54}$$

We have applied Jensen's inequality in the second step. We will control $\mathbb{E}\kappa_j^{2(b+1)}$. Let κ_j be the coefficient of the activation function approximating f_i. By the preceding the theorem statement, item 2 in Lemma 29 and the definition of $\hat{g}_j$ given Equation 40, which gives the κ_j corresponding to ω_j, T_j, it is clear that $|\kappa_j| \preceq \frac{B_a m\sqrt{M_i}}{N^2} \sum_{s=1}^{\overline{(a+1)m}} (1 + |T_s'|^2)$ where T_s' are chosen i.i.d. from μ_l and $\preceq$ denotes stochastic domination. Here the extra factor of N/m in the denominator is due to the fact that when we construct the estimator $\hat{g}(x) := \frac{1}{N'}\sum_{j=1}^{N'} \hat{g}_j(x)$ - there is a division by N'. Therefore,

$$\begin{aligned} \mathbb{E}|\kappa_j|^{2(b+1)} &\leq \mathbb{E}\frac{B_a^{2(b+1)} M_i^{(b+1)} m^{2(b+1)}}{N^{4(b+1)}} \left(\sum_{s=1}^{\frac{N}{m(a+1)}} (1 + |T_s'|^2) \right)^{2(b+1)} \\ &= \mathbb{E}\frac{B_a^{2(b+1)} M_i^{(b+1)}}{N^{2(b+1)}(a+1)^{2(b+1)}} \left(\frac{(a+1)m}{N} \sum_{s=1}^{\frac{N}{m(a+1)}} (1 + |T_s'|^2) \right)^{2(b+1)} . \end{aligned}$$

Now by Jensen's inequality and the fact that $\mathbb{E}(1+|T_1'|^2)^{2(b+1)} < \infty$ by our choice $l \geq 3a+3$, the above is

$$\begin{aligned} &\leq \frac{B_a^{2(b+1)} M_i^{(b+1)}}{N^{2(b+1)}} \mathbb{E}(1+|T_1'|^2)^{2(b+1)} \\ &\leq \frac{B_a^{2(b+1)} M_i^{(b+1)}}{N^{2(b+1)}} C(l,a) \\ &= \frac{B_a^{2(b+1)} M_i^{(b+1)}}{N^{2(b+1)}} \,. \end{aligned} \tag{55}$$

In the last step we absorbed factors not depending on m or N into B_a. Using Equation (55) in Equation (54), we have

$$\mathbb{E}K^{b+1} \leq \sum_{i=1}^{m} \frac{B_a^{2(b+1)} M_i^{(b+1)}}{N^{b+1} m} \,,$$

where we have used that fact that there are exactly N/m coefficients κ_j which corresponding to the activation functions which approximate f_i for any $i \in [m]$. By an application of Markov's inequality, for any $t \geq 0$,

$$\mathbb{P}(K \geq t) \leq \frac{\mathbb{E}K^{b+1}}{t^{b+1}} \,.$$

Setting the RHS above to δ completes the proof. ■

Appendix E. Proofs of Lemmas

E.1. Proof of Lemma 3

The first item follows from the definition of F and the triangle inequality. For the second item, observe that $|F(\xi)|^2 = \sum_{j=1}^{n} f(x_j)^2 + \sum_{j \neq k} f(x_j) f(x_k) e^{i\langle \xi, x_j - x_k \rangle}$. Taking expectation on both sides, we obtain

$$\mathbb{E}|F(\xi)|^2 \leq \|f\|_2^2 + \|f\|_1^2 \exp\left(-\tfrac{\sigma^2\theta^2}{2}\right) \leq \|f\|_2^2 + \frac{\|f\|_1^2}{n^s} \,.$$

The third item follows directly from the definition of $\tilde{f}$ and the choice of σ. ■

E.2. Proof of Lemma 5

$$
\begin{aligned}
\tilde{f}(x_k) &= \mathbb{E}F(\xi)e^{-i\langle \xi, x_k\rangle} = \mathbb{E}|F(\xi)|e^{-i\phi(\xi)-i\langle \xi, x_k\rangle} = \mathbb{E}|F(\xi)|\cos\big(\langle \xi, x_k\rangle - \phi(\xi)\big) \\
&= \mathbb{E}|F(\xi)|\cos\big(\langle \xi, x_k\rangle - \phi(\xi)\big)\mathbb{1}(\mathcal{A}) \\
&\quad + C\mathbb{E}|F(\xi)|(1+\tfrac{4s^2\log^2 n}{\theta^2})\eta(T;\alpha,\psi)\mathsf{ReLU}\Big(\theta\tfrac{\langle \xi, x_k\rangle}{2s\log n} - T\Big)\mathbb{1}(\mathcal{A}^c) \\
&= O(\|f\|_1\mathbb{P}(\mathcal{A})) - C\mathbb{E}|F(\xi)|(1+\tfrac{4s^2\log^2 n}{\theta^2})\eta(T;\alpha,\psi)\mathsf{ReLU}\Big(\theta\tfrac{\langle \xi, x_k\rangle}{2s\log n} - T\Big)\mathbb{1}(\mathcal{A}) \\
&\quad + C\mathbb{E}|F(\xi)|(1+\tfrac{4s^2\log^2 n}{\theta^2})\eta(T;\alpha,\psi)\mathsf{ReLU}\Big(\theta\tfrac{\langle \xi, x_k\rangle}{2s\log n} - T\Big) \\
&= O(\|f\|_1\mathbb{P}(\mathcal{A})) + O\Big(\tfrac{s\|f\|_1\log n}{\theta}\mathbb{E}|\langle \xi, x_k\rangle|\mathbb{1}(\mathcal{A})\Big) \\
&\quad + C\mathbb{E}|F(\xi)|(1+\tfrac{4s^2\log^2 n}{\theta^2})\eta(T;\alpha,\psi)\mathsf{ReLU}\Big(\theta\tfrac{\langle \xi, x_k\rangle}{2s\log n} - T\Big) \\
&= O(\|f\|_1\mathbb{P}(\mathcal{A})) + O\Big(\tfrac{s^{3/2}\|f\|_1\log^{3/2} n}{\theta^2}\sqrt{\mathbb{P}(\mathcal{A})}\Big) \\
&\quad + C\mathbb{E}|F(\xi)|(1+\tfrac{4s^2\log^2 n}{\theta^2})\eta(T;\alpha,\psi)\mathsf{ReLU}\Big(\theta\tfrac{\langle \xi, x_k\rangle}{2s\log n} - T\Big) \\
&= O\Big(\tfrac{s^{3/2}\|f\|_1\log^{3/2} n}{\theta^2 n^{s/2}}\Big) + C\mathbb{E}|F(\xi)|\big(1+\tfrac{4s^2\log^2 n}{\theta^2}\big)\eta(T;\alpha,\psi)\mathsf{ReLU}\Big(\theta\tfrac{\langle \xi, x_k\rangle}{2s\log n} - T\Big).
\end{aligned}
$$

Steps three through five are justified by Item 1 of Lemma 3 to bound $|F(\xi)|$, the fact that $\mathsf{ReLU}(x) \le |x|$ and Item 1 of Lemma 3, and an application of the Cauchy-Schwarz inequality to show that $\mathbb{E}|\langle \xi, x_k\rangle|\mathbb{1}(\mathcal{A}) \le \sqrt{\mathbb{P}(\mathcal{A})}\sqrt{\mathbb{E}|\langle \xi, x_k\rangle|^2} \le \sigma\sqrt{\mathbb{P}(\mathcal{A})}$. ■

E.3. Proof of Lemma 6

We begin with a chain of inequalities, justified right afterward:

$$
\begin{aligned}
\mathbb{E}(f(x_k) - \hat{f}(x_k))^2 &= \frac{\mathbb{E}(\hat{f}_1(x_k))^2 - \Big(\mathbb{E}\hat{f}_1(x_k)\Big)^2}{N_0} + (\tilde{f}(x_k) - \mathbb{E}\hat{f}_1(x_k))^2 + (f(x_k) - \tilde{f}(x_k))^2 \\
&\le \frac{\mathbb{E}(\hat{f}_1(x_k))^2}{N_0} + (f(x_k) - \tilde{f}(x_k))^2 + (\tilde{f}(x_k) - \mathbb{E}\hat{f}_1(x_k))^2 \\
&\le \frac{\mathbb{E}(\hat{f}_1(x_k))^2}{N_0} + \frac{\|f\|_1^2}{n^{2s}} + (\tilde{f}(x_k) - \mathbb{E}\hat{f}_1(x_k))^2 \\
&\le \frac{C\mathbb{E}s^4\log^4 n|F(\xi_1)|^2}{N_0\theta^4}\Big(1+\theta^2\tfrac{|\langle \xi_1, x_k\rangle|^2}{s^2\log^2 n}\Big) + \frac{\|f\|_1^2}{n^{2s}} + (\tilde{f}(x_k) - \mathbb{E}\hat{f}_1(x_k))^2 \\
&\le \frac{C\mathbb{E}s^4\log^4 n|F(\xi_1)|^2}{N_0\theta^4}\Big(1+\theta^2\tfrac{|\langle \xi_1, x_k\rangle|^2}{s^2\log^2 n}\Big) + \frac{\|f\|_1^2}{n^{2s}} + C\frac{s^3\|f\|_1^2\log^3 n}{\theta^4 n^s} \\
&= \frac{Cs^4\log^4 n}{N_0\theta^4}\Big(\mathbb{E}|F(\xi_1)|^2 + \theta^2\mathbb{E}|F(\xi_1)|^2\tfrac{|\langle \xi_1, x_k\rangle|^2}{s^2\log^2 n}\Big) + \frac{\|f\|_1^2}{n^{2s}} + C\frac{s^3\|f\|_1^2\log^3 n}{\theta^4 n^s} \\
&\le \frac{Cs^4\log^4 n}{N_0\theta^4}\Big(\|f\|_2^2 + \frac{\|f\|_1^2}{n^s} + \theta^2\mathbb{E}|F(\xi_1)|^2\tfrac{|\langle \xi_1, x_k\rangle|^2}{s^2\log^2 n}\Big) + \frac{\|f\|_1^2}{n^{2s}} + C\frac{s^3\|f\|_1^2\log^3 n}{\theta^4 n^s}. \qquad (56)
\end{aligned}
$$

The first step is the bias-variance decomposition of the squared error. In the third step we have used item 3 of Lemma 3. In the fourth step we have used the fact that

$\mathsf{ReLU}\left(\theta\frac{\langle\xi,x_k\rangle}{2s\log n}-T\right)\leq 1+\theta\frac{|\langle\xi,x_k\rangle|}{2s\log n}$ almost surely and have absorbed this into the constant C. In the fifth step we have used Lemma 5.

We will now bound $\mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2$ to obtain the stated result. By Gaussian concentration, we have for some universal constant $c>0$ and every $t\geq 0$ that

$$\mathbb{P}\left(|\langle\xi_1,x_k\rangle|\geq\sigma t\right)\leq 2e^{-ct^2}\,.$$

Consider the event $A_t=\{|\langle\xi_1,x_k\rangle|\leq\sigma t\}$ for some $t>0$. Decomposing based on A_t gives

$$\begin{aligned}
\mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2 &= \mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2\mathbb{1}(A_t)+\mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2\mathbb{1}(A_t^c)\\
&\leq \mathbb{E}\sigma^2t^2|F(\xi_1)|^2\mathbb{1}(A_t)+\mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2\mathbb{1}(A_t^c)\\
&\leq \mathbb{E}\sigma^2t^2|F(\xi_1)|^2+\mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2\mathbb{1}(A_t^c)\\
&\leq \sigma^2t^2\mathbb{E}|F(\xi_1)|^2+\|f\|_1^2\mathbb{E}|\langle\xi_1,x_k\rangle|^2\mathbb{1}(A_t^c)\\
&\leq \sigma^2t^2\mathbb{E}|F(\xi_1)|^2+\|f\|_1^2\sqrt{\mathbb{E}|\langle\xi_1,x_k\rangle|^4}\sqrt{\mathbb{P}(A_t^c)}\\
&\leq \sigma^2t^2\left(\|f\|_2^2+\frac{\|f\|_1^2}{n^s}\right)+C\|f\|_1^2\sigma^2e^{-ct^2} \qquad (57)
\end{aligned}$$

In the second step we have used the fact that $|\langle\xi_1,x_k\rangle|\leq\sigma t$ whenever $\mathbb{1}(A_t)=1$. In the third step we have used the fact that $|\mathbb{1}(A_t)|\leq 1$. In the fourth step we have used item 1 of Lemma 3. In the fifth step we have used the Cauchy-Schwarz inequality. In the sixth step we have used item 2 of Lemma 3 to bound $\mathbb{E}|F(\xi_1)|^2$ and the fact that for Gaussian random variables $\mathbb{E}|\langle\xi_1,x_k\rangle|^4\leq C\sigma^4$ for some universal constant C. We have also used the Gaussian concentration inequality to conclude that $\mathbb{P}(A_t^c)\leq 2e^{-ct^2}$ for some universal constant c and redefined and absorbed universal constants where necessary. We take $t=\sqrt{\frac{2s\log n}{c}}$ where c is the constant in the exponent of Equation (57) and $\sigma=\frac{\sqrt{2s\log n}}{\theta}$ to get

$$\mathbb{E}|\langle\xi_1,x_k\rangle|^2|F(\xi_1)|^2\leq\frac{Cs^2\log^2 n}{\theta^2}\left(\|f\|_2^2+\frac{\|f\|_1^2}{n^s}\right)\,. \qquad (58)$$

Using Equation (56) along with Equation (58) gives

$$\mathbb{E}(f(x_j)-\hat{f}(x_j))^2\leq\frac{Cs^4\log^4 n}{\theta^4N_0}\left(\|f\|_2^2+\frac{\|f\|_1^2}{n^s}\right)+\frac{\|f\|_1^2}{n^{2s}}+C\frac{s^3\|f\|_1^2\log^3 n}{\theta^4n^s}\,.$$

Clearly, $\|f\|_1^2\leq n\|f\|_2^2$. Plugging this into the equation above completes the proof. ■

E.4. Proof of Lemma 18

We first prove the following estimates before delving into the proof of Lemma 18.

Lemma 33 *The following holds almost surely:*

$$\int dx|\hat{g}_j(x;R)|\leq\begin{cases}\beta_{g,k}r(1+w_0-T)^2\mathrm{vol}(B_{q-1}^2(2R)) & \text{when } \omega_j\neq 0\\ \beta_{g,k}|1+w_0-T|\mathrm{vol}(B_q^2(2R)) & \text{otherwise}\,,\end{cases} \qquad (59)$$

where $B^2_{q-1}(2R)$ *is seen as a subset of* $\mathbb{R}^{q-1}$ *and* vol *denotes the Lebesgue measure of the set. Whenever* $T_j \leq 1 + w_0$,

$$\sup_{t\in\mathbb{R}} |\mathsf{S}\Delta_k(t;T_j)| \leq 1 + w_0 - T_j \,.$$

Proof When $T_j > 1 + w_0$, the bound above holds trivially since $\hat{g}_j = 0$. Now assume $T_j \leq 1+w_0$. We first note that $\int_{-\infty}^{\infty} |\Delta(t/r;T_j)|dt = r(1+w_0-T_j)^2$ and $\sup_{t\in\mathbb{R}} |\Delta(t/r;T_j)| = 1 + w_0 - T_j$. Since $\mathsf{S}\Delta = \lambda^{\alpha_0}_{k,w_0} * \Delta$ and $\lambda^{\alpha_0}_{k,w_0}$ is a probability density function, we apply Jensen's inequality to conclude the following inequalities:

1. $$\int_{-\infty}^{\infty} |\mathsf{S}\Delta_k(t/r;T_j)|dt \leq r(1 + w_0 - T_j)^2 \,,$$
2. $$\sup_{t\in\mathbb{R}} |\mathsf{S}\Delta_k(t;T_j)| \leq 1 + w_0 - T_j \,.$$

To prove the inequality on the L^1 norm of $\hat{g}_j$, we first consider the case $\omega_j = 0$. We conclude the corresponding bound by noting that $\theta_j \in \{-1, 1\}$ (recall θ_j from the definition of $\hat{g}_j$), $0 \leq \gamma(\frac{\|x\|^2}{R^2}) \leq 1$ and $\gamma(\frac{\|x\|^2}{R^2}) = 0$ when $x \notin B^2_q(2R)$ and $\sup_{t\in\mathbb{R}} |\mathsf{S}\Delta_k(t;T_j)| \leq 1 + w_0 - T_j$.

Now consider the case $\omega_j \neq 0$ and $T_j \leq 1 + w_0$. Clearly, $\gamma^{\perp}_{\omega_j}$ is a function of only the component of x perpendicular to ω_j. Therefore, we decompose the Lebesgue measure dx over $\mathbb{R}^q$ into the product measure $dx_{\omega_j} \times dx^{\perp}_{\omega_j}$ where dx_{ω_j} is the lebesgue measure over $\text{span}(\omega_j)$ and $dx^{\perp}_{\omega_j}$ is the Lebesgue measure over the space perpendicular to ω_j, which is isomorphic to $\mathbb{R}^{q-1}$. The following bound holds:

$$\|\hat{g}_j\|_1 \leq \beta_{g,k} \int |\mathsf{S}\Delta_k(x_{\omega_j}/r;T_j)|dx_{\omega_j} \int \gamma^{\perp}_{\omega_j}(x)dx^{\perp}_{\omega_j} \,.$$

We conclude the result using the fact that $0 \leq \gamma^{\perp}_{\omega_j}(x) \leq 1$, and it vanishes outside $B^2_{q-1}(2R)$ and the fact that $\int_{-\infty}^{\infty} |\mathsf{S}\Delta_k(t/r;T_j)|dt \leq r(1 + w_0 - T_j)^2$ as shown above. ■

Proof of Lemma 18

1. Follows from Lemma 17 and the preceding discussion.

2. From definition, it is clear that $\hat{g}_j(\,\cdot\,;R)$ has compact support almost surely. Therefore $\hat{g}_j(\,\cdot\,;R) \in L^1(\mathbb{R}^q)$ almost surely. To show that $g(\,\cdot\,;R) \in L^1(\mathbb{R}^q)$, it is sufficient to show that $\hat{g}_j(x;R)$ is integrable with respect to the measure $\mu_l \times \nu_{g,k} \times dx$ where dx denotes the Lebesgue measure over $\mathbb{R}^q$. First consider the case $\omega_j \neq 0$:

$$\begin{aligned}
&\int |\hat{g}_j(x;R)|\mu_l(dT_j) \times \nu_{g,k}(d\omega_j) \times dx = \int \left(\int |\hat{g}_j(x;R)|dx \right) \mu_l(dT_j) \times \nu_{g,k}(d\omega_j) \\
&\leq \int \beta_{g,k} r(1 + w_0 - T_j)^2 \text{vol}(B^2_{q-1}(2R))\mu_l(dT_j) \times \nu_{g,k}(d\omega_j) \\
&< \infty \,.
\end{aligned}$$

We have used Fubini's theorem for positive functions in the first step, Lemma 33 in the second step and we have used the fact that $\mathbb{E}|T_j|^2 < \infty$ in the third step. This shows that $g(\,\cdot\,;R) \in L^1(\mathbb{R}^q)$. The case $\omega_j = 0$ follows similarly.

3. The case $T_j > 1 + w_0$ is trivial. The case $\omega_j = 0$ and $T_j \leq 1 + w_0$ is simple to prove from the definitions. We now consider the case $\omega_j \neq 0$, $T_j \leq 1 + w_0$ and $q > 1$. The $q = 1$ case is similar to the one below, but we set $\gamma^{\perp}_{\omega_j}(x) = 1$ all along. We first note that $\omega_j \perp x^{\perp}_{\omega_j}$ and that $\gamma^{\perp}_{\omega_j}(x)$ is a function of $x^{\perp}_{\omega_j}$ only. Therefore, we decompose the Lebesgue measure dx over $\mathbb{R}^d$ into the product measure $dx_{\omega_j} \times dx^{\perp}_{\omega_j}$, where dx_{ω_j} is the Lebesgue measure over $\text{span}(\omega_j)$ and $dx^{\perp}_{\omega_j}$ is the Lebesgue measure over the space perpendicular to ω_j, which is isomorphic to $\mathbb{R}^{q-1}$:

$$\begin{aligned}
\hat{G}_j(\xi; R) &= \beta_{g,k}\theta_j \int \gamma^{\perp}_{\omega_j}(x) \mathsf{S}\Delta_k\left(\tfrac{x_{\omega_j}}{r}, T_j\right) e^{i\langle x,\xi\rangle} dx_{\omega_j} \times dx^{\perp}_{\omega_j} \\
&= \beta_{g,k}\theta_j \int \gamma^{\perp}_{\omega_j}(x) e^{i\langle x^{\perp}_{\omega_j}, \xi^{\perp}_{\omega_j}\rangle} \mathsf{S}\Delta_k\left(\tfrac{x_{\omega_j}}{r}, T_j\right) e^{i x_{\omega_j}\xi_{\omega_j}} dx_{\omega_j} \times dx^{\perp}_{\omega_j} \\
&= \beta_{g,k}\theta_j \int \gamma^{\perp}_{\omega_j}(x) e^{i\langle x^{\perp}_{\omega_j}, \xi^{\perp}_{\omega_j}\rangle} dx^{\perp}_{\omega_j} \int \mathsf{S}\Delta_k\left(\tfrac{x_{\omega_j}}{r}, T_j\right) e^{i x_{\omega_j}\xi_{\omega_j}} dx_{\omega_j} \,. \qquad (60)
\end{aligned}$$

In the third step, we have used the fact that $\gamma^{\perp}_{\omega_j}$ depends only on $x^{\perp}_{\omega_j}$ and that $\mathsf{S}\Delta\left(\frac{x_{\omega_j}}{r}, T_j\right)$ depends only on x_{ω_j}. Now, we consider $\mathsf{S}\Delta_k$ and its Fourier transform. For ease of notation, we replace x_{ω_j} by just $t \in \mathbb{R}$. Let $1 + w_0 \geq T \in \mathbb{R}$. Consider the function

$$\Delta(t; T) := \mathsf{ReLU}(t - T) - 2\mathsf{ReLU}(t - 1 - w_0) + \mathsf{ReLU}(t - 2 - 2w_0 + T) \,.$$

It is simple to check that the Fourier transform of $\Delta(t/r; T)$ is $\frac{4e^{i(1+w_0)\xi r}}{\xi^2 r}\sin^2((1 + w_0 - T)\xi r/2)$. $\mathsf{S}\Delta(x/r, T)$ is obtained from $\Delta(x/r; T)$ by convolving it with $\lambda^{\alpha_0}_{k,w_0}$. Therefore, from the convolution theorem we conclude that the Fourier transform of $\mathsf{S}\Delta(x/r; T)$ is $\frac{4e^{i(1+w_0)\xi r}}{\xi^2 r}\sin^2((1 + w_0 - T)\xi r/2)\Lambda^{\alpha_0}_{k,w_0}(\xi)$.

Now, $\gamma^{\perp}_{\omega_j}(x)$ is a function of $x^{\perp}_{\omega_j}$ only. Therefore, we can see this as a function with domain $\mathbb{R}^{q-1}$. In Equation (60), we conclude that the first integral, involving $\gamma^{\perp}_{\omega_j}$ infact gives its Fourier transform over $\Gamma_{q-1,R}$. Using these results in Equation (60), we obtain

$$\hat{G}_j(\xi; R) = \beta_{g,k}\theta_j \Gamma_{q-1,R}(\|\xi^{\perp}_{\omega_j}\|)\Lambda^{\alpha_0}_{k,w_0}(\xi_{\omega_j})\left[\frac{4e^{i(1+w_0)r\xi_{\omega_j}}}{\xi^2_{\omega_j} r}\sin^2((1 + w_0 - T)\xi_{\omega_j} r/2)\right] .$$

4. The fact that $\hat{G}_j \in L^1(\mathbb{R}^q)$ follows from item 3. The fact that $G(\xi; R) = \mathbb{E}\hat{G}_j(\xi, R)$ follows from Fubini's theorem after checking that $|\hat{g}_j|$ is integrable with respect to the product measure $\mu_l \times \nu_{g,k} \times dx$ (where dx denotes the Lebesgue measure over $\mathbb{R}^q$) as shown in the proof of item 2. Similar to the proof of item 2, we will conclude that $G(\xi; R) \in L^1(\mathbb{R}^q)$ by showing that $|\hat{G}_j(\xi; R)|$ is integrable with respect to the measure $\mu_l \times \nu_{g,k} \times d\xi$. In the cases $T_j > 1 + w_0$, $|\hat{G}_j(\,\cdot\,; R)| = 0$. When $T_j \leq 1 + w_0$ and $\omega_j = 0$, we know that $\Gamma_{q,R}(\|\xi\|) \in \mathcal{S}(\mathbb{R}^q)$ and therefore an L^1 function. Using the fact that $|\mathsf{S}\Delta(0; T_j)| \leq 1 + w_0 - T_j$, we conclude that in this case: $\int_{\mathbb{R}^q} |\hat{G}_j(\xi; R)| d\xi \leq \beta_{g,k}\|\Gamma_{q,R}\|_1(1 + w_0 - T_j)$. Now consider the case $T_j \leq 1 + w_0$ and $\omega_j \neq 0$. We first note an inequality which follows from elementary considerations for every $a > 0$ and $\xi \in \mathbb{R}$:

$$\frac{\sin^2(a\xi)}{\xi^2} \leq \min\left(a^2, \frac{1}{\xi^2}\right) . \qquad (61)$$

By Lemma 21,

$$|\Lambda^{\alpha_0}_{k,w_0}(\xi_{\omega_j})| \leq \frac{C(k,\omega_0,\alpha_0)}{1+\xi^{2k}_{\omega_j}}\,. \tag{62}$$

Using Equations (61) and (62), along with the expression for $\hat{G}_j(\,\cdot\,;R)$ in item 3, we have

$$|\hat{G}_j(\xi;R)| \leq \beta_{g,k}\Gamma_{q-1,R}(\|\xi^{\perp}_{\omega_j}\|)\frac{C(k,\omega_0,\alpha_0)}{1+\xi^{2k}_{\omega_j}}\min\left(r(1+w_0-T_j)^2, \frac{1}{r\xi^2_{\omega_j}}\right)\,.$$

Integrating this over $\mathbb{R}^q$, we get

$$\|\hat{G}_j(\,\cdot\,;R)\|_1 \leq C(k,\omega_0,\alpha_0,r)\beta_{g,k}\|\Gamma_{q-1,R}\|_1|1+w_0-T_j|\,.$$

Here we have abused notation to denote by $\|\Gamma_{q-1,R}\|_1$ the L^1 norm of $\Gamma_{q-1,R}$ when seen as a function over $\mathbb{R}^{q-1}$.

Combining the various cases, we conclude that $\hat{G}_j(\,\cdot\,;R)$ is integrable with respect to the measure $\mu_l \times \nu_{g,k} \times dx$ if $\mathbb{E}|1+w_0-T_j| < \infty$. This is true since we have chosen $l \geq 2$ in the statement of the lemma. ■

E.5. Proof of Lemma 19

We first state the following useful lemma before delving into the proof of Lemma 19.

Lemma 34 *Let Z be uniformly distributed on the sphere $\mathbb{S}^{q-1}$ for $q \geq 2$ and let $\rho > 0$ and $a, b \in \mathbb{R}^+$ be such that $b > \frac{q-1}{2}$. Let Z_1 denote the component of Z along the direction of the standard basis vector e_1. Then*

$$\int_{\mathbb{S}^{q-1}} \frac{1}{1+\rho^{2a}Z_1^{2a}}\frac{1}{1+(1-Z_1^2)^b\rho^{2b}}p_\theta(dZ) \leq C(q,a,b)\left[\frac{1}{1+\rho^{2b}} + \frac{\rho^{-q+1}}{1+\rho^{2a}}\right]\,.$$

Proof From standard results, it is clear that Z_1 is distributed over $[-1,1]$ with the density function $\psi_q(x) := C_q(1-x^2)^{\frac{q-3}{2}}$. Here C_q is the normalizing constant. Therefore, the integral in the statement of the lemma becomes

$$\begin{aligned}
&\int_{-1}^{1} \frac{\psi_q(x)}{1+\rho^{2a}x^{2a}}\frac{dx}{1+(1-x^2)^b\rho^{2b}} = 2\int_0^1 \frac{\psi_q(x)}{1+\rho^{2a}x^{2a}}\frac{dx}{1+(1-x^2)^b\rho^{2b}} \\
&= 2\int_0^{1/2} \frac{\psi_q(x)}{1+\rho^{2a}x^{2a}}\frac{dx}{1+(1-x^2)^b\rho^{2b}} + 2\int_{1/2}^1 \frac{\psi_q(x)}{1+\rho^{2a}x^{2a}}\frac{dx}{1+(1-x^2)^b\rho^{2b}} \\
&\leq \frac{2}{1+2^{-2b}\rho^{2b}}\int_0^{1/2}\psi_q(x)dx + \int_{1/2}^1 \frac{\psi_q(x)}{1+\rho^{2a}2^{-2a}}\frac{dx}{1+(1-x^2)^b\rho^{2b}} \\
&\leq \frac{C(q,b)}{1+\rho^{2b}} + \frac{C(q,a)}{1+\rho^{2a}}\int_{1/2}^1 \frac{(1-x^2)^{\frac{q-3}{2}}}{1+(1-x^2)^b\rho^{2b}}dx\,.
\end{aligned}$$

In the integral from $1/2$ to 1, $2x \geq 1$. Therefore, from the equation above,

$$\int_{-1}^{1} \frac{\psi_q(x)}{1+\rho^{2a}x^{2a}}\frac{dx}{1+(1-x^2)^b\rho^{2b}} \leq \frac{C(q,b)}{1+\rho^{2b}} + \frac{C(q,a)}{1+\rho^{2a}}\int_{1/2}^1 \frac{(1-x^2)^{\frac{q-3}{2}}}{1+(1-x^2)^b\rho^{2b}}2xdx\,.$$

We now make the change of variable $t = (1-x^2)\rho^2$, yielding

$$\int_{-1}^{1} \frac{\psi_q(x)}{1+\rho^{2a}x^{2a}} \frac{dx}{1+(1-x^2)^b\rho^{2b}} \leq \frac{C(q,b)}{1+\rho^{2b}} + \frac{C(q,a)\rho^{-q+1}}{1+\rho^{2a}} \int_0^{\frac{3\rho^2}{4}} \frac{t^{\frac{q-3}{2}}}{1+t^b} dt$$
$$\leq \frac{C(q,b)}{1+\rho^{2b}} + \frac{C(q,a)\rho^{-q+1}}{1+\rho^{2a}} \int_0^{\infty} \frac{t^{\frac{q-3}{2}}}{1+t^b} dt\,.$$

The integral on the RHS is finite if $b > \frac{q-1}{2}$ for every $q \geq 2$. Using this fact in the equation above, we conclude the statement of the lemma. ■

Proof (of Lemma 19) To begin, Fubini's theorem for positive functions and Jensen's inequality imply that

$$\mathbb{E}C^{(s)}_{g^{\mathsf{rem}}} = \mathbb{E}\int_{\mathbb{R}^q} \|\xi\|^s |G^{\mathsf{rem}}(\xi)| d\xi$$
$$= \int_{\mathbb{R}^q} \|\xi\|^s \mathbb{E}|G^{\mathsf{rem}}(\xi)| d\xi$$
$$\leq \int_{\mathbb{R}^q} \|\xi\|^s \sqrt{\mathbb{E}|G^{\mathsf{rem}}(\xi)|^2} d\xi\,. \tag{63}$$

By linearity of Fourier transform, we have $G^{\mathsf{rem}}(\xi) = G(\xi;R) - \frac{1}{N}\sum_{i=1}^N \hat{G}_j(\xi;R)$. By item 4 of Lemma 18, we know that for every $\xi \in \mathbb{R}^q$,

$$\mathbb{E}|G^{\mathsf{rem}}(\xi)|^2 = \frac{1}{N}\left[\mathbb{E}\left|\hat{G}_j(\xi;R)\right|^2 - |G(\xi;R)|^2\right] \leq \frac{1}{N}\left[\mathbb{E}\left|\hat{G}_j(\xi;R)\right|^2\right]\,.$$

Using this in Equation (63), we have that

$$\mathbb{E}C^{(s)}_{g^{\mathsf{rem}}} \leq \frac{1}{\sqrt{N}} \int_{\mathbb{R}^q} \|\xi\|^s \sqrt{\mathbb{E}|\hat{G}_j(\xi;R)|^2} d\xi\,. \tag{64}$$

We use the polar decomposition of $\mathbb{R}^q$. Let p_θ be the uniform probability measure on $\mathbb{S}^{q-1}$, the sphere embedded in $\mathbb{R}^q$. Continuing Equation (64),

$$\mathbb{E}C^{(s)}_{g^{\mathsf{rem}}} \leq \frac{1}{\sqrt{N}} \int_{\mathbb{R}^q} \|\xi\|^s \sqrt{\mathbb{E}|\hat{G}_j(\xi;R)|^2} d\xi$$
$$= \frac{C(q)}{\sqrt{N}} \int_{\rho=0}^{\infty} \int_{\mathbb{S}^{q-1}} \rho^{s+q-1} \sqrt{\mathbb{E}|\hat{G}_j(\rho Z;R)|^2} p_\theta(dZ) d\rho$$
$$\leq \frac{C(q)}{\sqrt{N}} \int_{\rho=0}^{\infty} \rho^{s+q-1} \sqrt{\int_{\mathbb{S}^{q-1}} \mathbb{E}|\hat{G}_j(\rho Z;R)|^2 p_\theta(dZ)} d\rho$$
$$= \frac{C(q)}{\sqrt{N}} \int_{\rho=0}^{\infty} \rho^{s+q-1} \sqrt{\mathbb{E}\int_{\mathbb{S}^{q-1}} |\hat{G}_j(\rho Z;R)|^2 p_\theta(dZ)} d\rho\,. \tag{65}$$

The third step above follows from Jensen's inequality applied to the probability measure p_θ. We first consider the case $q \geq 2$. We will now upper bound $\int_{\mathbb{S}^{q-1}} |\hat{G}_j(\rho Z;R)|^2 p_\theta(dZ)$ as a function of ρ. We note the following inequalities:

1. Now, by definition of the Schwartz space, for every integer n, there exists a constant $C(n,q,R)$ such that for every $\xi \in \mathbb{R}^d$
$$|\Gamma_{q-1,R}(\xi_\omega^\perp)| \leq \frac{C(n,q,R)}{1+\|\xi_\omega^\perp\|^n}\,.$$

2. From Lemma 21, we have
$$|\Lambda_{k,w_0}^{\alpha_0}(\xi_\omega)| \leq \frac{C_2}{1+\xi_\omega^{2k}}\,.$$

3. Similar to item 1, we have for every $\xi \in \mathbb{R}^q$:
$$|\Gamma_{q,R}(\xi)| \leq \frac{C(n,q,R)}{1+\|\xi\|^n}\,.$$

From the proof of item 4 of Lemma 18, we have

$$|\hat{G}_j(\xi)| \leq \begin{cases} \beta_{g,k}|\Gamma_{q,R}(\xi)||1+w_0-T_j| & \text{when } \omega_j = 0 \\ C(k)\beta_{g,k}|\Gamma_{q-1,R}(\xi_\omega^\perp)|\Lambda_{k,w_0}^{\alpha_0}(\xi)|| \min\left(r(1+w_0-T_j)^2, \frac{1}{r\xi_{\omega_j}^2}\right) & \text{when } \omega_j \neq 0\,. \end{cases} \tag{66}$$

We use the inequality $\min(a^2, \frac{1}{x^2}) \leq \frac{1+a^2}{1+x^2}$ along with the inequalities above to show that for every $\xi \in \mathbb{R}^q$

$$|\hat{G}_j(\xi)| \leq \begin{cases} \frac{C\beta_{g,k}}{1+\|\xi\|^n}|1+w_0-T_j| & \text{when } \omega_j = 0 \\ \frac{C\beta_{g,k}}{1+\|\xi_\omega^\perp\|^n}\frac{1+(1+w_0-T)^2}{1+\xi_{\omega_j}^{2k+2}} & \text{when } \omega_j \neq 0\,, \end{cases} \tag{67}$$

where C depends on k,q,n,R and r. Therefore

$$\int_{\mathbb{S}^{q-1}} |\hat{G}_j(\rho Z;R)|^2 p_\theta(dZ) \leq \begin{cases} \frac{C\beta_{g,k}^2(1+w_0-T_j)^2}{1+\rho^{2n}} \text{ when } \omega_j = 0 \\ C\beta_{g,k}^2(1+(1+w_0-T_j)^4)\int_{\mathbb{S}^{q-1}} \frac{p_\theta(dZ)}{1+\rho^{4k+4}Z_{\omega_j}^{4k+4}}\frac{1}{1+(1-Z_{\omega_j}^2)^n\rho^{2n}} \\ \text{when } \omega_j \neq 0\,. \end{cases} \tag{68}$$

Using the rotational invariance of p_θ, we invoke Lemma 34 and conclude that when $n > \frac{q-1}{2}$

$$\int_{\mathbb{S}^{q-1}} |\hat{G}_j(\rho Z;R)|^2 p_\theta(dZ) \leq \begin{cases} \frac{C\beta_{g,k}^2(1+w_0-T_j)^2}{1+\rho^{2n}} \text{ when } \omega_j = 0 \\ C\beta_{g,k}^2\left(1+(1+w_0-T_j)^4\right)\left[\frac{1}{1+\rho^{2n}} + \frac{\rho^{-q+1}}{1+\rho^{4k+4}}\right] \\ \text{when } \omega_j \neq 0\,. \end{cases} \tag{69}$$

Since n can be arbitrarily large (and this only changes the multiplicative constant), we can pick $n = 2k+2+q-1$. Now taking expectation with respect to T and noting that when $l \geq 3$, $\mathbb{E}T^4 < \infty$, we have that

$$\mathbb{E}\int_{\mathbb{S}^{q-1}} |\hat{G}_j(\rho Z;R)|^2 p_\theta(dZ) \leq C\beta_{g,k}^2\left[\frac{\rho^{-q+1}}{1+\rho^{4k+4}}\right]\,. \tag{70}$$

Consider the case $q = 1$: it is easy to show from the techniques above that the same bound as in Equation (70) holds. Plugging this into Equation (65), we conclude that

$$\mathbb{E}C^{(s)}_{g^{\mathsf{rem}}} \leq \frac{C\beta_{g,k}}{\sqrt{N}} \int_0^\infty d\rho \frac{\rho^{s+\frac{q-1}{2}}}{1+\rho^{2k+2}} \,.$$

Now, it is clear that the integral on the RHS is finite when $s < \frac{3-q}{2} + 2k$. Using the definition of $\beta_{g,k}$ we conclude the result. ■

E.6. Proof of Lemma 28

The first 2 items are similar as in the proof of Lemma 18. We will show items 3 and 4 below.

3. In Equation (41), $\gamma^{\perp}_{\omega_j}(x)$ is infinitely differentiable. Therefore, to show that $\hat{g}_j(\,\cdot\,; R) \in C^{2k}(\mathbb{R}^q)$, it is sufficient to show that $\mathsf{S}\Delta_k\left(\frac{\langle \omega_j, x\rangle}{r\|\omega_j\|}, T_j\right)$ is $2k$ times continuously differentiable. This reduces to showing that $t \to \mathsf{S}\Delta_k(t,T) \in C^{2k}(\mathbb{R})$ for $T \leq 1 + w_0$. (We only need to worry about the case $T_j \leq 1 + w_0$ because otherwise $\hat{g}_j(x; R) = 0$ identically). Consider the Fourier transform of $\mathsf{S}\Delta_k(t,T)$:

$$\mathsf{S}\Delta^F_k(v) = \frac{4e^{i(1+w_0)v}}{v^2} \sin^2((1+w_0-T)v/2)\Lambda^{\alpha_0}_{k,w_0}(v) \,.$$

Using the upper bounds on $\Lambda^{\alpha_0}_{k,w_0}(v)$ in Lemma 21, $v^{2k}\mathsf{S}\Delta^F_k(v)$ is a L^1 function with respect to Lebesgue measure. By duality between multiplication by v of the Fourier transform and differentiation of the function, we conclude that $\mathsf{S}\Delta_k(t,T)$ is $2k$ times continuously differentiable and and hence that $\hat{g}_j(x; R) \in C^{2k}(\mathbb{R}^q)$ almost surely. Further, for every $l \leq 2k$, we have

$$D^{(l)}\mathsf{S}\Delta_k(t;T) = \frac{1}{2\pi}\int (-i)^l(v)^l \mathsf{S}\Delta^F_k(v)e^{-ivt}dv \,.$$

Therefore,

$$\begin{aligned}
\sup_{t\in\mathbb{R}} |D^{(l)}\mathsf{S}\Delta_k(t;T)| &\leq \frac{1}{2\pi}\int |v|^l |\mathsf{S}\Delta^F_k(v)|dv \\
&\leq \int_{-\infty}^{\infty} B^0_k \min\left((1+w_0-T)^2, \frac{1}{v^2}\right)\frac{|v|^l}{1+|v|^{2k}}dv \\
&\leq B^0_k|1+w_0-T| \leq B^0_k(1+|T|)\,, \qquad (71)
\end{aligned}$$

where $B^0_k < \infty$ is a constant depending only on α_0, w_0 and k. We have absorbed constants involving α_0, w_0 and k into other constants throughout and used the inequality $\frac{\sin^2(v(1+w_0-T)/2)}{v^2} \leq \min\left((1+w_0-T)^2, \frac{1}{v^2}\right)$ and the upper bound on $\Lambda^{\alpha_0}_{k,w_0}(v)$ in Lemma 21. We can in fact improve this bound further because of the fact that $\mathsf{S}\Delta(t;T)$ is supported between $[T-w_0, 2+3w_0-T]$. Therefore, $|D^{(l)}\mathsf{S}\Delta_k(t;T)|$ is non zero only when $t \in [T-w_0, 2+3w_0-T]$. That is when $T - w_0 \leq t \leq 2+3w_0-T$. These inequalities along with the assumption that $T \leq 1+w_0$ imply that $|D^{(l)}\mathsf{S}\Delta_k(t;T)|$ is non-zero only when $T \leq -|t| + 2 + 3w_0$. Therefore, from Equation (71), we conclude:

$$|D^{(l)}\mathsf{S}\Delta_k(t;T)| \le B_k^0(1+|T|)\mathbb{1}(T \le -|t|+2+3w_0)\,. \tag{72}$$

Consider the element wise partial order $\le$ on $(\mathbb{N}\cup\{0\})^q$ where $\mathbf{a}\le\mathbf{b}$ iff $a_i \le b_i$ for $i\in[q]$. By the chain rule, we conclude that $\partial^{\mathbf{b}}\hat{g}_j(x;R)$ is a finite linear combination of terms of the form

$$\beta_{g,k}^S\theta_j\partial^{\mathbf{a}}\left(\mathsf{S}\Delta_k\left(\tfrac{\langle\omega_j,x\rangle}{r\|\omega_j\|},T_j\right)\right)\partial^{\mathbf{b}-\mathbf{a}}\gamma_{\omega_j}^{\perp}(x)\,, \tag{73}$$

for every $\mathbf{a}\le\mathbf{b}$ such that the coefficients depend only on $\mathbf{a}$ and $\mathbf{b}$. Now,

$$\beta_{g,k}^S\theta_j\partial^{\mathbf{a}}\left(\mathsf{S}\Delta_k\left(\tfrac{\langle\omega_j,x\rangle}{r\|\omega_j\|},T_j\right)\right) = \beta_{g,k}^S\theta_j\frac{\prod_{s=1}^q\langle\omega_j,e_s\rangle^{a_s}}{r^{|\mathbf{a}|}\|\omega_j\|^{|\mathbf{a}|}}D^{|\mathbf{a}|}\mathsf{S}\Delta_k\left(\tfrac{\langle\omega_j,x\rangle}{r\|\omega_j\|},T_j\right)\,. \tag{74}$$

From Equation (72), the quantity above is nonzero only when $|x_{\omega_j}| \le 2r+3rw_0-rT_j$. $\gamma_{\omega^\perp}(x)$ is a C^∞ function which vanishes when $\|x_{\omega^\perp}\|\ge 2R$, we conclude that $\partial^{\mathbf{b}-\mathbf{a}}\gamma_{\omega_j}^{\perp}(x)$ also vanishes when $\|x_{\omega^\perp}\|\ge 2R$. Therefore, we conclude that $\partial^{\mathbf{b}}\hat{g}_j(x)$ is continuous and compactly supported almost surely and hence in $L^1(\mathbb{R}^q)$.

Now for the bound on $\partial^{\mathbf{b}}\hat{g}_j(x)$, we proceed as above by noting that this is a linear combination of the terms of the form given in Equation (73) for $\mathbf{a}\le\mathbf{b}$. Now, $\partial^{\mathbf{b}-\mathbf{a}}\gamma_{\omega_j}^{\perp}(x)$ is bounded uniformly by a constant H_k for every x and a where H_k doesn't depend on ω_j. The function $\partial^{\mathbf{b}-\mathbf{a}}\gamma_{\omega_j}^{\perp}(x)$ vanishes when $\|x_{\omega_j}^{\perp}\|\ge 2R$. From Equations (72) and (74) we get that

$$\left|\beta_{g,k}^S\theta_j\partial^{\mathbf{a}}\left(\mathsf{S}\Delta_k\left(\tfrac{\langle\omega_j,x\rangle}{r\|\omega_j\|},T_j\right)\right)\right| \le \beta_{g,k}^S B_k(1+|T_j|)\mathbb{1}(rT_j\le -|x_{\omega_j}|+2r+3w_0r)\,.$$

Here B_k depends on α_0,q,r,k,R and w_0 but not on g, T_j or ω_j. Therefore, we obtain the desired bound (where we have absorbed all the constants into B_k, redefining as necessary):

$$|\partial^{\mathbf{b}}\hat{g}_j(x;R)| \le \beta_{g,k}^S B_k(1+|T_j|)\mathbb{1}(rT_j\le -|x_{\omega_j}|+2r+3w_0r)\mathbb{1}(\|x_{\omega_j}^{\perp}\|\le 2R)\,. \tag{75}$$

4. The proof follows through an induction over $|\mathbf{b}|$ and use of item 3. We will show this for one differentiation here but the argument can be extended to $2k$ times differentiation. By standard results in probability theory, $\frac{\partial g(x;R)}{\partial x_1}$ exists and equal to $\mathbb{E}\frac{\partial\hat{g}_j(x;R)}{\partial x_1}$ if $\frac{\partial\hat{g}_j(x;R)}{\partial x_1}$ exists and for every x, $|\frac{\partial\hat{g}_j(x;R)}{\partial x_1}|\le Z$ for some integrable random variable Z. From item 3, we conclude that $\frac{\partial\hat{g}_j(x;R)}{\partial x_1}$ exists and take $Z=\beta_{g,k}^S B_k(1+|T_j|)$ where $\beta_{g,k}^S B_k$ are constants as used in the statement of item 3. This shows that $\frac{\partial g(x;R)}{\partial x_1}=\mathbb{E}\frac{\partial\hat{g}_j(x;R)}{\partial x_1}$. We show that it is continuous by using dominated convergence theorem after noting the fact that $\frac{\partial\hat{g}_j(x;R)}{\partial x_1}$ is continuous and dominated by $Z=\beta_{g,k}^S B_k(1+|T_j|)$, which is integrable.

To show that $\partial^{\mathbf{b}}g(x;R)\in L^1(\mathbb{R}^q)$, it is sufficient to show that $\partial^{\mathbf{b}}\hat{g}_j(x;R)$ is integrable with respect to the measure $\mu_l\times\nu_0\times dx$ where dx denotes the Lebesgue measure over $\mathbb{R}^q$. From Fubini's theorem for positive functions, we conclude that

$$\int|\partial^{\mathbf{b}}\hat{g}_j(x;R)|\mu_l(dT_j)\times\nu_0(d\omega_j)\times dx = \int\mu_l(dT_j)\times\nu_0(d\omega_j)\int|\partial^{\mathbf{b}}\hat{g}_j(x;R)|dx\,.$$

Integrating Equation (75) over $\mathbb{R}^q$, we conclude that $\int|\partial^{\mathbf{b}}\hat{g}_j(x;R)|dx\le C(1+|T_j|^2)$ for some non-random constant C. Since $\mathbb{E}|T_j|^2<\infty$ by assumption in the statement of the lemma, we conclude that $\partial^{\mathbf{b}}\hat{g}_j(\,\cdot\,;R)$ is integrable with respect to $\mu_l\times\nu_0\times dx$ which implies the desired result. ■

Appendix F. Proof of Main Theorems

F.1. Proof of Theorem 8

We now prove Theorem 8. For the case $a = 0$, we can obtain this error using a ReLU network as shown in Theorem 16. By Equation (17), $|\kappa_j| \leq \beta_{g,0} \leq \frac{1}{N} C_1 \big(C_g^0 + C_g^{(2)}\big)$ almost surely and the bound on $\sum_{j=1}^N |\kappa_j|$ follows. Now we let $a \geq 1$. For the sake of clarity, we will assume that $\frac{N}{a+1}$ is an integer.

Item 2 of Theorem 7 implies that there exists a two-layer SReLU_{k_a} network with $N/(a+1)$ activation functions with output $\hat{g}^0(x)$ and there exists a remainder function $g^{\mathsf{rem},0} : \mathbb{R}^q \to \mathbb{R}$ such that for every $x \in B_q^2(r)$, we have $g^{\mathsf{rem},0}(x) = g(x) - \hat{g}^{(0)}(x)$ and

$$C_{g^{\mathsf{rem},0}}^{(0)} + C_{g^{\mathsf{rem},0}}^{(2k_{a-1}+2)} \leq C \frac{\big(C_g^{(0)} + C_g^{(2k_a+2)}\big)}{\sqrt{N}} .$$

Supposing that $\hat{g}^{(0)}(x) = \sum_{j=1}^{N/(a+1)} \kappa_j^a \mathsf{SReLU}_{k_a}(\langle \omega_j^a, x\rangle - T_j^a)$, by similar considerations as the $a = 0$ case we conclude that $\sum_{j=1}^{N/a} |\kappa_j^a| \leq C_1 \big(C_g^{(0)} + C^{(2k_a+2)}\big)$ almost surely. The fact that $\|\omega_j^a\| \leq 1/r$ follows from Equation (17), which is used to construct the estimators in Theorem 7.

Invoking Theorem 7 again, we conclude that we can approximate $g^{\mathsf{rem},0}$ by $\hat{g}^{(1)}$, which is the output two-layer $\mathsf{SReLU}_{k_{a-1}}$ network with $\frac{N}{a+1}$ non-linear activation functions and there exists $g^{\mathsf{rem},1} : \mathbb{R}^q \to \mathbb{R}$ such that $g^{\mathsf{rem},1}(x) = g^{\mathsf{rem},0}(x) - \hat{g}^{(1)}(x)$ and

$$C_{g^{\mathsf{rem},1}}^{(2k_{a-2}+2)} + C_{g^{\mathsf{rem},1}}^{(0)} \leq C \frac{C_{g^{\mathsf{rem},0}}^{(0)} + C_{g^{\mathsf{rem},0}}^{(2k_{a-1}+2)}}{\sqrt{N}} \leq C \frac{\big(C_g^{(0)} + C_g^{(2k_a+2)}\big)}{N} .$$

Continuing similarly, for $1 \leq b \leq a-1$ we obtain $\hat{g}^{(b)}$ which is the output of some $\mathsf{SReLU}_{k_{a-b}}$ units with $\frac{N}{a+1}$ neurons and remainders $g^{\mathsf{rem},b} : \mathbb{R}^q \to \mathbb{R}$ such that for every $x \in B_q^2(r)$, we have $g^{\mathsf{rem},b}(x) = g^{\mathsf{rem},b-1}(x) - \hat{g}^{(b)}(x)$ and

$$C_{g^{\mathsf{rem},b}}^{(2k_{a-b-1}+2)} + C_{g^{\mathsf{rem},b}}^{(0)} \leq C \frac{\big(C_g^{(0)} + C_g^{(2k_a+2)}\big)}{N^{\frac{b+1}{2}}} .$$

Now, writing $\hat{g}^{(b)}(x) = \sum_{j=1}^{N/(a+1)} \kappa_j^{a-b} \mathsf{SReLU}_{k_{a-b}}(\langle \omega_j^{a-b}, x\rangle - T_j^{a-b})$, we conclude that $\|\omega_j^{a-b}\| \leq 1/r$ and

$$\sum_{j=1}^{N/(a+1)} |\kappa_j^{a-b}| \leq C_1 \frac{\Big(C_g^{(0)} + C_g^{(2k_a+2)}\Big)}{N^{b/2}} .$$

In particular, we have $g^{\mathsf{rem},a-1}$ such that $C_{g^{\mathsf{rem},a-1}}^{(2)} + C_{g^{\mathsf{rem},a-1}}^{(0)} \leq C(C_g^{(0)} + C_g^{2k_a+2})/(N^{\frac{a}{2}})$. Therefore, by Theorem 16, there exists a random ReLU network with $N/(a+1)$ neurons which approximates $g^{\mathsf{rem},a-1}$ with output $\hat{g}^{(a)}$ such that:

1.
$$\mathbb{E}\int \left(g^{\mathsf{rem},a-1}(x)-\hat{g}^{(a)}(x)\right)^2\zeta(dx) \le C\frac{\left(C^{(2)}_{g^{\mathsf{rem},a-1}}+C^{(0)}_{g^{\mathsf{rem},a-1}}\right)^2}{N} \le C\frac{\left(C^{(0)}_g+C^{2k_a+2}_g\right)^2}{N^{a+1}}\,.$$

2.
$$g^{\mathsf{rem},a-1}-\mathbb{E}\hat{g}^{(a)}_j(x)=0\,,$$
where $\hat{g}^{(a)}_j$ is the j-th component of $\hat{g}^{(a)}$.

3. Assuming $\hat{g}^{(a)}(x)=\sum_{j=1}^{N/(a+1)}\kappa^0_j\mathsf{ReLU}(\langle\omega^0_j,x\rangle-T^0_j)$, it is clear that $\|\omega^0_j\|\le 1/r$:
$$\sum_{j=1}^{N/(a+1)}|\kappa^0_j|\le C_1\frac{C^{(0)}_g+C^{(2k_a+2)}_g}{N^{a/2}}\,.$$

We note that we have chosen the SReLU_{k_b} units in a non-random fashion through Theorem 7 whereas we have chosen the last $\frac{N}{a+1}$ ReLU units randomly using Theorem 16. Therefore, the expectation above is only with respect to the randomness of the ReLU units. It is clear that $g^{\mathsf{rem},a-1}(x)-\hat{g}^{(a)}(x)=g(x)-\left(\sum_{b=0}^a\hat{g}^{(b)}(x)\right)$ whenever $x\in B^2_q(r)$ and $\sum_{b=0}^a\hat{g}^{(b)}(x)$ is the output of a two-layer network with N non-linear units containing ReLU and SReLU_k units for $k\in\{k_1,\ldots,k_a\}$. We conclude items 1 and 2 in the statement of the lemma. The sum of the absolute values of the coefficients is $\sum_{b=0}^a\sum_{j=1}^{N/(a+1)}|\kappa^b_j|\le C_1(C^{(0)}_g+C^{(2k_a+2)}_g)$ as is clear from the discussion above. ■

F.2. Proof of Theorem 9

We first note that whenever $x\in B^2_d(r)$, $\langle x,B_i\rangle\in B^2_q(r)$. We assume that $N/(a+1)m$ is an integer. In Theorem 8, we take $g=f_i$ and replace N with N/m. We pick the weights ω_i inside the SReLU_k and ReLU units to be in $\mathrm{span}(B_i)$ instead of $\mathbb{R}^q$ and the replace the distribution $\zeta(dx)$ by $\zeta(\langle dx,B_i\rangle)$, which is the measure induced by ζ over $\mathrm{span}(B_i)$. We conclude that there exists a random neural network NN_i with 1 nonlinear layer whose output is $\hat{f}_i(x)$ such that:

1. For every $x\in B^2_q(r)$,
$$\mathbb{E}\hat{f}_i(x)=f_i(\langle x,B_i\rangle)$$

2.
$$\mathbb{E}\int\left(f_i(\langle x,B_i\rangle)-\hat{f}_i(x)\right)^2\zeta(dx)\le C_0\frac{Mm^{a+1}}{N^{a+1}}\,.$$

We construct the random neural networks NN_i independently for $i\in[m]$. We juxtapose these m neural networks and average their outputs to obtain the estimator $\hat{f}(x):=$

$\frac{1}{m}\sum_{i=1}^{m}\hat{f}_i(x)$. Now

$$\begin{aligned}
&\mathbb{E}\int\Big(\frac{1}{m}\sum_{i=1}^{m}f_i(\langle x,B_i\rangle)-\hat{f}_i(x)\Big)^2\zeta(dx)\\
&=\frac{1}{m^2}\sum_{i,j\in[m]}\mathbb{E}\int\Big(f_i(\langle x,B_i\rangle)-\hat{f}_i(x)\Big)\Big(f_j(\langle x,B_i\rangle)-\hat{f}_j(x)\Big)\zeta(dx)\\
&=\frac{1}{m^2}\sum_{i,j\in[m]}\int\mathbb{E}\Big(f_i(\langle x,B_i\rangle)-\hat{f}_i(x)\Big)\Big(f_j(\langle x,B_i\rangle)-\hat{f}_j(x)\Big)\zeta(dx)\\
&=\frac{1}{m^2}\sum_{i\in[m]}\int\mathbb{E}\Big(f_i(\langle x,B_i\rangle)-\hat{f}_i(x)\Big)^2\zeta(dx)\\
&\leq C_0\frac{m^aM}{N^{a+1}}\,. \qquad (76)
\end{aligned}$$

In the fourth step we have used the fact that $\hat{f}_j(x)$ and $\hat{f}_i(x)$ are independent when $i\neq j$. Because the above bound holds in expectation, it must hold for some configuration. ■

F.3. Proof of Theorem 10

Consider the low dimensional polynomial defined in Equation (2). Define the following orthonormal set associated with each V in the summation:

1. $B_V=\{e_j : V(j)\neq 0\}$ where e_j are the standard basis vectors in $\mathbb{R}^d$, if $|\{e_j : V(j)\neq 0\}|=q$.
2. Otherwise, let $w=q-|\{e_j : V(j)\neq 0\}|$. Otherwise, draw distinct $e_{j_1},\ldots,e_{j_w}\notin\{e_j : V(j)\neq 0\}$ from some arbitrary fixed procedure and define $B_V=\{e_j : V(j)\neq 0\}\cup\{e_{j_1},\ldots,e_{j_w}\}$. This ensures that $|B_V|=q$.

Clearly, p_V can be seen as a function over $\text{span}(B_V)$ which is isomorphic to $\mathbb{R}^q$. Since we are only interested in $x\in[0,1]^d$, it follows that $\langle x,B_V\rangle\in[0,1]^q\subseteq B_q^2(\sqrt{q})$. We can also modify $p_V(x)$ to $p_V(x)\gamma\left(\|\langle B_V,x\rangle\|^2/q\right)$ where $\gamma\in\mathcal{S}(\mathbb{R})$ is the bump function defined in Section B such that $\gamma(t)=1$ for $t\in[-1,1]$, $\gamma\geq 0$ and $\gamma(t)=0$ for $|t|\geq 2$. Therefore, $p_V(x)\gamma\left(\|\langle B_V,x\rangle\|^2/q\right)$, when seen as a function over $\text{span}(B_V)$, is itself a Schwartz function and it is equal to $p_V(x)$ whenever $\langle x,B_V\rangle\in B_q^2(\sqrt{q})$. Without any loss, we replace $p_V(x)$ with $p_V(x)\gamma\left(\|\langle B_V,x\rangle\|^2/q\right)$ in Equation (2). We note that the low degree polynomials defined above are an instance of the low dimensional function defined in Equation (1), but without the factor of m. In Theorem 32, we will just multiply throughout by a factor m - for both f and the estimator $\hat{f}$. The only change which occurs in the guarantees is that the error is multiplied by m^2 and the co-efficients κ_j in the statement of the theorem are multiplied by m. In this case, we take $m=\binom{q+d}{q}$. Fix an $a\in\mathbb{N}\cup\{0\}$ and take $N\geq(a+1)m$ such that $N/(a+1)m\in\mathbb{N}$. Consider the Fourier norm of p_V when seen as a function over $\text{span}(B_V)$. Clearly p_V is a Schwartz function and the Fourier norm defined in Equation (38) exists and is finite for every $l=2k_a^S+2$ (l is as used in Equation (38)). Therefore, we set $H:=\sup_V(S^0_{p_V}+S^{(2k_a^S+2)}_{p_V})^2<\infty$. It is clear that H depends only on q and a. Now, the corresponding squared Fourier norms for J_Vp_V, denoted by M_V satisfies $M_V\leq HJ_V^2$ (where

M_V is the analogue of M_i as defined in Theorem 32). Consider the sampling procedure given in Theorem 32: since the bases B_V (the analogues of B_i in the statement of the theorem) are known explicitly, this sampling can be done without the knowledge of the polynomial. Now, by a direct application of Theorem 32, we conclude the statement of Theorem 10. ■