NEURAL TANGENT KERNELS, TRANSPORTATION MAPPINGS, AND UNIVERSAL APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper establishes rates of universal approximation for the shallow *neural tangent kernel (NTK)*: network weights are only allowed microscopic changes from random initialization, which entails that activations are mostly unchanged, and the network is nearly equivalent to its linearization. Concretely, the paper has two main contributions: a generic scheme to approximate functions with the NTK by sampling from *transport mappings* between the initial weights and their desired values, and the construction of transport mappings via Fourier transforms. Regarding the first contribution, the proof scheme provides another perspective on how the NTK regime arises from rescaling: redundancy in the weights due to resampling allows individual weights to be scaled down. Regarding the second contribution, the most notable transport mapping asserts that roughly $1/\delta^{10d}$ nodes are sufficient to approximate continuous functions, where δ depends on the continuity properties of the target function. By contrast, nearly the same proof yields a bound of $1/\delta^{2d}$ for shallow ReLU networks; this gap suggests a tantalizing direction for future work, separating shallow ReLU networks and their linearization.

1 MAIN RESULT AND OVERVIEW

Consider functions computed by a single ReLU layer, meaning

$$x \mapsto \sum_{j=1}^{m} s_j \sigma\left(\langle w_j, x\rangle + b_j\right), \tag{1.1}$$

where $\sigma(z) := \max\{0, z\}$. While shallow networks are celebrated as being *universal approximators* (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989) — they approximate continuous functions arbitrarily well over compact sets — what is more shocking is that gradient descent can learn the parameters to these networks, and they generalize (Zhang et al., 2016).

Working towards an understanding of gradient descent on shallow (and deep!) networks, researchers began investigating the *neural tangent kernel (NTK)* (Jacot et al., 2018; Du et al., 2018; Allen-Zhu et al., 2018), which replaces a network with its linearization at initialization, meaning

$$x \mapsto \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^{m} s_j \langle \tau_j, \tilde{x}\rangle \sigma'\left(\langle \tilde{w}_j, \tilde{x}\rangle\right), \quad \text{where } \tilde{x} = (x, 1) \in \mathbb{R}^{d+1},\ \tilde{w} = (w, b) \in \mathbb{R}^{d+1}; \tag{1.2}$$

here each $\tilde{w}_j = (w_j, b_j)$ is frozen at Gaussian initialization (henceforth the bias is collapsed in for convenience), and each *transported weight* τ_j is microscopically close to the corresponding initial weight $\tilde{w}_j$, concretely $\|\tau_j - \tilde{w}_j\| = \mathcal{O}(1/\epsilon\sqrt{m})$, where $\epsilon > 0$ is a parameter and the scaling $\epsilon/\sqrt{m}$ is conventional in this literature (Allen-Zhu et al., 2018).

As eq. (1.2) is merely *affine* in the parameters, it is not outlandish that gradient descent can be analyzed. What *is* outlandish is firstly that gradient descent on eq. (1.1) with small learning rate will track the behavior of eq. (1.2), and secondly the weights hardly change *as a function of* m, specifically $\|\tau_j - \tilde{w}_j\|_2 = \mathcal{O}(1/\epsilon\sqrt{m})$.

Contributions. This work provides rates of function approximation for the NTK as defined in eq. (1.2), moreover in the "NTK setting": the transported weights must be near initialization, meaning $\|\tau_j - \tilde{w}_j\| = \widetilde{\mathcal{O}}(1/\epsilon\sqrt{m})$. In more detail:

Continuous functions (cf. Theorem 1.5). The main theorem packages the primary tools here to say: the NTK can approximate continuous functions arbitrarily well, so long as the width is at least $1/\delta^{10d}$, where δ depends on continuity properties of the target function; moreover, the transports satisfy $\|\tau_j - \tilde{w}_j\| = \mathcal{O}(1/\epsilon\sqrt{m})$, and the ReLU network (eq. (1.1)) and its linearization (eq. (1.2)) stay close. Re-using many parts of the proof, a nearly-optimal rate $1/\delta^{2d}$ is exhibited for ReLU networks in Theorem E.1; this gap between ReLU networks and their NTK poses a tantalizing gap for future work.

Approximation via sampling of transport mappings. The first component of the proof of Theorem 1.5, detailed in Section 2, is a procedure which starts with an infinite width network, and describes how sampling introduces redundancy in the weights, and automatically leads to the desired microscopic transports $\|\tau_j - \tilde{w}_j\| = \mathcal{O}(1/\epsilon\sqrt{m})$. As detailed in Theorem 2.1, the error between the infinite width and sampled networks is $\widetilde{\mathcal{O}}(\epsilon + 1/\sqrt{m})$. In this way, the analysis provides another perspective on the scaling behavior and small weight changes of the NTK.

Construction of transport mappings. The second component of the proof of Theorem 1.5, detailed in Section 3, is to construct explicit transport mappings for various types of functions. As detailed in Lemma 3.3, approximating continuous functions proceeds by constructing an infinite width network not directly for the target function f, but instead its convolution $f * G_\alpha$ with a Gaussian G_α with tiny variance α^2. Care is needed in order to obtain a rate of the form $1/\delta^{\mathcal{O}(d)}$, rather than, say, $1/\delta^{\mathcal{O}(d/\delta)}$. The main constructions are based on Fourier transforms.

Rounding out the organization of this paper: this introduction will state the main summarizing result and its intuition, and then close with related work; Section 4 will describe certain odds and ends for approximating continuous functions which were left out from the main tools in Section 2 and Section 3; Section 5 sketches abstract approaches to constructing transport mappings, including ones based on a corresponding Hilbert space; Section 6 will conclude with open problems and related discussion. Proofs are sketched in the paper body, but details are deferred to the appendices.

1.1 Basic notation, intuition, and main result

The NTK views networks as finite width realizations of intrinsically *infinite width* objects. In order to transport an infinite number of parameters away from their initialization, one option is to use a *transport mapping* $\mathcal{T} : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ to show where weights should go:

$$x \;\mapsto\; \mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \Phi(x;\tilde{w}) \right\rangle \;=\; \mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \tilde{x} \right\rangle \sigma'(\langle \tilde{w}, \tilde{x} \rangle) \;=\; \mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \tilde{x} \right\rangle \mathbb{1}\left[\langle \tilde{w}, \tilde{x} \rangle \geq 0\right],$$

where $\Phi(x;\tilde{w}) = \tilde{x}\sigma'(\langle \tilde{w}, \tilde{x} \rangle)$ is a *random feature* representation of x (Rahimi & Recht, 2008). This abstracts the individual transported weights $(\tau_j)_{j=1}^m$ from before into transported weights defined over arbitrary weights $\tilde{w} \in \mathbb{R}^{d+1}$. These (augmented) weights $\tilde{w} = (w, b)$ (with weight $w \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$) will always be distributed according to a standard Gaussian with identity covariance, with G denoting the density and probability law simultaneously.

A key message of this work, developed in Section 2, is (a) the infinite width network can be sampled to give rates of approximation by finite width networks, (b) the microscopic adjustments of the NTK setting arise naturally from the sampling process! Indeed, letting $s \in \{-1, +1\}$ denote a uniformly distributed random sign,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \Phi(x;\tilde{w}) \right\rangle &= \mathop{\mathbb{E}}_{\tilde{w},s} \left\langle s^2 \mathcal{T}(\tilde{w}) + s\tilde{w}\boldsymbol{\epsilon\sqrt{m}}, \Phi(x;\tilde{w}) \right\rangle && \because \mathbb{E}\, s^2 = 1, \mathbb{E}\, s = 0 \\
&\approx \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{j=1}^m \left\langle s_j^2 \mathcal{T}(\tilde{w}_j) + s_j \tilde{w}_j \boldsymbol{\epsilon\sqrt{m}}, \Phi(x;\tilde{w}_j) \right\rangle && \text{sampling } (w_j, s_j) \\
&= \frac{\boldsymbol{\epsilon}}{\boldsymbol{\sqrt{m}}} \sum_{j=1}^m \left\langle \frac{s_j \mathcal{T}(\tilde{w}_j)}{\boldsymbol{\epsilon\sqrt{m}}} + \tilde{w}_j, s_j \Phi(x;\tilde{w}_j) \right\rangle.
\end{aligned} \tag{1.3}$$

As highlighted by the bolded terms: increasing the width m corresponds to resampling, and allows the transported weights to be scaled down! Indeed, the distance moved is $\mathcal{O}(1/\epsilon\sqrt{m})$ *by construction*.

To this end, for convenience define

$$\begin{aligned} \tau_j &:= \mathcal{T}_{\boldsymbol{\epsilon}}(\tilde{w}_j, s_j) := \frac{s_j \mathcal{T}(\tilde{w}_j)}{\boldsymbol{\epsilon}\sqrt{\boldsymbol{m}}} + \tilde{w}_j \mathbb{1}[\|\tilde{w}_j\| \le R], \\ \phi_j(x) &:= \Phi_{\boldsymbol{\epsilon}}(x; \tilde{w}_j, s_j) := \frac{\boldsymbol{\epsilon} s_j}{\sqrt{\boldsymbol{m}}} \Phi(x; \tilde{w}_j) = \frac{\boldsymbol{\epsilon} s_j}{\sqrt{\boldsymbol{m}}} \tilde{x} \sigma'(\langle \tilde{w}_j, \tilde{x} \rangle), \end{aligned} \tag{1.4}$$

where R is a truncation radius included for purely technical reasons. The transport mappings constructed in Section 3 satisfy $B := \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\| < \infty$, and thus $\max_j \|\tau_j - \tilde{w}_j\| \le {}^{B}/_{\epsilon\sqrt{m}}$ by construction as promised (with high probability).

The key message of Section 2 is to control the deviations of this process, culminating in Theorem 2.1 and also Theorem 1.5 below, which yields upper bounds on the width necessary to approximate infinite width networks. The notion of approximation here will follow (Barron, 1993) and use the $L_2(P)$ metric, where P is a probability measure on the ball $\{x \in \mathbb{R}^d : \|x\| \le 1\}$:

$$\|h\|_{L_2(P)} = \sqrt{\int h(x)^2 \, \mathrm{d}P(x)}.$$

Additionally $\|h\|_{L_2} = \sqrt{\int h(x)^2 \, \mathrm{d}x}$ and $\|h\|_{L_1} = \int |h(x)| \, \mathrm{d}x$ will respectively denote the usual L_2 and L_1 metrics over functions on $\mathbb{R}^d$.

Theorem 1.5 (Simplification of Theorems 2.1 and 4.3)**.** *Let continuous function $f : \mathbb{R}^d \to \mathbb{R}$ be given, along with $\delta \in (0, 1]$ so that $|f(x) - f(x')| \le \epsilon$ whenever $\|x - x'\|_2 \le \delta$ and $\max\{\|x\|, \|x'\|\} \le 1 + \delta$. Let P be any probability distribution over $\|x\| \le 1$. Then there exists a transport mapping $\mathcal{T}$ (defining $\mathcal{T}_\epsilon$ and τ_j as in eq.* (1.4)*) and associated scalars*

$$B := \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2 = \widetilde{\mathcal{O}}\left(\frac{M^5 d^{(5d+9)/2}}{\epsilon^4 \delta^{5(d+1)}} \right), \qquad \text{where } M := \sup_{\|x\| \le 1+\delta} |f(x)|,$$

so that with probability at least $1 - 3\eta$ over Gaussian weights $(\tilde{w}_j)_{j=1}^m$ and uniform signs $(s_j)_{j=1}^m$, then $\max_j \|\tau_j - \tilde{w}_j\| \le {}^{B}/_{\epsilon\sqrt{m}}$, and

$$\left\| f - \sum_{j=1}^m \langle \tau_j, \phi_j(\cdot) \rangle \right\|_{L^2(P)} \le \widetilde{\mathcal{O}}\left(\left[\frac{B}{\sqrt{m}} + \epsilon\sqrt{d} \right] \sqrt{\ln(1/\eta)} \right),$$

$$\left\| \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^m s_j \sigma(\langle \tau_j, \tilde{x} \rangle) - \sum_{j=1}^m \langle \tau_j, \phi_j(\cdot) \rangle \right\|_{L_2(P)} \le \widetilde{\mathcal{O}}\left(\left[\frac{B^2}{\epsilon m^{3/2}} + \frac{B\sqrt{d}}{m} + \frac{B}{\sqrt{m}} + \epsilon\sqrt{d} \right] \sqrt{\ln(1/\eta)} \right).$$

In words: given an arbitrary target function f and associated continuity parameter δ, width $(B/\epsilon)^2 = \widetilde{\mathcal{O}}({}^{d^{5d+9}}/_{\epsilon^{10}\delta^{10(d+1)}})$ suffices for error $\widetilde{\mathcal{O}}(\epsilon)$, parameters are close to initialization, and the NTK and the original network behave similarly. The randomized construction does not merely give existence, but holds with high probability: the sampling process is thus in a sense robust, and may be used algorithmically!

As provided in Theorem 4.5, elements of the proof of Theorem 1.5 can be extracted and converted into a direct approximation rate of continuous functions by ReLU networks, and the rate becomes $\widetilde{\mathcal{O}}({}^{d^{d+2}}/_{\epsilon^2\delta^{2d+2}})$. Since this rate is nearly tight, together these rates pose an interesting question: is there a purely approximation-theoretic gap between shallow ReLU networks and their NTK?

1.2 Related work

Optimization literature; the NTK. This work is motivated and inspired by the optimization literature, which introduced the NTK to study gradient descent in a variety of nearly-parallel works (Jacot et al., 2018; Du et al., 2018; Du et al., 2018; Allen-Zhu et al., 2018; Arora et al., 2019; Oymak & Soltanolkotabi, 2019; Li & Liang, 2018; Cao & Gu, 2019). These works require the network width to be polynomial in n, the size of the training set; by contrast, the analysis here studies closeness in function space, and the width instead scales with properties of the target function.

One close relative to the present work is that of (Chizat & Bach, 2019), which provides an abstract proof scheme following the preceding works, and explains the microscopic change of the weights as a consequence of the scaling $\epsilon/\sqrt{m}$. This is consistent with the resampling perspective here, as summarized in eq. (1.3).

Random features and the mean-field perspective. The *random features* perspective (Rahimi & Recht, 2008) studies a related *convex* problem: similarly to the NTK, the activations $\sigma'(\langle \tilde{w}_j, \tilde{x} \rangle)$ are held fixed, and what are trained are scalar weights $a_j \in \mathbb{R}$ on each feature. The Fourier transport map construction used both for the NTK here in Theorem 1.5 and for shallow networks in Theorem E.1 proceeds by constructing exactly such a reweighting, and thus the present work also establishes universal approximation properties of random features. A related perspective is presented in the mean-field literature, which relate gradient descent on $(\tilde{w}_j)_{j=1}^m$ to a Wasserstein flow in the space of distributions on these features (Chizat & Bach; Mei et al., 2018). The analysis here does not have any explicit ties to the mean-field literature, however it is interesting and suggestive that transport mappings appear in both.

Approximation literature. The closest prior work is due to Barron (1993), who gave good rates of approximation for functions $f : \mathbb{R}^d \to \mathbb{R}$ when the associated quantity $\int \|w\| \cdot |\hat{f}(w)| \, \mathrm{d}w$ is small, where $\hat{f}$ denotes the Fourier transform of f. The proofs in Section 3 will use elements from the proofs in (Barron, 1993), but with many distinct components, and thus it is interesting that the same quantity $\int \|w\| \cdot |\hat{f}(w)| \, \mathrm{d}w$ arises once again. Like the work of (Barron, 1993), the present work also chooses to approximate in the $L_2(P)$ metric. Standard classical works in this literature are general universal approximation guarantees without rates or attention to the weight magnitudes (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989; Leshno et al., 1993). The rate given here of roughly $1/\delta^{2d}$ in Theorem 4.5 does not seem to appear rigorously in prior work, though it is mentioned as a consequence of a proof in (Mhaskar & Micchelli, 1992), who also take the approach of approximation via Gaussian convolutions; the use of convolutions is not only standard (Wendland, 2004), it is moreover classical, having been used in Weierstrass's original proof (Weierstrass, 1885).

Many related works use a RKHSes directly. Sun et al. (2018) prove universal approximation (with rates) via an RKHS, however they do not consider the NTK (or the NTK setting of small weight changes). Bach (2017a) (see also (Bach, 2017b; Basri et al., 2019; Bietti & Mairal, 2019)) studies a variety of questions related to function fitting with the random features model, including establishing rates of approximation for Lipschitz functions on the surface of the sphere (with a few further conditions); the rates are better than those here (roughly $\Theta(1/\delta^{d/2})$), however they do not consider the NTK setting, meaning either the setting of small changes from initialization nor the linearization.

Another close parallel work studies exact representation power of infinite width networks, developing representations for functions with $\Omega(d)$ derivatives (Ongie et al., 2019); similarly, the constructions here use an exact representation result for Gaussian convolutions, as developed in Section 3.

Regarding lower bounds from the literature, there are two lower bounds of the form $1/\delta^{d/2}$ for general shallow networks, not necessarily in the NTK setting (Yarotsky, 2016; Bach, 2017a). Interestingly, Yarotsky (2016) also presents a lower bound of $1/\delta^d$ for approximations whose parameters vary *continuously* with the target function; this seems to hold for the Fourier constructions here in Section 3, though an argument needs to be made for the sampling step.

2 Sampling from a transport

This section establishes that by sampling from an infinite width NTK, the resulting finite width NTK is close in $L_2(P)$ both to the infinite width idealization, and also to the finite width non-linearized ReLU network; moreover, the sampling process introduces redundancy in the weights, allowing them to be scaled down and lie close to initialization.

Theorem 2.1. *Suppose* $B \geq \max\left\{2, \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2\right\}$*, and set* $R := \sqrt{d+1} + 2\sqrt{\ln(m/\eta)}$*. With probability at least* $1 - 3\eta$*, then* $\max_j \|\tau_j - \tilde{w}_j\| \leq B/\epsilon\sqrt{m}$*, and*

$$\left\| \sum_{j=1}^{m} \left\langle \tau_j, \phi_j(\cdot) \right\rangle - \mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \Phi(\cdot; \tilde{w}) \right\rangle \right\|_{L^2(P)} \leq 2\left(\frac{B}{\sqrt{m}} + \epsilon R\right)\left[1 + \sqrt{\ln(1/\eta)}\right], \tag{2.2}$$

$$\left\| \sum_{j} \left\langle \tau_j, \phi_j(\cdot) \right\rangle - \sum_{j} \frac{s_j \epsilon}{\sqrt{m}} \sigma(\langle \tau_j, \cdot \rangle) \right\|_{L_2(P)} \leq 2\left(\frac{B^2}{\epsilon m^{3/2}} + \frac{BR}{m} + \frac{B}{\sqrt{m}} + \epsilon R\right)\left[1 + \sqrt{\ln(1/\eta)}\right]. \tag{2.3}$$

As discussed in the introduction, $\max_j \|\tau_j - \tilde{w}_j\| \leq B/\epsilon\sqrt{m}$ is essentially by construction. Next, recall the sampling derivation in eq. (1.3), restated here as a lemma for convenience, the notation $(\mathcal{W}, S)$ collecting all random variables together, meaning $\mathcal{W} = (\tilde{w}_1, \ldots, \tilde{w}_m)$ and $S = (s_1, \ldots, s_m)$.

Lemma 2.4. $\mathop{\mathbb{E}}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \Phi(x; \tilde{w}) \right\rangle = \mathop{\mathbb{E}}_{\widetilde{W}, S} \sum_j \left\langle \mathcal{T}_\epsilon(\tilde{w}_j, s_j), \Phi_\epsilon(x; \tilde{w}_j, s_j) \right\rangle = \mathop{\mathbb{E}}_{\widetilde{W}, S} \sum_j \left\langle \tau_j, \phi_j(x) \right\rangle.$

The proof of eq. (2.2) now follows from the classical Maurey sampling lemma (Pisier, 1980), which was also used in the related work by Barron (1993). The following version additionally includes a high probability control, which results from an application of McDiarmid's inequality. Applying the following sampling lemma to the present setting, the deviations will scale with $B := \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2$.

Lemma 2.5 (Maurey)**.** *Let functions* $\{g(\cdot; v) : v \in \mathcal{V}\}$ *be given, where* $\mathcal{V} \subseteq \mathbb{R}^p$ *is a set of possible parameters. Let* ν *be a probability measure over* $\mathcal{V}$*, let* $(v_1, \ldots, v_m)$ *be an iid random draw from* ν*, and define*

$$f(x) := \mathop{\mathbb{E}}_{v \sim \nu} g(x; v) \qquad \text{and} \qquad g_j(x) := g(x; v_j).$$

Then

$$\mathop{\mathbb{E}}_{((s_j, v_j))_{j=1}^m} \left\| f - \frac{1}{m} \sum_{j=1}^{m} g_j \right\|_{L^2(P)}^2 \leq \frac{1}{m} \mathop{\mathbb{E}}_{v} \left\| g(\cdot; v) \right\|_{L_2(P)}^2 \leq \frac{1}{m} \sup_{v \in \mathcal{V}} \left\| g(\cdot; v) \right\|_{L_2(P)}^2,$$

and with probability at least $1 - \eta$*,*

$$\left\| f - \frac{1}{m} \sum_{j=1}^{m} g_j \right\|_{L^2(P)} \leq \sup_{v \in \mathcal{V}} \|g(\cdot; v)\|_{L_2(P)} \left[\frac{1 + \sqrt{2\ln(1/\eta)}}{\sqrt{m}}\right].$$

Concretely, here $g_j(x) = m \left\langle \tau_j, \phi_j(x) \right\rangle$, and $\sup_{v \in \mathcal{V}} \|g(\cdot; \tilde{w})\|_{L_2(P)} \leq \sqrt{2} \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2 = \mathcal{O}(B + R\epsilon\sqrt{m})$ by Cauchy-Schwarz. Before continuing, note also that there are other proof schemes attaining similar bounds (Bach, 2017a;b, Proposition 1), and that similar bounds are possible for the uniform norm, albeit with more sensitivity to the basis functions g (cf. Lemma B.2).

Turning now to the final bound in eq. (2.3), the first step is to note by positive homogeneity of the ReLU that $\sigma(\langle \tau_j, \tilde{x} \rangle) = \langle \tau_j, \tilde{x} \rangle \, \sigma'(\langle \tau_j, \tilde{x} \rangle)$, thus

$$\sum_{j=1}^{m} \left\langle \tau_j, \phi_j \right\rangle - \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^{m} s_j \sigma(\langle \tau_j, \tilde{x} \rangle) = \sum_{j=1}^{m} \left\langle \tau_j, \phi_j - \frac{s_j \epsilon}{\sqrt{m}} \tilde{x} \sigma'(\langle \tau_j, \tilde{x} \rangle) \right\rangle,$$

which boils down to checking the difference in activations, namely $\sigma'(\langle \tilde{w}_j, \tilde{x} \rangle) - \sigma'(\langle \tau_j, \tilde{x} \rangle)$. As is standard in the NTK literature, since $\tau_j - \tilde{w}_j$ is (with high probability) microscopic compared to $\langle \tilde{w}_j, \tilde{x} \rangle$, the activations should also be close. The following lemma makes this precise.

Lemma 2.6. *For any* $x \in \mathbb{R}^p$*, if* $R \geq \sqrt{d} + 2\sqrt{\ln\left(\frac{\epsilon\sqrt{m\pi}}{B\sqrt{2}}\right)}$ *(as used in eq.* (1.4)*), then*

$$\mathop{\mathbb{E}}_{\tilde{w}} \left| \sigma'(\langle \tilde{w}, \tilde{x} \rangle) - \sigma'(\langle \mathcal{T}_\epsilon(\tilde{w}), \tilde{x} \rangle) \right| \leq \frac{2B\sqrt{2}}{\epsilon\sqrt{m\pi}}.$$

From here, the eq. (2.3) can be established with another application of Lemma 2.5. This completes the proof of Theorem 2.1 after an application of Gaussian concentration to ensure $\max_j \|\tilde{w}_j\| \leq R$. This also establishes the first half of Theorem 1.5.

3 CONCRETE TRANSPORT MAPPINGS VIA FOURIER TRANSFORMS

The previous section showed function approximation in the NTK setting *assuming* the existence of an infinite width NTK defined by a transport mapping $\mathcal{T}$; this section will close the gap by providing a variety of transport maps.

The transport mappings here will be constructed via Fourier transforms, with convention

$$\hat{f}(x) = \int \exp\left(-2\pi i x^\intercal w\right) f(x)\,\mathrm{d}x;$$

a few general properties are summarized in Appendix A. Interestingly, these transports are all *random feature transports*: they have the form $\mathcal{T}(\tilde{w}) = (0, \cdots, 0, p(\tilde{w}))$ where p is a *signed density* over random features, and $\mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \Phi(x;\tilde{w}) \right\rangle = \mathbb{E}_{\tilde{w}} p(\tilde{w}) \sigma'(\langle \tilde{w}, \tilde{x} \rangle)$. This perspective of a signed density will be used to prove universal approximation — again via sampling! — of shallow ReLU networks (and random features) later in Theorems 4.5 and E.1. (For constructions which are not based on random features, see Section 5.)

The first steps of the approach here follow a derivation due to Barron (1993). Specifically, the *inverse* Fourier transform gives a way to rewrite a function as an infinite with network with complex-valued activations $x \mapsto \exp(2\pi i x^\intercal w)$:

$$f(x) = \int \exp(2\pi i x^\intercal w) \hat{f}(w)\,\mathrm{d}w.$$

A key trick due to Barron (1993) is to force the right hand side to be real (since the left hand side is real): specifically, letting $\hat{f}(w) = |\hat{f}(w)| \exp(2\pi i \theta_f(w))$ with $|\theta_f(w)| \leq 1$ denote the radial decomposition of $\hat{f}$,

$$\begin{aligned}
\mathrm{Re} f(x) &= \mathrm{Re} \int \exp(2\pi i x^\intercal w) \hat{f}(w)\,\mathrm{d}w \\
&= \mathrm{Re} \int \exp(2\pi i x^\intercal w + 2\pi i \theta_f(w)) |\hat{f}(w)|\,\mathrm{d}w \\
&= \int \cos\left(2\pi(x^\intercal w + \theta_f(w))\right) |\hat{f}(w)|\,\mathrm{d}w.
\end{aligned}$$

After this step, the proofs diverge: the approach here is to use the fundamental theorem of calculus to rewrite $\cos$ in terms of σ':

$$\cos(z) - \cos(0) = -\int_0^z \sin(b)\,\mathrm{d}b = -\int_0^\infty \sin(b) \mathbb{1}[z - b \geq 0]\,\mathrm{d}b, = -\int_0^\infty \sin(b) \sigma'(z-b)\,\mathrm{d}b;$$

plugging this back in gives an explicit representation of f in terms of an infinite width threshold network! A similar approach can be used to obtain an infinite width ReLU network.

This is summarized in the following lemma, which includes a calculation of the error incurred by truncating the weights; this truncation is necessary when applying the sampling of Section 2. Interestingly, this truncation procedure leads to the quantity $\int \|w\| \cdot |\hat{f}(w)|\,\mathrm{d}w$, which was *explicitly* introduced as a key quantity by Barron (1993) via a different route, namely of introducing a factor $\|w\|$ to enforce decay on $\cos$.

Lemma 3.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be given with Fourier transform $\hat{f}$ and truncation radius $r \in [0, \infty]$.*

1. *Define infinite width threshold network*

$$\begin{aligned}
F_r(x) := f(0) &+ \int |\hat{f}(w)| \cos\left(2\pi(\theta_f(w) - \|w\|)\right)\mathrm{d}w \\
&+ 2\pi \int \sigma'(\langle \tilde{w}, \tilde{x} \rangle) |\hat{f}(w)| \sin(2\pi(\theta_f(w) - b)) \mathbb{1}[|b| \leq \|w\| \leq r]\,\mathrm{d}\tilde{w}.
\end{aligned}$$

For any $\|x\| \leq 1$, $F_\infty = f$ and $\left|f(x) - F_r(x)\right| \leq 4\pi \int_{\|w\|>r} \|w\| \cdot |\hat{f}(w)|\,\mathrm{d}w$.

2. *Define infinite width ReLU network*

$$Q_r(x) := f(0) + \int |\hat{f}(w)| \left[\cos(2\pi(\theta_f(w) - \|w\|)) - 2\pi\|w\| \sin(2\pi(\theta_f(w) - \|w\|))\right] \mathrm{d}w$$
$$+ x^\intercal \int w|\hat{f}(w)| \,\mathrm{d}w$$
$$+ 4\pi^2 \int \sigma(\tilde{w}^\intercal \tilde{x})|\hat{f}(w)| \cos(2\pi(\|w\| - b))\mathbb{1}[|b| \le \|w\| \le r] \,\mathrm{d}\tilde{w}.$$

For any $\|x\| \le 1$, $Q_\infty = f$ *and* $\left|f(x) - Q_r(x)\right| \le 12\pi^2 \int_{\|w\|>r} \|w\|^2 \cdot |\hat{f}(w)| \,\mathrm{d}w$.

(The second part of the shows that the same technique allows functions to be written with *equality* as ReLU networks; this is included as a curiosity and used in a few places in the appendices, but is not part of the main NTK story.)

The preceding constructions immediately yield transport mappings from Gaussian initialization to the function f in a brute-force way: by introducing the fraction $G(\tilde{w})/G(\tilde{w})$, calling the numerator part of the integration measure, and the denominator part of the integrand. As stated before, these transport maps are random feature maps: they zero out the coordinates corresponding to x!

Lemma 3.2. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be given with Fourier transform* $\hat{f}$. *For any* $r \in [0, \infty]$, *define transport mapping* $\mathcal{T}_r(w, b) := (0, \ldots, 0, p_r(\tilde{w}))$ *with*

$$\mathcal{T}_r(w, b)_{d+1} = p_r(\tilde{w}) := 2\left[f(0) + \int |\hat{f}(v)| \cos(2\pi(\theta_f(v) - \|v\|)) \,\mathrm{d}v\right]$$
$$+ 2\pi \left(\frac{|\hat{f}(w)|}{G(\tilde{w})}\right) \cos(2\pi(\theta_f(w) - b))\mathbb{1}[|b| \le \|w\| \le r].$$

By this choice, for any $\|x\| \le 1$, $f(x) = \mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}_\infty(\tilde{w}), \Phi(x; \tilde{w})\right\rangle$, *and*

$$\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\|_2 \le 2\left|f(0)\right| + 2\int |\hat{f}(v)| \,\mathrm{d}v + 2\pi \sup_{\substack{\|w\| \le r \\ |b| \le \|w\|}} \frac{|\hat{f}(w)|}{G(\tilde{w})},$$
$$\left|f(x) - \mathbb{E}\left\langle \mathcal{T}(\tilde{w}), \Phi(x; \tilde{w})\right\rangle\right| \le 4\pi \int_{\|w\|>r} |\hat{f}(w)| \cdot \|w\| \,\mathrm{d}w.$$

The preceding construction may seem general, however it is quite loose, noting the final supremum term within $\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\|_2$; indeed, attempting to plug this construction into Theorem 2.1 does not yield the $1/\delta^{\mathcal{O}(d)}$ rate in Theorem 1.5, but instead a rate $1/\delta^{\mathcal{O}(d/\delta)}$, which is disastrously larger!

Interestingly, a fix is possible for special functions of the form $f * G_\alpha$, namely convolutions with Gaussians of coordinate-wise variance α^2. These are exactly the types of functions used in Section 4 to approximate continuous functions. The fix is simply to apply a change of variable so that, in a sense, the target function and the initialization distribution have similar units.

Lemma 3.3. *Let function* f, *variance* $\alpha^2 > 0$, *and* $r \in [0, \infty]$ *be given, and define* $f_\alpha := f * G_\alpha$ *and* $\phi := (2\pi\alpha)^{-1}$, *and transport mapping* $\mathcal{T}_r(w, b) := (0, \ldots, 0, p_r(\tilde{w}))$ *with*

$$\mathcal{T}_r(w, b)_{d+1} = p_r(\tilde{w}) := 2\left[f_\alpha(0) + \int |\hat{f}_\alpha(v)| \cos(2\pi(\theta_{f_\alpha}(v) - \|v\|)) \,\mathrm{d}v\right]$$
$$+ 2\pi(2\pi\phi^2)^{(d+1)/2} |\hat{f}(\phi w)| e^{b^2/2} \sin(2\pi(\theta_{f_\alpha}(\phi w) - b))\mathbb{1}[|b| \le \|w\| \le r].$$

Then $f_\alpha(x) = \mathbb{E}_{\tilde{w}} \left\langle \mathcal{T}_\infty(\tilde{w}), \Phi(x; \tilde{w})\right\rangle$ *for* $\|x\| \le 1$, *and for* $r \in [\sqrt{d}, \infty)$,

$$\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\| \le 2\left[M + (2\pi\phi^2)^{d/2} M_f \left(1 + \sqrt{2\pi^3\phi^2} e^{r^2/2}\right)\right],$$

where $M := \sup_x |f(x)|$, *and* $M_f = 1$ *when* $f_\alpha = G_\alpha$ *and* $M_f = \|f(\phi\cdot)\|_{L_1}$ *otherwise, and*

$$\sup_{\|x\| \le 1} \left|f(x) - \mathbb{E}\left\langle \mathcal{T}_r(\tilde{w}), \Phi(x; \tilde{w})\right\rangle\right| \le 4\pi(2\pi\phi^2)^{(d+1)/2} M_f(\sqrt{d} + 3) \exp\left(-(r - \sqrt{d})^2/4\right).$$

4 APPROXIMATING CONTINUOUS FUNCTIONS

The final piece needed to prove Theorem 1.5 is to show that a function f is close to its Gaussian convolution $f * G_\alpha$, at least when $\alpha > 0$ is chosen appropriately. This is a classical topic (Wendland, 2004), and indeed it was used in the original proof of the Weierstrass approximation theorem (Weierstrass, 1885). The treatment here will include enough detail necessary to yield explicit rates.

The following definition will be used to replace the usual (ϵ, δ) conditions associated with continuous functions with an exact quantity.

Definition 4.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be given, and define *modulus of continuity* ω_f as

$$\omega_f(\delta) := \sup\left\{f(x) - f(x') : \max\{\|x\|, \|x'\|\} \le 1 + \delta, \|x - x'\| \le \delta\right\}. \qquad \Diamond$$

If f is continuous, then ω_f (defined here over a compact set) is not only finite for all inputs, but moreover $\lim_{\delta\to 0} \omega_f(\delta) \to 0$. It is also possible to use this definition with discontinuous functions; note additionally that the convolution bounds in Section 3 only required an L_1 bound on the pre-convolution function f, and therefore the tools throughout may be applied to discontinuous functions, albeit with some care to their Fourier transforms!

Lemma 4.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$ be given, and define*

$$M := \sup_{\|x\|\le 1+\delta} |f(x)|, \qquad f_{|\delta}(x) := f(x)\mathbb{1}[\|x\| \le 1 + \delta], \qquad \alpha := \frac{\delta}{\sqrt{d} + \sqrt{2\ln(2M/\omega_f(\delta))}}.$$

Let G_α denote a Gaussian with the preceding variance α^2. Then

$$\sup_{\|x\|\le 1}\left|f - f_{|\delta} * G_\alpha\right| \le 2\omega_f(\delta).$$

The proof splits the integrand into two parts: points close to x, and points far from it. Points close to x must behave like $f(x)$ due to continuity, whereas points far from x are rare and do not matter due to the Gaussian convolution. The full details are in the appendix.

Lemma 4.2 can be combined with the transport for $f * G_\alpha$ from Section 3 to give a transport for approximating continuous functions.

Theorem 4.3. *As in Lemma 4.2, let $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$ be given, and define*

$$M := \sup_{\|x\|\le 1+\delta} |f(x)|, \qquad f_{|\delta}(x) := f(x)\mathbb{1}[\|x\| \le 1 + \delta],$$

$$\alpha := \frac{\delta}{\sqrt{d} + \sqrt{2\ln(2M/\omega_f(\delta))}} = \widetilde{\mathcal{O}}(\delta/\sqrt{d}), \qquad r := \sqrt{d} + 2\sqrt{\ln\left(\frac{4\pi M_f(\sqrt{d}+3)}{(2\pi\alpha^2)^{(d+1)/2}\omega_f(\delta)}\right)}.$$

Let G_α denote a Gaussian with the preceding variance α^2, and let $\mathcal{T}_r$ denote the truncated Fourier map constructed in Lemma 3.3 for $f_{|\delta} * G_\alpha$, with preceding truncation choice r. Then

$$\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\| = \widetilde{\mathcal{O}}\left(\|f_{|\delta}\|_{L_1}^5 \left(\frac{\sqrt{d}}{\delta}\right)^{5(d+1)} \left[\frac{\sqrt{d}}{\omega_f(\delta)}\right]^4\right),$$

$$\sup_{\|x\|\le 1}\left|f - \mathbb{E}_{\tilde{w}}\left\langle \mathcal{T}_r(\tilde{w}), \Phi(x;\tilde{w})\right\rangle\right| \le \omega_f(\delta).$$

This completes all the pieces needed to prove Theorem 1.5.

Proof of Theorem 1.5. Let f be given, and let $\mathcal{T}_r$ denote the transport mapping provided by Theorem 4.3 for $f_{|\delta} * G_\alpha$, whose various parameters match those in the statement of Theorem 1.5. The proof is completed by plugging $\mathcal{T}_r$ into Theorem 2.1, and simplifying by noting that $\epsilon \ge \omega_f(\delta)$ by definition, and $\|f_{|\delta}\|_{L_1} = \mathcal{O}(M)$ since $\delta \le 1$. $\square$

As mentioned earlier, the infinite width network constructed in Lemma 3.1 via inverse Fourier transforms can be used to succinctly prove (via Lemma 2.5 and Lemma 4.2) that threshold and ReLU networks are universal approximators, with a rate vastly improving upon that of Theorem 1.5.

Before stating the result, one more tool is needed: a sampling semantics for signed densities; see also (Bach, 2017a;b) for further development and references.

Definition 4.4. A sample from a signed (Lebesgue) density $p : \mathbb{R}^{d+1} \to \mathbb{R}$ with $\|p\|_{L_1} < \infty$ is a pair $(\tilde{w}, s)$ where $\tilde{w}$ is sampled from the probability density $|p|/\|p\|_{L_1}$, and $s := \operatorname{sgn}(p(\tilde{w}))$. Let $\mathbb{E}_p$ denote the corresponding expectation over $(\tilde{w}, s) \sim p$. $\Diamond$

This notion of signed sampling also has a corresponding Maurey lemma, and an analogue for the uniform norm; both are provided in Appendix B. The full detailed universal approximation theorems for threshold and ReLU networks are provided in Appendix E; a simplified form for threshold networks alone is as follows. In either case, the proof proceeds by applying signed density sampling bounds (e.g., appropriate generalizations of Lemma 2.5) to the infinite width networks constructed in Lemma 3.1. Curiously, the simplified bound stated here for threshold networks for the uniform norm is only a multiplicative factor $\sqrt{d}$ larger than the $L_2(P)$ bound in Theorem E.1.

Theorem 4.5 (Simplification of Theorem E.1)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$ be given, and define*

$$M := \sup_{\|x\| \le 1+\delta} |f(x)|, \qquad f_{|\delta}(x) := f(x)\mathbb{1}[\|x\| \le 1+\delta], \qquad \alpha := \frac{\delta}{\sqrt{d} + \sqrt{2\ln(2M/\omega_f(\delta))}}.$$

Then there exist $c \in \mathbb{R}$ and $p : \mathbb{R}^{d+1} \to \mathbb{R}$ with

$$|c| \le M + \|f_{|\delta}\|_{L_1}(2\pi\alpha^2)^{d/2}, \qquad \text{and} \qquad \|p\|_{L_1} \le 2\|f_{|\delta}\|_{L_1}\sqrt{\frac{2\pi d}{(2\pi\alpha^2)^{d+1}}},$$

so that, with probability $\ge 1 - 3\eta$ over $((s_j, \tilde{w}_j))_{j=1}^m$ drawn from p (cf. Definition 4.4),

$$\sup_{\|x\|\le 1}\left| f(x) - \left[c_1 + \frac{\|p\|_{L_1}}{m}\sum_{j=1}^m s_j\sigma'(\langle \tilde{w}_j, x\rangle)\right]\right| \le 2\omega_f(\delta) + \frac{\|p\|_{L_1}}{\sqrt{m}}\left[8\sqrt{d\ln(m)} + \sqrt{\ln(1/\eta)}\right].$$

5 Abstract transport mappings, and an RKHS

Section 3 provided *concrete* transport mappings via Fourier transforms: it was, for instance, easy to use these constructions to develop approximation rates for continuous functions. These constructions had a major weakness: they were random feature transport mappings, meaning they arguably did not fully utilize the transport sampling provided in Section 2. This section will develop one abstract approach via RKHSes, but first will revisit the random feature constructions.

Suppose $f(x) = \int p(\tilde{w})\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w}$ for some density p; as in the proof of Lemma 3.2, introducing the term $G(\tilde{w})/G(\tilde{w})$ gives $f(x) = \int \frac{p(\tilde{w})}{G(\tilde{w})}\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}G(\tilde{w})$, which is now in the desired form, however the ratio term can be large (and a truncation is needed to make it finite in Lemma 3.2). The refined construction in Lemma 3.3 achieved a better bound on $\sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|$ by being careful about the scaling of the Gaussian, and then standardizing it with a change-of-variable transformation, but still it yields a random feature transport.

Another approach would be to start from the second construction in Lemma 3.1, which writes $f(x) = \int p(\tilde{w})\sigma(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} = \int p(\tilde{w})\,\langle \tilde{w}, \tilde{x}\rangle\,\sigma'(\langle \tilde{w}, \tilde{x}\rangle)$, and thus build a transport around $\tilde{w} \mapsto p(\tilde{w})\tilde{w}$, which now uses all coordinates. This transport mapping is still just a rescaling, however, and does not lead to improvements when plugged into the other parts of this work.

Consider the following approach to building a general $\mathcal{T}$ and an associated *Reproducing Kernel Hilbert Space (RKHS)*. To start, define an inner product $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$ via $\|\mathcal{T}\|_{\mathcal{H}}^2 = \langle \mathcal{T}, \mathcal{T}\rangle_{\mathcal{H}} = \int \|\mathcal{T}(\tilde{w})\|_2^2\,\mathrm{d}G(\tilde{w}) = \|\mathcal{T}\|_{L_2(G)}^2$; to justify this Hilbert space, note that it gives rise to the usual kernel product (Cho & Saul, 2009), namely

$$(x, x') \;\mapsto\; \mathbb{E}_{\tilde{w}}\Phi(x;\tilde{w})^{\mathsf{T}}\Phi(x';\tilde{w}) = \left\langle \Phi(x;\cdot), \Phi(x';\cdot)\right\rangle_{\mathcal{H}},$$

and moreover our earlier predictors can be written as $\mathbb{E}_{\tilde{w}}\left\langle \mathcal{T}(\tilde{w}), \Phi(x;\tilde{w})\right\rangle = \left\langle \mathcal{T}, \Phi(x;\cdot)\right\rangle_{\mathcal{H}}$.

The utility of these definitions is highlighted in the following bounds; specifically, while a given $\mathcal{T}$ may have $\sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2 = \infty$, truncation can make this quantity finite (and thus Theorem 2.1 may be applied), and the approximation error can be bounded with $\|\mathcal{T}\|_{\mathcal{H}}$.

Proposition 5.1. *1. The* output-truncated *transport* $\mathcal{T}_B(\tilde{w}) := \mathcal{T}(\tilde{w})\mathbb{1}\left[\|\mathcal{T}(\tilde{w})\|_2 \le B\right]$ *has approximation error*

$$\sup_{\|x\|\le 1}\left|\left\langle \mathcal{T}_B, \Phi(x;\cdot)\right\rangle_{\mathcal{H}} - \left\langle \mathcal{T}, \Phi(x;\cdot)\right\rangle_{\mathcal{H}}\right| \le \frac{\|\mathcal{T}\|_{\mathcal{H}}^2\sqrt{2}}{B^2}.$$

2. The input-truncated *transport* $\mathcal{T}_r(\tilde{w}) := \mathcal{T}(\tilde{w})\mathbb{1}[\|\tilde{w}\| \le r]$ *with* $r > \sqrt{d}$ *has approximation error*

$$\sup_{\|x\|\le 1}\left|\left\langle \mathcal{T}_r, \Phi(x;\cdot)\right\rangle_{\mathcal{H}} - \left\langle \mathcal{T}, \Phi(x;\cdot)\right\rangle_{\mathcal{H}}\right| \le \frac{\|\mathcal{T}\|_{\mathcal{H}}\sqrt{2}}{e^{(r-\sqrt{d})^2/4}}.$$

Unfortunately, this formalism can not be applied to the pre-truncation mapping $\mathcal{T}_\infty$ from Lemma 3.3, since $\|\mathcal{T}_\infty\|_{\mathcal{H}} = \infty$. Consequently, this approach is left as an interesting direction for future work.

6 Open problems

The main open question is: how much can the rates $1/\delta^{2d}$ for ReLU networks and $1/\delta^{10d}$ for their NTK be tightened, and is there a genuine gap? If so, this separation would constitute an interesting, purely approximation-theoretic manifestation of a "beyond NTK" regime (Allen-Zhu & Li, 2019).

Another direction is to use the Fourier tools of Section 2, as well as other tools for constructing transportation maps, and identify function classes with good approximation rates by the NTK and by shallow networks, in particular rates with a merely polynomial dependence on dimension.

Connecting back to the optimization literature, the referenced NTK optimization works for the squared loss seem to require a width which scales with n, and the test error sometimes scales with detailed functions of the observed labels, which require a further argument to go to 0 (see, e.g., $y^{\mathsf{T}}(H^\infty)^{-1}y$ in (Arora et al., 2019)). Perhaps such quantities can be replaced with a function space or other approximation theoretic perspective on the conditional mean function (and samples thereof)?

Lastly, what are connections to *optimal* transport? It seems natural to choose $\mathcal{T}$ as an optimal transport, in which case one would hope the parameter $B := \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2$ can be small, and moreover easily bounded by the optimal transport cost, ideally in ways similarly easy to the bounding by the Hilbert norm in Proposition 5.1.

References

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? 2019. `arXiv:1905.10337 [cs.LG]`.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. *arXiv e-prints*, art. arXiv:1811.04918, Nov 2018.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. 2018. `arXiv:1811.03962 [cs.LG]`.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. *arXiv e-prints*, art. arXiv:1901.08584, Jan 2019.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. 2019. `arXiv:1901.08584 [cs.LG]`.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017a.

Francis R. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, 18:21:1–21:38, 2017b.

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

Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *NeurIPS*, 2019.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *NeurIPS*, 2019.

Yuan Cao and Quanquan Gu. Generalization Error Bounds of Gradient Descent for Learning Over-parameterized Deep ReLU Networks. 2019. `arXiv:1902.01384 [cs.LG]`.

Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. art. arXiv:1805.09545. `arXiv:1805.09545 [cs.LG]`.

Lenaic Chizat and Francis Bach. A Note on Lazy Training in Supervised Differentiable Programming. `arXiv:1812.07956v2 [math.OC]`, 2019.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *NIPS*, 2009.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. 2018. `arXiv:1811.03804 [cs.LG]`.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. 2018. `arXiv:1810.02054 [cs.LG]`.

Gerald B. Folland. *Real analysis: modern techniques and their applications*. Wiley Interscience, 2 edition, 1999.

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

Leonid Gurvits and Pascal Koiran. Approximation and learning of convex superpositions. In Paul Vitányi (ed.), *Computational Learning Theory*, pp. 222–236. Springer, 1995.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, July 1989.

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

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

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A Mean Field View of the Landscape of Two-Layers Neural Networks. art. arXiv:1804.06561, 2018. `arXiv:1804.06561 [cs.LG]`.

Hrushikesh N Mhaskar and Charles A Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied mathematics*, 13(3):350–373, 1992.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. 2019. `arXiv:1910.01635v1 [cs.LG]`.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. 2019. arXiv:1902.04674 [cs.LG].

Gilles Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire Analyse fonctionnelle (dit)*, pp. 1–12, 1980.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, pp. 1177–1184. 2008.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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

Martin J. Wainwright. UC Berkeley Statistics 210B, Lecture Notes: Basic tail and concentration bounds, Jan 2015. URL https://www.stat.berkeley.edu/~mjwain/stat210b/.

Karl Weierstrass. Über die analytische darstellbarkeit sogenannter willkürlicher functionen einer reellen veränderlichen. *Sitzungsberichte der Akademie zu Berlin*, pp. 633–639, 789–805, 1885.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. 2016. arXiv:1610.01145 [cs.LG].

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. arXiv:1611.03530 [cs.LG], 2016.

A TECHNICAL LEMMAS

Gaussian Concentration. The following lemma collects a few properties of Gaussian concentration needed throughout.

Lemma A.1. *Let $w \sim G_d$ be a standard Gaussian in $\mathbb{R}^d$, and let $r \geq \sqrt{d}$ be given.*

1. $\int \|w\| \, \mathrm{d}G(w) \leq \sqrt{d}$.
2. $\mathbb{P}[\|w\| > r] \leq \exp(-(r-\sqrt{d})^2/2)$*; alternatively, with probability at least* $1-\eta$, $\|w\| \leq \sqrt{d} + \sqrt{2\ln(1/\eta)}$.
3. $\int_{\|w\|>r} \|w\| \, \mathrm{d}G(w) \leq (r+2)\exp\left(-(r-\sqrt{d})^2/2\right) \leq 2(\sqrt{d}+3)\exp\left(-(r-\sqrt{d})^2/4\right)$.
4. $\int_{\|w\|>r} \|w\|^2 \, \mathrm{d}G(w) \leq 2(r+2)^2\exp\left(-(r-\sqrt{d})^2/2\right) \leq 2(\sqrt{d}+7)^2\exp\left(-(r-\sqrt{d})^2/4\right)$.

The more convenient form of some of the inequalities will need to following technical lemma.

Lemma A.2. *Given $b \geq 0$ and $c > 0$ and $a \geq 0$ with $a + b \geq 2c$ and $x \geq b$, then*

$$(x+a)\exp(-(x-b)^2/c) \leq (a+b)\exp(-(x-b)^2/(2c)),$$

and if moreover $a + b \geq 4c$,

$$(x+a)^2\exp(-(x-b)^2/c) \leq (a+b)^2\exp(-(x-b)^2/(2c))$$

Proof. Since $\ln(x+a) \le \ln(b+a) + (x-b)/(b+a)$,

$$\begin{aligned}(x+a)\exp(-(x-b)^2/c) &\le (a+b)\exp(-(x-b)^2/c + (x-b)/(b+a)) \\ &\le (a+b)\exp(-(x-b)^2/c + (x-b)^2/(2c)) \\ &\le (a+b)\exp(-(x-b)^2/(2c)).\end{aligned}$$

Similarly, multiplying the preceding Taylor expansion by 2,

$$\begin{aligned}(x+a)^2\exp(-(x-b)^2/c) &\le (a+b)^2\exp(-(x-b)^2/c + 2(x-b)/(b+a)) \\ &\le (a+b)^2\exp(-(x-b)^2/c + 2(x-b)^2/(4c)) \\ &\le (a+b)^2\exp(-(x-b)^2/(2c)).\end{aligned}$$

□

Proof of Lemma A.1.

1. By Jensen's inequality, $\int \|w\| \,\mathrm{d}G(w) \le \sqrt{\int \|w\|^2 \,\mathrm{d}G(w)} = \sqrt{d}$.

2. The claim follows from Gaussian concentration with Lipschitz mappings (Wainwright, 2015, Theorem 2.4), specifically since $w \mapsto \|w\|$ is 1-Lipschitz, meaning
$$\left|\|w\| - \|w'\|\right| \le \|w - w'\|,$$
and since $\mathbb{E}\,\|w\| < \sqrt{d}$.

3. Note that
$$\int_{\|w\|>r} \|w\| \,\mathrm{d}G(w) \le \sum_{i=0}^{\infty} \int_{r+i<\|w\|\le r+i+1} (r+i+1)\,\mathrm{d}G(w),$$
whereas the Gaussian concentration from the preceding part grants
$$\mathbb{P}[\|w\| > r+i] \le \exp(-(r+i-\sqrt{d})^2/2) \le \exp(-(r-\sqrt{d})^2/2)\exp(-i^2/2),$$
whereby
$$\begin{aligned}\int_{\|w\|>r} \|w\| \,\mathrm{d}G(w) &\le (r+1)\int_{\|w\|>r} \mathrm{d}G(w) + \sum_{i=0}^{\infty} i \int_{\|w\|>r+i} \mathrm{d}G(w) \\ &\le (r+1)\exp(-(r-\sqrt{d})^2/2) + \sum_{i=0}^{\infty} i\exp(-(r-\sqrt{d})^2/2)\exp(-i^2/2) \\ &\le \exp(-(r-\sqrt{d})^2/2)\left[r+1+\sum_{i=0}^{\infty} i\exp(-i^2/2)\right] \\ &\le \exp(-(r-\sqrt{d})^2/2)\,[r+2]\,.\end{aligned}$$
The final inequality follows by applying Lemma A.2 with $(a,b,c,x) = (3,\sqrt{d},2,r)$

4. Proceeding similarly,
$$\begin{aligned}\int_{\|w\|>r} \|w\| \,\mathrm{d}G(w) &\le \sum_{i=0}^{\infty} \int_{r+i<\|w\|\le r+i+1} (r+i+1)^2 \,\mathrm{d}G(w), \\ &\le 2(r+1)^2 \int_{\|w\|>r} \mathrm{d}G(w) + 2\sum_{i=0}^{\infty} i^2 \int_{\|w\|>r+i} \mathrm{d}G(w) \\ &\le 2(r+1)^2\exp(-(r-\sqrt{d})^2/2) + 2\sum_{i=0}^{\infty} i^2\exp(-(r-\sqrt{d})^2/2)\exp(-i^2/2) \\ &\le 2\exp(-(r-\sqrt{d})^2/2)\left[(r+1)^2 + \sum_{i=0}^{\infty} i^2\exp(-i^2/2)\right] \\ &\le 2\exp(-(r-\sqrt{d})^2/2)\,[r+2]^2\,.\end{aligned}$$

The final inequality follows by applying Lemma A.2 with $(a, b, c, x) = (7, \sqrt{d}, 2, r)$.

□

Fourier transforms. The convention for the Fourier transform used here is

$$\hat{f}(w) = \int f(x) \exp(2\pi i w^{\mathsf{T}} x) \, \mathrm{d}x;$$

see for instance (Folland, 1999, Section 8.8) for a discussion of other conventions, and the resulting tradeoffs. Note also the polar decomposition notation $\hat{f}(w) = |\hat{f}(w)| \exp(2\pi i \theta_f(w))$ with $|\theta_f(w)| \leq 1$. The following lemma collects a few properties used throughout.

Lemma A.3. *1.* $|\hat{f}| \leq \|f\|_{L_1}$.

2. $\widehat{f * g} = \hat{f}\hat{g}$ *and* $|\widehat{f * g}| \leq \|f\|_{L_1} |\hat{g}|$.

3. Let $\alpha > 0$ *be given and define* $\phi := (2\pi\alpha)^{-1}$. *Then* $|\hat{G}_\alpha| = \hat{G}_\alpha$ *(meaning* $\hat{G}_\alpha$ *has no radial component, thus* $\theta_{G_\alpha}(w) = 0$*), and*

$$\hat{G}_\alpha(w) = (2\pi\alpha^2)^{-d/2} G_\phi(w) = (2\pi\phi^2)^{d/2} G_\phi(w) = (2\pi)^{d/2} G(w/\phi).$$

Proof. 1. Directly,

$$|\hat{h}(w)| \leq \int |h(x)| \cdot |\exp(2\pi i w^{\mathsf{T}} x)| \, \mathrm{d}x \leq \|h\|_{L_1},$$

2. The first equality is standard (Folland, 1999, Theorem 8.22c), and the inequality combines it with the preceding bound.

3. The form of $\hat{G}_\alpha$ and the first displayed inequality are standard (Folland, 1999, Proposition 8.24). The second and third inequalities use the choice of ϕ and the form of G_ϕ.

□

ReLU representation. Lastly, the exact ReLU representation constructions (e.g., Lemma 3.1) will use the following folklore lemma to write a univariate twice continuously differentiable function as an infinite width ReLU network.

Lemma A.4. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be given with* $f(0) = f'(0) = 0$ *and* f'' *continuous. For any* $z \geq 0$,

$$f(z) = \int_0^\infty \sigma(z - b) f''(b) \, \mathrm{d}b.$$

Proof. Using integration by parts,

$$\begin{aligned}
\int_0^\infty \sigma(z-b) f''(b) \, \mathrm{d}b &= \int_0^z (z-b) f''(b) \, \mathrm{d}b \\
&= z \int_0^z f''(b) \, \mathrm{d}b - \int_0^z b f''(b) \, \mathrm{d}b \\
&= z f'(b)|_0^z - \left(b f'(b)|_0^z - \int_0^z f'(b) \, \mathrm{d}b \right) \\
&= z f'(z) - z f'(0) - \big(z f'(z) - 0 - f(z) + f(0) \big) \\
&= f(z).
\end{aligned}$$

□

B SAMPLING TOOLS: MAUREY'S LEMMA AND CO-VC DIMENSION

This section collects various sampling tools used as a basis for Section 2. First is a proof of Lemma 2.5, which here is combined with an application of McDiarmid's inequality to give a high probability guarantee.

Proof of Lemma 2.5. Following the usual Maurey scheme (Pisier, 1980),

$$\begin{aligned}
\mathop{\mathbb{E}}_{(v_j)_{j=1}^m} \left\| f - \frac{1}{m}\sum_j g_j \right\|_{L_2(P)}^2 &= \frac{1}{m^2} \mathop{\mathbb{E}}_{(v_j)_{j=1}^m} \left\| \sum_j \left(f - g_j\right) \right\|_{L_2(P)}^2 \\
&= \frac{1}{m^2} \mathop{\mathbb{E}}_{(v_j)_{j=1}^m} \sum_j \left\| f - g_j \right\|_{L_2(P)}^2 \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{v_1} \left\| f - g_1 \right\|_{L_2(P)}^2 \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{v_1} \left(\left\| g_1 \right\|_{L_2(P)}^2 - \left\| f \right\|_{L_2(P)}^2 \right) \\
&\leq \frac{1}{m} \mathbb{E}_{v_1} \left\| g_1 \right\|_{L_2(P)}^2 \\
&\leq \frac{1}{m} \sup_{v \in \mathcal{V}} \left\| g(\cdot; v) \right\|_{L_2(P)}^2 .
\end{aligned}$$

The high probability bound will follow from McDiarmid's inequality. To establish the bounded differences property, define

$$F(V) := F((v_1, \ldots, v_m)) = \left\| f - \frac{1}{m}\sum_j g(\cdot; v_j) \right\|_{L_2(P)},$$

and note from the general metric space inequality $\left| \|p\| - \|q\| \right| \leq \|p - q\|$ that for any $V = (v_1, \ldots, v_m)$ and $V' = (v_1', \ldots, v_m')$ differing only on a single $v_k \neq v_k'$,

$$\begin{aligned}
\left\| F(V) - F(V') \right\| &\leq \left\| \frac{1}{m}\sum_j g(\cdot; v_j) - \frac{1}{m}\sum_j g(\cdot; v_j') \right\|_{L_2(P)} \\
&= \frac{1}{m} \left\| g(\cdot; v_k) - g(\cdot; v_k') \right\|_{L_2(P)} \\
&\leq \frac{1}{m} \left(\left\| g(\cdot; v_k) \right\|_{L_2(P)} + \left\| g(\cdot; v_k') \right\|_{L_2(P)} \right) \\
&\leq \frac{2}{m} \sup_{v \in \mathcal{V}} \left\| g(\cdot; v) \right\|_{L_2(P)} .
\end{aligned}$$

Thus, with probability at least $1 - \eta$, McDiarmid's inequality grants,

$$F(V) \leq \mathbb{E}_V F(V) + \sup_{v \in \mathcal{V}} \|g(\cdot; v)\|_{L_2(P)} \sqrt{\frac{2\ln(1/\eta)}{m}},$$

and the statement follows by Jensen's inequality, specifically $\mathbb{E}_V F(V) \leq \sqrt{\mathbb{E}_V F(V)^2}$. □

Maurey's lemma also applies to sampling from signed densities in the sense of Definition 4.4.

Lemma B.1. *Let $f(x) = \int p(\tilde{w}) g(\langle \tilde{w}, \tilde{x} \rangle) \, \mathrm{d}\tilde{w}$ be given with $\|p\|_{L_1} < \infty$ and p is supported on a ball of radius B, and let $((s_j, \tilde{w}_j))_{j=1}^m$ be sampled from p as in Definition 4.4, and define $g_j(x) := g(\langle \tilde{w}, \tilde{x} \rangle)$. With probability at least $1 - \eta$,*

$$\left\| f - \frac{\|p\|_{L_1}}{m} \sum_{j=1}^m s_j g_j \right\|_{L^2(P)} \leq \sup_{\|\tilde{w}\| \leq B} \|g(\langle \tilde{w}, \cdot \rangle)\|_{L_2(P)} \|p\|_{L_1} \left[\frac{1 + \sqrt{2\ln(1/\eta)}}{\sqrt{m}} \right].$$

Proof. Since

$$\int p(\tilde{w})g(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} = \|p\|_{L_1}\int \mathrm{sgn}(p)\frac{|p(\tilde{w})|}{\|p\|_{L_1}}g(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} = \|p\|_{L_1}\mathop{\mathbb{E}}_{s,\tilde{w}} sg(\langle \tilde{w}, \tilde{x}\rangle),$$

the sampling procedure indeed provides an unbiased estimate of the integral, and thus by Maurey's Lemma (cf. Lemma 2.5), with probability at least $1-\eta$,

$$\begin{aligned}\left\| f - \frac{\|p\|_{L_1}}{m}\sum_{j=1}^m s_j g_j\right\|_{L^2(P)} &= \|p\|_{L_1}\left\|\mathbb{E}_{s,\tilde{w}} sg(\langle \tilde{w}, \cdot\rangle) - \frac{1}{m}\sum_{j=1}^m s_j g_j\right\|_{L^2(P)} \\ &\le \|p\|_{L_1}\sup_{\substack{s\in\{\pm 1\}\\ \|\tilde{w}\|\le B}}\|sg(\langle \tilde{w}, \cdot\rangle)\|_{L_2(P)}\left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}}\right] \\ &= \|p\|_{L_1}\sup_{\|\tilde{w}\|\le B}\|g(\langle \tilde{w}, \cdot\rangle)\|_{L_2(P)}\left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}}\right].\end{aligned}$$

□

Lastly, here is a uniform norm analog of the preceding $L_2(P)$ signed density sampling bound. Interestingly, the bound only gives a $\sqrt{d}$ degradation with σ', and no degradation for σ. The method of proof is to use uniform convergence, but with data and parameters switched; consequently, this has been called "co-VC dimension" (Gurvits & Koiran, 1995; Sun et al., 2018). The proof is somewhat more complicated than the proof of the Maurey lemma, and in particular needs to be a bit more attentive to the fine-grained structure of the functions being sampled.

Lemma B.2. *Let density $p:\mathbb{R}^{d+1}\to\mathbb{R}$ with $\|p\|_{L_1}<\infty$, and let $((s_j,w_j))_{j=1}^m$ be a sample from p in the sense of Definition 4.4.*

1. *With probability at least $1-2\eta$,*

$$\sup_{\|x\|\le 1}\left|\int p(\tilde{w})\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} - \frac{\|p\|_{L_1}}{m}\sum_j s_j\sigma'(\langle \tilde{w}_j, \tilde{x}\rangle)\right| \le \frac{\|p\|_{L_1}}{\sqrt{m}}\left[\sqrt{8(d+1)\ln(m+1)}+\sqrt{\ln(1/\eta)}\right].$$

2. *Suppose p is supported on the set $\mathcal{W} := \{\tilde{w}\in\mathbb{R}^{d+1} : \|w\|\le r, |b|\le\|w\|\}$. With probability at least $1-2\eta$,*

$$\sup_{\|x\|\le 1}\left|\int p(\tilde{w})\sigma(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} - \frac{\|p\|_{L_1}}{m}\sum_j s_j\sigma(\langle \tilde{w}_j, \tilde{x}\rangle)\right| \le \frac{4r\|p\|_{L_1}}{\sqrt{m}}\left[1+\sqrt{\ln(1/\eta)}\right].$$

Proof of Lemma B.2. In both cases, letting g denote either of σ' or σ,

$$\begin{aligned}&\sup_{\|x\|\le 1}\left|\int p(\tilde{w})g(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} - \frac{\|p\|_{L_1}}{m}\sum_j s_j g(\langle \tilde{w}_j, \tilde{x}\rangle)\right| \\ &\qquad = \|p\|_{L_1}\sup_{\|x\|\le 1}\left|\mathop{\mathbb{E}}_{p} sg(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} - \frac{1}{m}\sum_j s_j g(\langle \tilde{w}_j, \tilde{x}\rangle)\right|,\end{aligned}$$

and at this point it is a classical uniform deviations problem, but with the role of parameter and data swapped, an approach which has been used before (sometimes under the heading "co-VC dimension" (Gurvits & Koiran, 1995; Sun et al., 2018)). Continuing, with probability at least $1-2\eta$, standard Rademacher complexity (Shalev-Shwartz & Ben-David, 2014) grants

$$\begin{aligned}\sup_{\|x\|\le 1}\left|\mathop{\mathbb{E}}_{p} sg(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} - \frac{1}{m}\sum_j s_j g(\langle \tilde{w}_j, \tilde{x}\rangle)\right| &\le 2\mathrm{Rad}\left(\left\{(s_j g(\langle \tilde{w}_j, \tilde{x}\rangle))_{j=1}^m : \|x\|\le 1\right\}\right) \\ &\quad + 3\sup_{\substack{\tilde{w}\in\mathcal{W}\\ \|x\|\le 1}}|g(\langle \tilde{w}, \tilde{x}\rangle)|\sqrt{\frac{\ln(1/\eta)}{2m}}.\end{aligned}$$

where $\mathcal{W}$ is a constraint set on $\tilde{w}$ (when $g = \sigma'$, it is $\mathbb{R}^{d+1}$, whereas with $g = \sigma$ it is $|b| \le \|w\| \le r$). To simplify further, note that a Rademacher random vector $(\epsilon_1, \ldots, \epsilon_m)$ is distributionally equivalent to $(s_1\epsilon_1, \ldots, s_m\epsilon_m)$ for any fixed vector of signs $(s_1, \ldots, s_m)$, and therefore

$$\begin{aligned}
\mathrm{Rad}\left(\left\{(s_j g(\langle \tilde{w}_j, \tilde{x}\rangle))_{j=1}^m : \|x\| \le 1\right\}\right) &= \frac{1}{n}\mathbb{E}_\epsilon \sup_{\|x\|\le 1} \sum_j s_j \epsilon_j g(\langle \tilde{w}, \tilde{x}\rangle) \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{\|x\|\le 1} \sum_j \epsilon_j g(\langle \tilde{w}, \tilde{x}\rangle) \\
&= \mathrm{Rad}\left(\left\{(g(\langle \tilde{w}_j, \tilde{x}\rangle))_{j=1}^m : \|x\| \le 1\right\}\right).
\end{aligned}$$

Combining these steps, with probability at least $1 - 2\eta$,

$$\begin{aligned}
&\sup_{\|x\|\le 1}\left|\int p(\tilde{w}) g(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}\tilde{w} - \frac{\|p\|_{L_1}}{m}\sum_j s_j g(\langle \tilde{w}_j, \tilde{x}\rangle)\right| \\
&\quad \le \|p\|_{L_1}\left[2\mathrm{Rad}\left(\left\{(g(\langle \tilde{w}_j, \tilde{x}\rangle))_{j=1}^m : \|x\| \le 1\right\}\right) + 3\sup_{\substack{\tilde{w}\in\mathcal{W}\\ \|x\|\le 1}} |g(\langle \tilde{w}, \tilde{x}\rangle)|\sqrt{\frac{\ln(1/\eta)}{2m}}\right].
\end{aligned}$$

The proof now splits into two cases $g \in \{\sigma', \sigma\}$, bounding the remaining terms.

1. Since the range of σ' is $\{0, 1\}$,
$$\sup_{\substack{\tilde{w}\in\mathcal{W}\\ \|x\|\le 1}} |\sigma'(\langle \tilde{w}, \tilde{x}\rangle)| \le 1,$$
and the Rademacher complexity is the VC dimension of linear predictors, thus
$$\mathrm{Rad}\left(\left\{(\sigma'(\langle \tilde{w}_j, \tilde{x}\rangle))_{j=1}^m : \|x\| \le 1\right\}\right) \le \sqrt{\frac{2(d+1)\ln(m+1)}{m}}.$$

2. In the case $g = \sigma$, since $\mathcal{W} := \{\tilde{w} \in \mathbb{R}^{d+1} : \|w\| \le r, |b| \le \|w\|\}$,
$$\sup_{\substack{\tilde{w}\in\mathcal{W}\\ \|x\|\le 1}} |\sigma(\langle \tilde{w}, \tilde{x}\rangle)| \le \sup_{\substack{\|w\|\le r\\ |b|\le\|w\|\\ \|x\|\le 1}} |w^\intercal x + b| \le 2r.$$
Moreover, the Rademacher complexity is a standard combination of the Lipschitz composition rule and linear prediction rules (Shalev-Shwartz & Ben-David, 2014), and thus
$$\mathrm{Rad}\left(\left\{(\sigma'(\langle \tilde{w}_j, \tilde{x}\rangle))_{j=1}^m : \|x\| \le 1\right\}\right) \le \mathrm{Rad}\left(\left\{(\langle \tilde{w}_j, \tilde{x}\rangle)_{j=1}^m : \|x\| \le 1\right\}\right) \le \frac{4r}{\sqrt{m}}.$$

□

C Deferred proofs from Section 2

For convenience throughout this appendix, define

$$B := \sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\| \qquad \text{and} \qquad B_\epsilon := \sup_{\tilde{w}, s} \|\mathcal{T}_\epsilon(\tilde{w}, s)\|_2 \le \frac{B}{\epsilon\sqrt{m}} + R.$$

The first step is to prove eq. (2.2), restated here as follows.

Lemma C.1. *With probability at least $1 - \eta$ over $(\widetilde{W}, S)$,*

$$\left\|\sum_{j=1}^m \left\langle \tau_j, \phi_j(\cdot)\right\rangle - \mathbb{E}_{\tilde{w}}\left\langle \mathcal{T}(\tilde{w}), \Phi(\cdot; \tilde{w})\right\rangle\right\|_{L^2(P)} \le \epsilon B_\epsilon\left[\sqrt{2} + 2\sqrt{\ln(1/\eta)}\right].$$

Proof. The proof proceeds by applying Maurey sampling (cf. Lemma 2.5) to the functions $g_j(x) := m\left\langle \tau_j, \phi_j(x)\right\rangle$, noting by Lemma 2.4 that

$$f(x) := \mathop{\mathbb{E}}_{\widetilde{W},S} \frac{1}{m}\sum_j g_j(x) = \mathop{\mathbb{E}}_{\widetilde{W},S} \sum_j \left\langle \tau_j, \phi_j(x)\right\rangle = \mathop{\mathbb{E}}_{\widetilde{W},S} \left\langle \mathcal{T}_\epsilon(\widetilde{W}, S), \Phi_\epsilon(x; \widetilde{W}, S)\right\rangle = \mathop{\mathbb{E}}_{\tilde{w}} \left\langle \mathcal{T}(\tilde{w}), \Phi(x;\tilde{w})\right\rangle.$$

Applying Lemma 2.5, with probability at least $1-\eta$,

$$\left\| f - \frac{1}{m}\sum_{j=1}^m g_j \right\|_{L^2(P)} \le \frac{\sup_{\tilde{w},s} m \| \left\langle \mathcal{T}_\epsilon(\tilde{w}, s), \Phi_\epsilon(\cdot; \tilde{w}, s)\right\rangle \|_{L_2(P)}}{\sqrt{m}} \left[1 + \sqrt{2\ln(1/\eta)}\right],$$

where

$$\sup_{\tilde{w},s} \| \left\langle \mathcal{T}_\epsilon(\tilde{w}, s), \Phi_\epsilon(\cdot; \tilde{w}, s)\right\rangle \|_{L_2(P)}^2 \le \sup_{\tilde{w},s} \mathbb{E}_x \left\|\mathcal{T}_\epsilon(\tilde{w})\right\|_2^2 \left\|\Phi_\epsilon(x;\tilde{w},s)\right\|_2^2 \le \frac{2\epsilon^2 B_\epsilon^2}{m}.$$

□

Next, the restatement of eq. (2.3) is as follows.

Lemma C.2. *With probability at least* $1-\eta$*, If* $R \ge \sqrt{d} + 2\sqrt{\ln\left(\frac{\epsilon\sqrt{m\pi}}{B\sqrt{2}}\right)}$*, then with probability at least* $1-\eta$*,*

$$\left\| \sum_j \left\langle \tau_j, \phi_j(\cdot)\right\rangle - \sum_j \frac{s_j \epsilon}{\sqrt{m}} \sigma(\left\langle \tau_j, \tilde{x}\right\rangle) \right\|_{L_2(P)} \le \frac{B_\epsilon B}{m\sqrt{\pi}} + \epsilon B_\epsilon \left[\sqrt{2} + 2\sqrt{\ln(1/\eta)}\right].$$

Recall that the proof of Lemma C.2, as discussed in the body, must calculate the fraction of activations which change, which was collected into Lemma 2.6.

Proof of Lemma 2.6. Consider an idealized $\mathcal{T}_\epsilon'$ which does not truncate, whereby

$$\left| |\left\langle \mathcal{T}_\epsilon'(\tilde{w}), \tilde{x}\right\rangle| - |\left\langle \tilde{w}, \tilde{x}\right\rangle| \right| \le \left| \frac{\left\langle \mathcal{T}(\tilde{w}), \tilde{x}\right\rangle}{\epsilon\sqrt{m}} + \left\langle \tilde{w}, \tilde{x}\right\rangle - \left\langle \tilde{w}, \tilde{x}\right\rangle \right| \le \frac{B\|\tilde{x}\|}{\epsilon\sqrt{m}}.$$

The event $\left[\mathrm{sgn}(\left\langle \tilde{w}, \tilde{x}\right\rangle) \ne \mathrm{sgn}(\left\langle \mathcal{T}_\epsilon'(\tilde{w}), \tilde{x}\right\rangle)\right]$ implies the event $\left[|\left\langle \tilde{w}, \tilde{x}\right\rangle| \le {}^{B\|\tilde{x}\|}/_{\epsilon\sqrt{m}}\right]$, and thus, additionally using rotational invariance of the Gaussian,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\tilde{w}} \left| \sigma'(\left\langle \tilde{w}, \tilde{x}\right\rangle) - \sigma'(\left\langle \mathcal{T}_\epsilon'(\tilde{w}), \tilde{x}\right\rangle) \right| &= \mathop{\mathbb{P}}_{\tilde{w}} \left[\mathrm{sgn}(\left\langle \tilde{w}, \tilde{x}\right\rangle) \ne \mathrm{sgn}(\left\langle \mathcal{T}_\epsilon'(\tilde{w}), \tilde{x}\right\rangle)\right] \\
&\le \mathop{\mathbb{P}}_{\tilde{w}} \left[|\left\langle \tilde{w}, \tilde{x}\right\rangle| \le {}^{B\|\tilde{x}\|}/_{\epsilon\sqrt{m}}\right] \\
&= \mathop{\mathbb{P}}_{\tilde{w}} \left[|w_1| \cdot \|\tilde{x}\| \le {}^{B\|\tilde{x}\|}/_{\epsilon\sqrt{m}}\right] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-B/(\epsilon\sqrt{m})}^{B/(\epsilon\sqrt{m})} e^{-z^2/2}\,\mathrm{d}z \\
&\le \frac{B}{\epsilon}\sqrt{\frac{2}{m\pi}}.
\end{aligned}$$

Returning to the general case with truncation, by Lemma A.1, using the assumed lower bound on R,

$$\mathbb{P}[\|\tilde{w}\| > R] \le \exp(-(R-\sqrt{d})^2/2) \le \frac{B}{\epsilon}\sqrt{\frac{2}{m\pi}},$$

which gives the final bound via triangle inequality. □

With Lemma 2.6 in hand, the proof of Lemma C.2 is now an application of Maurey's lemma, with an invocation of positive homogeneity to massage terms.

Proof of Lemma C.2. The approach is once again to apply Maurey sampling (cf Lemma 2.5). To this end, define

$$g(x;\tilde{w},s) := m\left(\left\langle \mathcal{T}_\epsilon(\tilde{w},s), \Phi_\epsilon(x;\tilde{w},s)\right\rangle - \frac{s\epsilon}{\sqrt{m}}\sigma\left(\left\langle \mathcal{T}_\epsilon(\tilde{w},s),\tilde{x}\right\rangle\right)\right) \quad \text{and} \quad f(x) = \mathop{\mathbb{E}}_{\tilde{w},s} g(x;\tilde{w},s),$$

as well as $g_j(x) := g(x;\tilde{w}_j,s_j)$. Using this notation, the goal of this proof is to upper bound

$$\left\|\sum_j \left\langle \tau_j, \phi_j(\cdot)\right\rangle - \sum_j \frac{s_j\epsilon}{\sqrt{m}}\sigma(\left\langle \tau_j,\cdot\right\rangle)\right\|_{L_2(P)} = \left\|\frac{1}{m}\sum_j g_j\right\|_{L_2(P)}.$$

By Lemma 2.5, with probability at least $1-\eta$,

$$\begin{aligned}\left\|\frac{1}{m}\sum_j g_j\right\|_{L_2(P)} &\le \|f\|_{L_2(P)} + \left\|f - \frac{1}{m}\sum_j g_j\right\|_{L_2(P)} \\ &\le \|f\|_{L_2(P)} + \sup_{\tilde{w},s}\|g(\cdot;\tilde{w},s)\|_{L_2(P)}\left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}}\right].\end{aligned}$$

To control these terms, fixing any $(\tilde{w},s)$, it holds by positive homogeneity of σ that

$$\left\|g(x;\tilde{w},s)\right\|_{L_2(P)}^2 = m^2 \mathop{\mathbb{E}}_x\left\langle \mathcal{T}_\epsilon(\tilde{w},s), \Phi_\epsilon(x;\tilde{w},s) - \frac{s\epsilon}{\sqrt{m}}\Phi(x;\mathcal{T}_\epsilon(\tilde{w},s))\right\rangle^2 \le m^2 B_\epsilon^2 \mathop{\mathbb{E}}_x \frac{\epsilon^2\|x\|^2}{m} \le 2m\epsilon^2 B_\epsilon^2.$$

On the other hand, by Lemma 2.6, for any $\|x\|\le 1$,

$$\begin{aligned}|f(x)| &\le \mathbb{E}_{\tilde{w},s}\left|\left\langle \mathcal{T}_\epsilon(\tilde{w},s), \Phi_\epsilon(x;\tilde{w},s) - \frac{s\epsilon}{\sqrt{m}}\Phi(x;\mathcal{T}_\epsilon(\tilde{w},s))\right\rangle\right| \\ &\le B_\epsilon \mathbb{E}_{\tilde{w},s}\frac{\epsilon\|\tilde{x}\|\left|\sigma'(\mathcal{T}_\epsilon(\tilde{w})^\intercal\tilde{x}) - \sigma'(\tilde{w}^\intercal\tilde{x})\right|}{\sqrt{m}} \\ &\le \frac{\epsilon B_\epsilon}{\sqrt{m}}\left(\frac{B}{\epsilon\sqrt{m\pi}}\right),\end{aligned}$$

which also upper bounds $\|f\|_{L_2(P)}$. □

The proof of Theorem 2.1 now follows by combining Lemmas C.1 and C.2.

Proof of Theorem 2.1. By Lemma A.1 and a union bound on $(\tilde{w}_1,\ldots,\tilde{w}_m)$, $\max_j\|\tilde{w}_j\|\le R$, thus

$$\max_j \|\mathcal{T}_\epsilon(\tilde{w}_j) - \tilde{w}_j\| \le \max_j \frac{\|\mathcal{T}(\tilde{w}_j)\|}{\epsilon\sqrt{m}} \le \frac{B}{\epsilon\sqrt{m}}.$$

Moreover, $R \ge \sqrt{d} + 2\sqrt{\ln\left(\frac{\epsilon\sqrt{m\pi}}{B\sqrt{2}}\right)}$, and thus the two other bounds are from Lemmas C.1 and C.2. □

D Deferred proofs from Section 3

The first core lemma shows how to write a target function f as an infinite-width network via its Fourier transform.

Proof of Lemma 3.1. The first steps are the same for σ and σ', and indeed match the initial steps of (Barron, 1993), namely

$$\begin{aligned}f(x) - f(0) &= \text{Re}\int \exp(2\pi i x^\intercal w)\hat{f}(w)\,\mathrm{d}w \\ &= \text{Re}\int \exp(2\pi i x^\intercal w + 2\pi i\theta_f(w))|\hat{f}(w)|\,\mathrm{d}w \\ &= \int \cos\left(2\pi(x^\intercal w + \theta_f(w))\right)|\hat{f}(w)|\,\mathrm{d}w.\end{aligned}$$

For convenience, define $h(z) := \cos(2\pi z)$, whereby

$$f(x) - f(0) = \int h(x^\intercal w + \theta_f(w))|\hat{f}(w)| \,\mathrm{d}w, \tag{D.1}$$

and the proofs not differ for both activations and from (Barron, 1993).

1. Consider first σ'. Since $\|x\| \leq 1$, by Cauchy-Schwarz it suffices to approximate h along the interval $[-\|w\| + \theta_f(w), \|w\| + \theta_f(w)]$. By the fundamental theorem of calculus,

$$\begin{aligned}
&h(\langle w, x\rangle + \theta_f(w)) - h\left(-\|w\| + \theta_f(w)\right) \\
&\quad = \int_{-\|w\|+\theta_f(w)}^{\langle w,x\rangle+\theta_f(w)} h'(b) \,\mathrm{d}b \\
&\quad = \int h'(b)\mathbb{1}[\langle w, x\rangle + \theta_f(w) \geq b]\mathbb{1}[b \geq -\|w\| + \theta_f(w)] \,\mathrm{d}b \\
&\quad = -\int h'(\theta_f(w) - b)\mathbb{1}[x^\intercal w + b \geq 0]\mathbb{1}[\|w\| \geq b] \,\mathrm{d}b, \qquad\qquad b \mapsto \theta_f(w) - b \\
&\quad = -\int h'(\theta_f(w) - b)\mathbb{1}[x^\intercal w + b \geq 0]\mathbb{1}[\|w\| \geq |b|] \,\mathrm{d}b,
\end{aligned}$$

where the last step follows since $\mathbb{1}[x^\intercal w + b \geq 0]$ implies $b \geq -\|w\|$. Plugging this back in to eq. (D.1) and still using $h(z) = \cos(2\pi z)$,

$$\begin{aligned}
f(x) - f(0) &= \int |\hat{f}(w)| \cos\left(2\pi(x^\intercal w + \theta_f(w))\right) \mathrm{d}w \\
&= \int |\hat{f}(w)| \left[h(-\|w\| + \theta_f(w)) - \int h'(\theta_f(w) - b)\mathbb{1}[x^\intercal w + b \geq 0]\mathbb{1}[\|w\| \geq |b|] \,\mathrm{d}b\right] \mathrm{d}w,
\end{aligned}$$

which after pushing more terms onto the left hand side gives

$$\begin{aligned}
&f(x) - f(0) - \int |\hat{f}(w)| h(\theta_f(w) - \|w\|) \,\mathrm{d}w \\
&\quad = -\iint |\hat{f}(w)| h'(\theta_f(w) - b)\mathbb{1}[x^\intercal w + b \geq 0]\mathbb{1}[\|w\| \geq |b|] \,\mathrm{d}b \,\mathrm{d}w \\
&\quad = 2\pi \int |\hat{f}(w)| \sin(2\pi(\theta_f(w) - b))\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\mathbb{1}[\|w\| \geq |b|] \,\mathrm{d}\tilde{w},
\end{aligned}$$

which gives $F_\infty = f$ for $\|x\| \leq 1$. To bound the error of F_r, note by the form of F_∞ for any $\|x\| \leq 1$ that

$$\begin{aligned}
\left|f(x) - F_r(x)\right| &= \left|2\pi \int_{\|w\|>r} \int |\hat{f}(w)| \sin(2\pi(\theta_f(w) - b))\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\mathbb{1}[\|w\| \geq |b|] \,\mathrm{d}b \,\mathrm{d}w\right| \\
&\leq 2\pi \int_{\|w\|>r} \int_{|b|\leq\|w\|} |\hat{f}(w)| |\sin(2\pi(\theta_f(w) - b))| \sigma'(\langle \tilde{w}, \tilde{x}\rangle) \,\mathrm{d}b \,\mathrm{d}w \\
&\leq 2\pi \int_{\|w\|>r} |\hat{f}(w)| \int_{|b|\leq\|w\|} \mathrm{d}b \,\mathrm{d}w \\
&= 4\pi \int_{\|w\|>r} \|w\| \cdot |\hat{f}(w)| \,\mathrm{d}w.
\end{aligned}$$

2. Now consider σ. Rather than using FTC as above, this proof replaces h with ReLUs via Lemma A.4, which requires a function which is both zero and flat at 0. To this end, define

$$H(b) = h(b + q) - \left(h(q) + bh'(q)\right) \quad \text{with } q = -\|w\| + \theta_f(w),$$

whereby $H(0) = 0 = H'(0)$. Invoking Lemma A.4 on H gives, for any $z := w^\intercal x + \theta_f(w) \geq q$,

$$\begin{aligned}
h(z) - \big(h(q) + (z-q)h'(q)\big) &= H(z-q) \\
&= \int H''(b)\sigma(z-q-b)\mathbb{1}[b \geq 0]\,\mathrm{d}b \\
&= \int H''(b)\sigma(w^\intercal x + \theta_f(w) + \|w\| - \theta_f(w) - b)\mathbb{1}[b \geq 0]\,\mathrm{d}b \\
&= -\int H''(\|w\| - b)\sigma(w^\intercal x + b)\mathbb{1}[\|w\| \geq b]\,\mathrm{d}b \qquad b \mapsto \|w\| - b \\
&= -\int H''(\|w\| - b)\sigma(w^\intercal x + b)\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b,
\end{aligned}$$

the final equality since $-b > \|w\|$ implies $w^\intercal x + b \leq \|w\| + b < 0$, thus $\sigma(\tilde{w}^\intercal \tilde{x}) = 0$ and this case has no effect. Plugging this back into eq. (D.1),

$$\begin{aligned}
f(x) - f(0) &= \int |\hat{f}(w)| h(z)\,\mathrm{d}w \\
&= \int |\hat{f}(w)| \left[h(q) + (z-q)h'(q) - \int H''(\|w\| - b)\sigma(\tilde{w}^\intercal \tilde{x})\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\right]\mathrm{d}w \\
&= \int |\hat{f}(w)| \left[h(q) + (w^\intercal x + \|w\|)h'(q) - \int H''(\|w\| - b)\sigma(\tilde{w}^\intercal \tilde{x})\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\right]\mathrm{d}w,
\end{aligned}$$

which gives $Q_\infty = f$ for $\|x\| \leq 1$ after expanding h and H. To bound the error of Q_r, for any $\|x\| \leq 1$

$$\begin{aligned}
\big|f(x) - Q_r(x)\big| &= \left|\int_{\|w\|>r} |\hat{f}(w)| \int H''(\|w\| - b)\sigma(\tilde{w}^\intercal \tilde{x})\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\right|\mathrm{d}w \\
&\leq \int_{\|w\|>r} |\hat{f}(w)| \int |H''(\|w\| - b)|\sigma(\tilde{w}^\intercal \tilde{x})\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\,\mathrm{d}w \\
&\leq 4\pi^2 \int_{\|w\|>r} |\hat{f}(w)| \int_{-\|w\|}^{\|w\|} \sigma(\tilde{w}^\intercal \tilde{x})\,\mathrm{d}b\,\mathrm{d}w \\
&\leq 4\pi^2 \int_{\|w\|>r} |\hat{f}(w)| \int_{-\|w\|}^{\|w\|} (\|w\| + |b|)\,\mathrm{d}b\,\mathrm{d}w \\
&\leq 12\pi^2 \int_{\|w\|>r} \|w\|^2 \cdot |\hat{f}(w)|\,\mathrm{d}w.
\end{aligned}$$

□

Next, Lemma 3.2 converts Lemma 3.1 into a (random feature) transport map by introducing the fraction $G(\tilde{w})/G(\tilde{w})$.

Proof of Lemma 3.2. Starting from the construction in Lemma 3.1, again using $h(z) = \cos(2\pi z)$ for convenience, and manually introducing a factor $G(\tilde{w})$,

$$\begin{aligned}
f(x) - f(0) - \int |\hat{f}(w)| h(\theta_f(w) - \|w\|)\,\mathrm{d}w \\
&= -\iint |\hat{f}(w)| h'(\theta_f(w) - b)\mathbb{1}[x^\intercal w + b \geq 0]\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\,\mathrm{d}w \\
&= -\int \frac{|\hat{f}(w)|}{G(\tilde{w})} h'(\theta_f(w) - b)\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}G(\tilde{w}).
\end{aligned}$$

To construct $\mathcal{T}_\infty$, rotational invariance of the Gaussian gives $\mathbb{E}\mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] = 1/2$, thus

$$f(0) + \int |\hat{f}(w)| h(\theta_f(w) - \|w\|)\,\mathrm{d}w = \int 2\left[f(0) + \int |\hat{f}(v)| h(\theta_f(v) - \|v\|)\,\mathrm{d}v\right]\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\,\mathrm{d}G(\tilde{w}),$$

and transport mapping is $\mathcal{T}_\infty(w,b) = (0, p_\infty(\tilde{w})) \in \mathbb{R}^d \times \mathbb{R}$ with

$$p_\infty(\tilde{w}) = 2\left[f(0) + \int |\hat{f}(v)| h(\theta_f(v) - \|v\|)\,\mathrm{d}v\right] - \frac{|\hat{f}(w)|}{G(\tilde{w})} h'(\theta_f(w) - b)\mathbb{1}[\|w\| \geq |b|],$$

By construction, $\mathbb{E}_{\tilde{w}}\left\langle \mathcal{T}_r(\tilde{w}), \Phi(x;\tilde{w})\right\rangle = F_r(x)$, and therefore Lemma 3.1 grants for all $\|x\| \leq 1$

$$f(x) = \mathbb{E}_{\tilde{w}}\left\langle \mathcal{T}_\infty(\tilde{w}), \Phi(x;\tilde{w})\right\rangle \quad \text{and} \quad \left|f(x) - \mathbb{E}\left\langle \mathcal{T}_r(\tilde{w}), \Phi(x;\tilde{w})\right\rangle\right| \leq 4\pi \int_{\|w\|>r} |\hat{f}(w)| \cdot \|w\|\,\mathrm{d}w.$$

□

With more care (in particular, a crucial change of variable), a much better bound is possible for convolutions with Gaussians.

Proof of Lemma 3.3. By Lemma A.3, setting $\phi := (2\pi\sigma)^{-1}$,

$$|\hat{f}_\alpha(w)| = |\hat{f}(w)|\hat{G}_\alpha(w) = (2\pi)^{d/2}|\hat{f}(w)|G(w/\phi).$$

Plugging this into Lemma 3.1 and again defining $h(z) := \cos(2\pi z)$ for convenience, but unlike Lemma 3.2 performing a change of variable to directly introduce $G(\tilde{w})$, and then manually introducing $G(b)$,

$$\begin{aligned}
&f_\alpha(x) - f_\alpha(0) - \int |\hat{f}_\alpha(w)| h(\theta_{f_\alpha}(w) - \|w\|)\,\mathrm{d}w \\
&\quad = -\iint |\hat{f}_\alpha(w)| h'(\theta_{f_\alpha}(w) - b)\sigma'(\tilde{w}^\intercal \tilde{x})\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\,\mathrm{d}w \\
&\quad = -(2\pi)^{d/2}\iint |\hat{f}(w)| G(w/\phi) h'(\theta_{f_\alpha}(w) - b)\sigma'(\tilde{w}^\intercal \tilde{x})\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}b\,\mathrm{d}w \\
&\quad = -(2\pi\phi^2)^{d/2}\phi \int |\hat{f}(\phi w)| G(w) h'(\theta_{f_\alpha}(\phi w) - b)\sigma'(\phi\langle \tilde{w}, \tilde{x}\rangle)\mathbb{1}[\phi\|w\| \geq \phi|b|]\,\mathrm{d}b\,\mathrm{d}w \\
&\quad = -(2\pi\phi^2)^{(d+1)/2} \int |\hat{f}(\phi w)| e^{b^2/2} h'(\theta_{f_\alpha}(\phi w) - b)\sigma'(\langle \tilde{w}, \tilde{x}\rangle)\mathbb{1}[\|w\| \geq |b|]\,\mathrm{d}G(\tilde{w}).
\end{aligned}$$

As in Lemma 3.2, the transport is constructed by using $\mathbb{E}\sigma'(\langle \tilde{w}, \tilde{x}\rangle) = 1/2$ to model constants: $\mathcal{T}_r(w,b) = (0,\ldots,0,p_r(\tilde{w}))$, where

$$\begin{aligned}
p_r(\tilde{w}) := 2&\left[f_\alpha(0) + \int |\hat{f}_\alpha(v)| h(\theta_{f_\alpha}(v) - \|v\|)\,\mathrm{d}v\right] \\
&- (2\pi\phi^2)^{(d+1)/2}|\hat{f}(\phi w)| e^{b^2/2} h'(\theta_{f_\alpha}(\phi w) - b)\mathbb{1}\left[|b| \leq \|w\| \leq r\right],
\end{aligned}$$

with $f(x) = \mathbb{E}_{\tilde{w}}\left\langle \mathcal{T}_\infty(\tilde{w}), \Phi(x;\tilde{w})\right\rangle$ for $\|x\| \leq 1$ by construction.

When $r < \infty$, by construction

$$\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\| \leq 2\big|f_\alpha(0)\big| + 2\int |\hat{f}_\alpha(v)|\,\mathrm{d}v + 2\pi(2\pi\phi^2)^{(d+1)/2} \sup_{\substack{\|w\| \leq r \\ |b| \leq \|w\|}} |\hat{f}(\phi w)| e^{b^2/2},$$

where $|\hat{f}(\phi w)| = 1$ when $f_\alpha = G_\alpha$ (meaning f itself is the Dirac at 0), and more generally Lemma A.3 grants $|\hat{f}(\phi w)| \leq \|f(\phi\cdot)\|_{L_1}$; as in the lemma statement, these cases are summarized with $|\hat{f}(\phi w)| \leq M_f$. Plugging this in and simplifying further via Lemma A.3,

$$\begin{aligned}
\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\| &\leq 2\left|\int f(x)G_\alpha(-x)\,\mathrm{d}x\right| + 2(2\pi)^{d/2}\int |\hat{f}(v)| G(v/\phi)\,\mathrm{d}v + 2\pi(2\pi\phi^2)^{(d+1)/2} M_f \sup_{|b| \leq r} e^{b^2/2} \\
&\leq 2\left[M + 2(2\pi\phi^2)^{d/2} M_f + 2\pi(2\pi\phi^2)^{(d+1)/2} M_f \sup_{|b| \leq r} e^{b^2/2}\right].
\end{aligned}$$

For the approximation estimate, for any $\|x\| \leq 1$, the preceding derivation and Lemma A.1 grant

$$\begin{aligned}
&\left| f(x) - \mathbb{E}\left\langle \mathcal{T}_r(\tilde{w}), \Phi(x;\tilde{w}) \right\rangle \right| \\
&\quad = \left| \mathbb{E}\left\langle \mathcal{T}_\infty(\tilde{w}) - \mathcal{T}_r(\tilde{w}), \Phi(x;\tilde{w}) \right\rangle \right| \\
&\quad \leq 2\pi(2\pi\phi^2)^{(d+1)/2} \int_{\|w\|>r} \int_{|b|\leq\|w\|} |\hat{f}(\phi w)| |\sin(2\pi(w^\intercal x + \theta_{f_\alpha}(w)))\sigma'(\tilde{w}^\intercal \tilde{x})| \,\mathrm{d}b \,\mathrm{d}G(w) \\
&\quad \leq 2\pi(2\pi\phi^2)^{(d+1)/2} M_f \int_{\|w\|>r} \int_{|b|\leq\|w\|} \mathrm{d}b \,\mathrm{d}G(w) \\
&\quad \leq 4\pi(2\pi\phi^2)^{(d+1)/2} M_f \int_{\|w\|>r} \|w\| \,\mathrm{d}G(w) \\
&\quad \leq 4\pi(2\pi\phi^2)^{(d+1)/2} M_f (\sqrt{d}+3) \exp\left(-(r-\sqrt{d})^2/4\right).
\end{aligned}$$

□

E DEFERRED PROOFS FROM SECTION 4

The first proof is of the approximation properties of Gaussian convolution; as stated in the body, the proof proceeds by splitting the error into two terms, one for nearby points, the other for distant points.

Proof of Lemma 4.2. Splitting the integral into two terms, for any $\|x\| \leq 1$,

$$\begin{aligned}
\left| f(x) - (f_{|\delta} * G_\alpha)(x) \right| &= \left| \int f_{|\delta}(x) G_\alpha(z) \,\mathrm{d}z - \int f_{|\delta}(z) G_\alpha(x-z) \,\mathrm{d}z \right| \\
&= \left| \int f_{|\delta}(x) G_\alpha(z) \,\mathrm{d}z - \int f_{|\delta}(x-z) G_\alpha(z) \,\mathrm{d}z \right| \\
&\leq \int \left| f_{|\delta}(x) - f_{|\delta}(x-z) \right| G_\alpha(z) \,\mathrm{d}z \\
&= \int_{\|z\|\leq\delta} \left| f_{|\delta}(x) - f_{|\delta}(x-z) \right| G_\alpha(z) \,\mathrm{d}z \\
&\quad + \int_{\|z\|>\delta} \left| f_{|\delta}(x) - f_{|\delta}(x-z) \right| G_\alpha(z) \,\mathrm{d}z.
\end{aligned}$$

Analyzing these terms separately, the definition of $\omega_f(\delta)$ gives

$$\int_{\|z\|\leq\delta} \left| f_{|\delta}(x) - f_{|\delta}(x-z) \right| G_\alpha(z) \,\mathrm{d}z \leq \int_{\|z\|\leq\delta} \omega_f(\delta) G_\alpha(z) \,\mathrm{d}z \leq \omega_f(\delta),$$

whereas Gaussian concentration (cf. Lemma A.1) gives

$$\int_{\|z\|>\delta} \left| f_{|\delta}(x) - f_{|\delta}(x-z) \right| G_\alpha(z) \,\mathrm{d}z \leq 2M\, \mathbb{P}[\|\alpha z\| > \delta] \leq 2M \exp(-(\delta/\alpha - \sqrt{d})^2/2) \leq \omega_f(\delta).$$

□

This now combines with Lemma 3.3 to prove Theorem 4.3.

Proof of Theorem 4.3. Plugging the choice of r into Lemma 3.3, for any $\|x\| \leq 1$,

$$\left| f(x) - \mathbb{E}\left\langle \mathcal{T}_r(\tilde{w}), \Phi(x;\tilde{w}) \right\rangle \right| \leq 4\pi(2\pi\phi^2)^{(d+1)/2} M_f (\sqrt{d}+3) \exp\left(-(r-\sqrt{d})^2/4\right) \leq \omega_f(\delta).$$

Moreover, plugging r into the estimate on $\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\|$ provided by Lemma 3.3 gives

$$\sup_{\tilde{w}} \|\mathcal{T}_r(\tilde{w})\| \leq 2\left[M + (2\pi\phi^2)^{d/2} M_f \left(1 + \sqrt{2\pi^3\phi^2} e^{r^2/2}\right) \right],$$

where Lemma A.3 and the choice of r give

$$M_f \leq \|f_{|\delta}\|_{L_1}, \qquad \text{and} \qquad e^{r^2/2} \leq e^d e^{(r-\sqrt{d})^2},$$

where

$$e^{(r-\sqrt{d})^2} = \left(\frac{4\pi(2\pi\phi^2)^{(d+1)/2} M_f(\sqrt{d}+3)}{\omega_f(\delta)}\right)^4 = \mathcal{O}\left(\left(\frac{M_f\sqrt{d}}{\omega_f(\delta)\alpha^{d+1}}\right)^4\right),$$

and noting moreover that $\alpha = \widetilde{\mathcal{O}}(\delta/\sqrt{d})$. □

To close this section comes the full version of Theorem 4.5, which gives explicit constructions for both threshold σ' and ReLU σ. Interestingly, in the case of σ', it is not necessary to truncate the density, as is the case everywhere else in this work.

Theorem E.1. *As in Lemma 4.2, let $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$ be given, and define*

$$M := \sup_{\|x\|\leq 1+\delta} |f(x)|, \qquad f_{|\delta}(x) := f(x)\mathbb{1}[\|x\| \leq 1+\delta], \qquad \alpha := \frac{\delta}{\sqrt{d} + \sqrt{2\ln(2M/\omega_f(\delta))}}.$$

Let G_α denote a Gaussian with the preceding variance α^2, and define $h := f_{|\delta} * G_\alpha$ with Fourier transform $\hat{h}$ satisfying radial decomposition $\hat{h}(w) = |\hat{h}(w)|\exp(2\pi i\theta_h(w)$. Lastly, let P be a probability measure supported on $\|x\| \leq 1$.

1. *Additionally define*

   $$c_1 := h(0) + \int |\hat{h}(w)|\cos\big(2\pi(\theta_h(w) - \|w\|)\big)\,\mathrm{d}w, \qquad p_1 := 2\pi|\hat{h}(w)|\sin(2\pi(\theta_h(w) - b))\mathbb{1}\left[|b| \leq \|w\|\right].$$

   Then

   $$|c_1| \leq M + \|f_{|\delta}\|_{L_1}(2\pi\alpha^2)^{d/2}, \qquad \text{and} \qquad \|p_1\|_{L_1} \leq 2\|f_{|\delta}\|_{L_1}\sqrt{\frac{2\pi d}{(2\pi\alpha^2)^{d+1}}},$$

   and with probability at least $1 - 3\eta$ over a draw of $((s_j, \tilde{w}_j))_{j=1}^m$ from p_1 (cf. Definition 4.4),

   $$\left\| f - \left[c_1 + \|p_1\|_{L_1}\sum_{j=1}^m s_j\sigma'(\langle \tilde{w}_j, \cdot\rangle)\right]\right\|_{L_2(P)} \leq 2\omega_f(\delta) + \|p_1\|_{L_1}\left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}}\right],$$

   $$\sup_{\|x\|\leq 1}\left| f(x) - \left[c_1 + \|p_1\|_{L_1}\sum_{j=1}^m s_j\sigma'(\langle \tilde{w}_j, x\rangle)\right]\right| \leq 2\omega_f(\delta) + \frac{\|p\|_{L_1}}{\sqrt{m}}\left[\sqrt{8(d+1)\ln(m+1)} + \sqrt{\ln(1/\eta)}\right].$$

2. *Additionally define*

   $$c_2 := f(0)f(0) + \int |\hat{h}(w)|\left[\cos(2\pi(\theta_h(w) - \|w\|)) - 2\pi\|w\|\sin(2\pi(\theta_h(w) - \|w\|))\right]\mathrm{d}w,$$

   $$a_2 := \int w|\hat{h}(w)|\,\mathrm{d}w,$$

   $$r_2 := \sqrt{d} + 2\sqrt{\ln\frac{24\pi^2(\sqrt{d}+7)^2\|f_{|\delta}\|_{L_1}}{\omega_f(\delta)}},$$

   $p_2(\tilde{w}) := 4\pi^2|\hat{h}(w)|\cos(2\pi(\|w\| - b))\mathbb{1}[|b| \leq \|w\| \leq r_2]$,
   and for convenience create fake (weight, bias, sign) triples
   $(w, b, s)_{m+1} := (0, |c_2|, m \cdot \mathrm{sgn}(c_2)), \quad (w, b, s)_{m+2} := (a_2, 0, +m), \quad (w, b, s)_{m+3} := (-a_2, 0, -m).$
   Then

   $$\|a_2\|_2 \leq \sqrt{d}\|f_{|\delta}\|_{L_1}\phi(2\pi\alpha^2)^{-d/2},$$

   $$\|p_2\|_{L_1} \leq 2\|f_{|\delta}\|_{L_1}\sqrt{\frac{(2\pi)^3 d}{(2\pi\alpha^2)^{d+1}}},$$

   $$|c_2| \leq M + 2\sqrt{d}\|f_{|\delta}\|_{L_1}(2\pi\alpha^2)^{-d/2},$$

and with probability at least $1-3\eta$ *over a draw of* $((s_j, \tilde{w}_j))_{j=1}^m$ *from* p_2 *(cf. Definition 4.4),*

$$\left\| f - \frac{1}{m}\sum_{j=1}^{m+3} s_j \sigma(\langle \tilde{w}_j, \cdot \rangle) \right\|_{L_2(P)} \leq 3\omega_f(\delta) + r_2 \|p\|_{L_1} \left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}} \right],$$

$$\sup_{\|x\|\leq 1} \left| f(x) - \frac{1}{m}\sum_{j=1}^{m+3} s_j \sigma(\langle \tilde{w}_j, \cdot \rangle) \right| \leq 3\omega_f(\delta) + \frac{4r_2\|p\|_{L_1}}{\sqrt{m}} \left[1 + \sqrt{\ln(1/\eta)} \right].$$

Proof. 1. By Lemma 3.1 and the choice of b_1, for any $\|x\| \leq 1$,

$$h(x) = c_1 + \int p_1(\tilde{w}) \sigma'(\langle \tilde{w}, x \rangle) \, \mathrm{d}\tilde{w},$$

thus by Lemma 4.2 and Lemma B.1, defining $h_j := \|p\|_{L_1} \sigma'(\langle \tilde{w}_j, \cdot \rangle)$ for convenience, with probability at least $1-\eta$,

$$\begin{aligned}
\left\| f - (c_1 + \sum_j h_j/m) \right\|_{L_2(P)} &\leq \|f-h\|_{L_2(P)} + \left\| h - (c_1 + \sum_j h_j/m) \right\|_{L_2(P)} \\
&\leq 2\omega_f(\delta) + \|p_1\|_{L_1} \sup_{\|\tilde{w}\|\leq r_2} \|\sigma'(\langle \tilde{w}, \cdot \rangle\|_{L_2(P)} \left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}} \right] \\
&\leq 2\omega_f(\delta) + \|p_1\|_{L_1} \left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}} \right].
\end{aligned}$$

Similarly, the uniform norm bound follows by Lemma 4.2 and Lemma B.2: with probability at least $1-2\eta$, for any $\|x\| \leq 1$,

$$\begin{aligned}
\left| f(x) - (c_1 + \sum_j h_j(x)/m) \right| &\leq \left| f(x) - h(x) \right| + \left| h(x) - (c_1 + \sum_j h_j(x)/m) \right| \\
&\leq 2\omega_f(\delta) + \frac{\|p_1\|_{L_1}}{\sqrt{m}} \left[\sqrt{8(d+1)\ln(m+1)} + \sqrt{\ln(1/\eta)} \right].
\end{aligned}$$

For the estimates on $|c_1|$ and $\|p_1\|_{L_1}$, note setting $\phi := (2\pi\alpha)^{-1}$, note by Lemma A.3 and a change of variable $w \mapsto \phi w$ and Lemma A.1 that

$$\begin{aligned}
\|p_1\|_{L_1} &\leq 2\pi \int |\widehat{f_{|\delta} * G_\alpha}(w)| \int \mathbb{1}[|b| \leq \|w\|] \, \mathrm{d}b \, \mathrm{d}w \\
&\leq 4\pi \|f_{|\delta}\|_{L_1} \int \|w\| (2\pi\phi^2)^{d/2} G_\phi(w) \, \mathrm{d}w \\
&= 4\pi (2\pi)^{d/2} \|f_{|\delta}\|_{L_1} \int \|\phi w\| \phi^d G(w) \, \mathrm{d}w \\
&\leq 4\pi (2\pi)^{d/2} \phi^{d+1} \|f_{|\delta}\|_{L_1} \int \|w\| G(w) \, \mathrm{d}w \\
&\leq 4\sqrt{d}\pi (2\pi)^{d/2} \phi^{d+1} \|f_{|\delta}\|_{L_1}, \\
&\leq 2\|f_{|\delta}\|_{L_1} \sqrt{\frac{2\pi d}{(2\pi\alpha^2)^{d+1}}}.
\end{aligned}$$

Similarly,

$$|c_1| \leq M + \|f_{|\delta}\|_{L_1} \int \hat{G}_\alpha(w) \, \mathrm{d}w \leq M + \|f_{|\delta}\|_{L_1} (2\pi\phi^2)^{d/2}.$$

2. By Lemma 3.1 and Lemma A.1 and the various chosen parameters, for any $\|x\| \le 1$,

$$\begin{aligned}\left|b_2 + \langle x, a_2\rangle + \int p_2(\tilde{w})\sigma(\langle \tilde{w}, x\rangle)\,\mathrm{d}\tilde{w} - h(x)\right| &\le 12\pi^2 \int_{\|w\|>r_2} \|w\|^2 |\hat{h}(w)|\,\mathrm{d}w \\ &\le 24\pi^2(\sqrt{d}+7)^2 \exp(-(r_2-\sqrt{d})^2/4) \\ &\le \omega_f(\delta).\end{aligned}$$

Thus by Lemma 4.2 and Lemma B.1, defining $h_j := \|p\|_{L_1} s_j \sigma(\langle \tilde{w}_j, \cdot\rangle)$ for convenience, with probability at least $1-\eta$,

$$\begin{aligned}\left\|f - \sum_{j=1}^{m+3} h_j/m\right\|_{L_2(P)} &\le \|f-h\|_{L_2(P)} + \left\|f - (b_2 + (\cdot)^\intercal c_2 + \mathbb{E}_{p_2} s_1 h_1\right\| + \left\|\mathbb{E}_{p_2} s_1 h_1 - \sum_{j=1}^{m} s_j h_j/m\right\|_{L_2(P)} \\ &\le 3\omega_f(\delta) + \|p_2\|_{L_1} \sup_{\|\tilde{w}\|\le r_2} \|\sigma(\langle \tilde{w}, \cdot\rangle\|_{L_2(P)} \left[\frac{1+\sqrt{2\ln(1/\eta)}}{\sqrt{m}}\right] \\ &\le 3\omega_f(\delta) + 2r_2\|p\|_{L_1}\left[\frac{1+\sqrt{\ln(1/\eta)}}{\sqrt{m}}\right].\end{aligned}$$

Similarly, the uniform norm bound follows by Lemma 4.2 and Lemma B.2: with probability at least $1-2\eta$, for any $\|x\| \le 1$,

$$\begin{aligned}\left|f(x) - \sum_{j=1}^{m+3} h_j(x)/m\right| &\le \left|f(x) - h(x)\right| + \left|f(x) - (b_2 + x^\intercal c_2 + \mathbb{E}_{p_2} s_1 h_1(x)\right| + \left|\mathop{\mathbb{E}}_{p_2} s_1 h_1(x) - \sum_{j=1}^{m} s_j h_j(x)/m\right| \\ &\le 3\omega_f(\delta) + \frac{4r_2\|p_2\|_{L_1}}{\sqrt{m}}\left[1+\sqrt{\ln(1/\eta)}\right].\end{aligned}$$

For the estimates on $|c_1|$ and $\|p_1\|_{L_1}$, note setting $\phi := (2\pi\alpha)^{-1}$, note by Lemma A.3 and a change of variable $w \mapsto \phi w$ and Lemma A.1 that

$$\begin{aligned}\|p_2\|_{L_1} &\le 4\pi^2 \int |\widehat{f_{|\delta} * G_\alpha}(w)| \int \mathbb{1}[|b| \le \|w\| \le r_2]\,\mathrm{d}b\,\mathrm{d}w \\ &\le 8\pi^2 \|f_{|\delta}\|_{L_1} \int_{\|w\|\le r_2} \|w\|(2\pi\phi^2)^{d/2} G_\phi(w)\,\mathrm{d}w \\ &= 8\pi^2(2\pi)^{d/2}\|f_{|\delta}\|_{L_1} \int_{\|\phi w\|\le r_2} \|\phi w\|\phi^d G(w)\,\mathrm{d}w \\ &\le 8\pi^2(2\pi)^{d/2}\phi^{d+1}\|f_{|\delta}\|_{L_1} \int_{\|\phi w\|\le r_2} \|w\| G(w)\,\mathrm{d}w \\ &\le 8\sqrt{d}\pi^2(2\pi)^{d/2}\phi^{d+1}\|f_{|\delta}\|_{L_1} \\ &\le 2\|f_{|\delta}\|_{L_1}\sqrt{\frac{(2\pi)^3 d}{(2\pi\alpha^2)^{d+1}}}.\end{aligned}$$

Similarly,

$$
\begin{aligned}
|c_2| &\le M + \|f_{|\delta}\|_{L_1} \int (1 + \|w\|)\hat{G}_\alpha(w)\,\mathrm{d}w \\
&\le M + \|f_{|\delta}\|_{L_1} (2\pi\phi^2)^{d/2} \int (1 + \|\phi w\|)\,\mathrm{d}G(w) \\
&\le M + 2\sqrt{d}\|f_{|\delta}\|_{L_1} (2\pi\phi^2)^{d/2}, \\
\|a_2\|_2 &= \left\| \int w|\hat{h}(w)|\,\mathrm{d}w \right\| \\
&\le \int \|w\||\hat{h}(w)|\,\mathrm{d}w \\
&\le \|f_{|\delta}\|_{L_1} (2\pi\phi^2)^{d/2} \int \|\phi w\||\hat{h}(w)|\,\mathrm{d}w \\
&\le \sqrt{d}\|f_{|\delta}\|_{L_1} \phi (2\pi\phi^2)^{d/2}.
\end{aligned}
$$

□

F DEFERRED PROOFS FROM SECTION 5

Lastly, the two short proofs leading to the RKHS bounds.

Proof of Proposition 5.1. 1. By Markov's inequality

$$
\int \mathbb{1}[\|\mathcal{T}(\tilde{w})\|_2 > B]\,\mathrm{d}G(\tilde{w}) \le \frac{\|\mathcal{T}\|_{\mathcal{H}}^2}{B^2},
$$

thus by Cauchy-Schwarz, for any $\|x\| \le 1$,

$$
\begin{aligned}
\left| \left\langle \mathcal{T}_B, \Phi(x;\cdot) \right\rangle_{\mathcal{H}} - \left\langle \mathcal{T}, \Phi(x;\cdot) \right\rangle_{\mathcal{H}} \right| &= \left| \int \left(\mathcal{T}_B(\tilde{w}) - \mathcal{T}(\tilde{w})\right)^\intercal \tilde{x}\sigma'(\langle \tilde{w}, \tilde{x} \rangle)\,\mathrm{d}G(\tilde{w}) \right| \\
&\le \int \left| \mathbb{1}\left[\|\mathcal{T}(\tilde{w})\|_2 > B\right] \right| \cdot \left|\mathcal{T}(\tilde{w})^\intercal \tilde{x}\right| \cdot \left|\sigma'(\langle \tilde{w}, \tilde{x} \rangle)\right| \mathrm{d}G(\tilde{w}) \\
&\le \sqrt{\int \mathbb{1}\left[\|\mathcal{T}(\tilde{w})\|_2 > B\right]^2 \mathrm{d}G(\tilde{w})} \cdot \sqrt{\int \left(\mathcal{T}(\tilde{w})^\intercal \tilde{x}\right)^2 \mathrm{d}G(\tilde{w})} \\
&\le \frac{\|\mathcal{T}\|_{\mathcal{H}}}{B} \cdot \sqrt{\int 2\|\mathcal{T}(\tilde{w})\|^2\,\mathrm{d}G(\tilde{w})} \\
&= \frac{\|\mathcal{T}\|_{\mathcal{H}}^2 \sqrt{2}}{B}.
\end{aligned}
$$

2. Proceeding similarly, but now using Lemma A.1 to control the indicator,

$$
\begin{aligned}
\left| \left\langle \mathcal{T}_r, \Phi(x;\cdot) \right\rangle_{\mathcal{H}} - \left\langle \mathcal{T}, \Phi(x;\cdot) \right\rangle_{\mathcal{H}} \right| &= \left| \int \left(\mathcal{T}_r(\tilde{w}) - \mathcal{T}(\tilde{w})\right)^\intercal \tilde{x}\sigma'(\langle \tilde{w}, \tilde{x} \rangle)\,\mathrm{d}G(\tilde{w}) \right| \\
&\le \int \left| \mathbb{1}\left[\|\tilde{w}\|_2 > r\right] \right| \cdot \left|\mathcal{T}(\tilde{w})^\intercal \tilde{x}\right| \cdot \left|\sigma'(\langle \tilde{w}, \tilde{x} \rangle)\right| \mathrm{d}G(\tilde{w}) \\
&\le \sqrt{\int \mathbb{1}\left[\|\tilde{w}\|_2 > r\right]^2 \mathrm{d}G(\tilde{w})} \cdot \sqrt{\int \left(\mathcal{T}(\tilde{w})^\intercal \tilde{x}\right)^2 \mathrm{d}G(\tilde{w})} \\
&\le \sqrt{\exp\left(-(r - \sqrt{d})^2/2\right)} \cdot \sqrt{\int 2\|\mathcal{T}(\tilde{w})\|^2\,\mathrm{d}G(\tilde{w})} \\
&= \frac{\|\mathcal{T}\|_{\mathcal{H}}\sqrt{2}}{e^{(r-\sqrt{d})^2/4}}.
\end{aligned}
$$

□