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 neural tangent kernel (NTK) in the standard setting of microscopic changes to initial weights. Concretely, given a target function f, a target width m, and a target approximation error $\epsilon > 0$, then with high probability, moving the initial weight vectors a distance $B_{f,\epsilon}/(\epsilon\sqrt{m})$ will give a linearized finite-width NTK which is $(\sqrt{\epsilon} + B_{f,\epsilon}/\sqrt{\epsilon m})^2$-close to both the target function f, and also the shallow network which this NTK linearized. The constant $B_{f,\epsilon}$ can be independent of ϵ — particular cases studied here include f having good Fourier transform or RKHS norm — though in the worse case it scales roughly as $1/\epsilon^d$ for general continuous functions. The method of proof is to rewrite f with *equality* as an infinite-width linearized network whose weights are a *transport mapping* applied to random initialization, and to then sample from this transport mapping. This proof therefore provides another perspective on the scaling behavior of the NTK: redundancy in the weights due to resampling allows weights to be scaled down. Since the approximation rates match those in the literature for shallow networks, this work implies that universal approximation is not reliant upon any behavior outside the NTK regime.

1 MAIN RESULT AND OVERVIEW

Consider functions computed by a single ReLU layer, meaning

$$x \mapsto \sum_{j=1}^{m} s_j \sigma_{\mathrm{r}}\left(w_j^{\mathsf{T}} x + b_j\right), \tag{1.1}$$

where $\sigma_{\mathrm{r}}(z) := \max\{0, z\}$. While shallow networks are celebrated as being *universal approximators* (Funahashi, 1989; Hornik et al., 1989; Barron, 1993; Leshno et al., 1993) — 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., 2018b; Allen-Zhu et al., 2018), which replaces a network with its Taylor expansion at initialization, meaning

$$x \mapsto \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^{m} s_j \left(\tilde{v}_j + \tilde{w}_j\right)^{\mathsf{T}} \tilde{x} \sigma_{\mathrm{r}}'\left(\tilde{w}_j^{\mathsf{T}} \tilde{x}\right), \qquad \text{where } \tilde{x} = (x, 1) \in \mathbb{R}^{d+1}; \tag{1.2}$$

here $\tilde{w}_j = (w_j, b_j) \in \mathbb{R}^{d+1}$ is frozen at Gaussian initialization (henceforth the bias is collapsed in for convenience), thus the increments $(\tilde{v}_j)_{j=1}^m$ are the genuine model parameters, and the dropping of the other Taylor terms as well as the $\epsilon/\sqrt{m}$ scaling are conventional in this literature.

As eq. (1.2) is merely *affine* in the parameters, it is not shocking that gradient descent can be analyzed. What *is* shocking is that: 1. gradient descent on eq. (1.1) with small learning rate will track the behavior of eq. (1.2), 2. the weights hardly change *as a function of* m, specifically $\|\tilde{v}_j\|_2 = \mathcal{O}(1/\sqrt{m})$.

Contribution. This work provides rates of universal approximation for the NTK as defined in eq. (1.2), moreover in the "NTK setting": the increments $\tilde{v}_j$ must be small, meaning $\|\tilde{v}_j\| \leq \widetilde{\mathcal{O}}(1/\epsilon\sqrt{m})$. Some further consequences:

- As sketched in the abstract and as detailed shortly (the main result is Theorem 1.3), the approximation rates are concrete, and improve with the good behavior of the target function. The classical universal approximation results to not yield better bounds, and therefore universal approximation is not something networks can do which the NTK can not.
- The target distance here is an L_2 metric, but over a function space; by contrast, the NTK optimization literature on regression problems has the network with scaling polynomially in the network width (Du et al., 2018a; Oymak & Soltanolkotabi, 2019; Li & Liang, 2018). The analysis here must rely upon properties of the target function, necessarily ones which will explode if the target is to simulate random labels. Consequently, the present work complements the optimization work.
- The analysis boils down to two steps: (a) constructing the transport mapping, (b) sampling from it. The sampling naturally introduces a factor ${}^{1}\!/\!{}_{m}$, which is then *split* into ${}^{1}\!/\!{}_{\sqrt{m}} \cdot {}^{1}\!/\!{}_{\sqrt{m}}$, and one ${}^{1}\!/\!{}_{m}$ is then pushed into the weights — this trivial algebra contributes the scaling behavior of the earlier $\tilde{v}_j$ reporting in all the bounds!

1.1 Main result

To state the main result, a little more notation is necessary. As before, the activation will always be the ReLU $\sigma_{\mathrm{r}}(z) = \max\{0, z\}$; converting to other activations incurs constant factors and is not considered here. Weights have the biases pushed in as before, e.g., $\tilde{w} = (w, b) \in \mathbb{R}^{d+1}$. Rather than writing down increments $\tilde{v} \in \mathbb{R}^{d+1}$, this work uses *transport mappings* on weight space, meaning

$$\mathcal{T} : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1};$$

in this way, an initial weight $\tilde{w}$ always has a new weight it is clearly associated with, namely $\mathcal{T}(\tilde{w})$, and the distance traveled is $\|\mathcal{T}(\tilde{w}) - \tilde{w}\|_2$.

The notion of function space distance used here is the L_2 metric: given a probability distribution P on $\{x \in \mathbb{R}^d : \|x\| \le 1\}$, define the $L_2(P)$ metric as

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

and L_2 will denote the usual (Lebesgue) integral over all of $\mathbb{R}^d$.

In various places, G will denote a standard Gaussian $\mathcal{N}(0, I)$ of appropriate dimension; somewhat sloppy, G will be used both as a function computing a density, and as a probability law (and appear as $\mathrm{d}G$ within integrals). G_σ means a Gaussian with coordinate-wise variance σ^2, thus $\mathcal{N}(0, \sigma^2 I)$, and $f * G_\sigma$ means the convolution of a function f with G_σ.

The initial distribution will always be uniform on the sign coefficients $s \in \{\pm 1\}$, and Gaussian on the weights and biases $\tilde{w}$. Due to positive homogeneity of the ReLU, the variances and scalings are effectively pushed into the first term; the scaling here is standard in the literature.

Theorem 1.3. *Let target function $f : \mathbb{R}^d \to \mathbb{R}$, target accuracy $\epsilon > 0$, target width m, and a probability measure P over $\|x\| \le 1$ be given. Then there exists a transport mapping $\mathcal{T}$ and constant $B_{f,\epsilon}$ so that, with probability at least $1 - \delta$ over the choice of initial weights,*

$$\max_j \|\mathcal{T}(\tilde{w}_j) - \tilde{w}_j\|_2 \le \widetilde{\mathcal{O}}\left(\frac{B_{f,\epsilon}}{\epsilon\sqrt{m}}\right),$$

$$\left\| x \mapsto f(x) - \frac{\epsilon}{\sqrt{m}} \sum_j s_j \mathcal{T}(\tilde{w}_j)^{\mathsf{T}} \tilde{x} \sigma_{\mathrm{r}}'(\tilde{w}_j^{\mathsf{T}} \tilde{x}) \right\|_{L_2(P)} \le \widetilde{\mathcal{O}}\left(\epsilon + \frac{B_{f,\epsilon}}{\sqrt{m}}\right),$$

$$\left\| x \mapsto \frac{\epsilon}{\sqrt{m}} \sum_j s_j \sigma_{\mathrm{r}}(\mathcal{T}(\tilde{w}_j)^{\mathsf{T}} \tilde{x}) - \frac{\epsilon}{\sqrt{m}} \sum_j s_j \mathcal{T}(\tilde{w}_j)^{\mathsf{T}} \tilde{x} \sigma_{\mathrm{r}}'(\tilde{w}_j^{\mathsf{T}} \tilde{x}) \right\|_{L_2(P)} \le \widetilde{\mathcal{O}}\left(\left[\sqrt{\epsilon} + \frac{B_{f,\epsilon}}{\sqrt{\epsilon m}}\right]^2\right),$$

and $B_{f,\epsilon}$ can be bounded as follows.

- *If $f = h * G_\sigma$, then $B_{f,\epsilon} = \widetilde{\mathcal{O}}(\|h\|_{L_2} \sigma^{-d})$.*

- *The corresponding RKHS is universal (cf. Section 3.3 for the space and its norm), thus f may be approximated by h with $B_{f,\epsilon} = \widetilde{\mathcal{O}}(\ln(\|h\|_{\mathcal{H}}/\epsilon))$.*
- *If f is continuous, then $B_{f,\epsilon} = \widetilde{\mathcal{O}}(\|f\|_{L_2}\delta^{-d})$, where $\delta := \omega_f^{-1}(\epsilon)$ (cf. Section 3.2).*

In terms of organization, this introduction will close with further related work. Section 2 shows how a network may be sampled, *given* a transport mapping. Example transport mappings are then constructed in Section 3:

- Section 3.1 uses Fourier transforms to construct transport mappings; the construction utilizes elements of the universal approximation proof due to Barron (1993). The bounds are particularly nice in the case of convolutions, and yield the corresponding part of Theorem 1.3.
- Section 3.2 then uses the convolution bound to give a quick transport mapping for continuous functions; the use of convolutions in function approximation is classical (Weierstrass, 1885; Wendland, 2004).
- Section 3.3 constructs a natural RKHS for the NTK, and from there derives an easy bound on transport mappings via the RKHS norm of the target function.

Section 4 concludes and gives some open problems.

1.2 Further related work

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 $\int |\hat{(w)}| \cdot \|w\|_2 \,\mathrm{d}w$ is small, where $\hat{f}$ denotes the Fourier transform of f. The proofs in Section 3.1 will use elements of the corresponding proof. Like the work of (Barron, 1993), the present work also chooses to approximate in the $L_2(P)$ metric. Another related work by Sun et al. (2018) uses an RKHS approach to universal approximation, however does not consider the NTK (or the NTK setting of small weight changes).

Optimization literature. This work is motivated and inspired by the optimization literature, which introduced the NTK to study gradient descent in a variety of parallel and nearly-parallel works (Jacot et al., 2018; Du et al., 2018b; Allen-Zhu et al., 2018; Arora et al., 2019). As stated above, these results rely on a width which is polynomial in the number of data points here, whereas the analysis here must rely on properties of the target function.

One close relative to the present work is that of (Chizat & Bach, 2019), which abstracts away many aspects of the preceding proofs in an attempt to explain, amongst other things, how it is that the weights can change so little. The provided perspective is that it is due to the scaling $\epsilon/\sqrt{m}$, a view corroborated by the work here: indeed, the work here relates this scaling constant to the $1/m$ which arises naturally via random sampling.

Another related line of papers on optimization are the mean-field analyses, which relate gradient descent to a Wasserstein flow in the space of distributions on parameters (Chizat & Bach, 2018; Mei et al., 2018). The analysis here does not have any explicit ties, however it is interesting and suggestive that transport mappings appear in both.

2 Sampling from a transport

This section will establish: *given* an initial transport $\mathcal{T}$ so that

$$f(x) = \mathbb{E}_{s,\tilde{w}} s\mathcal{T}(\tilde{w})^{\mathsf{T}}\tilde{x}\sigma_{\mathrm{r}}'(\tilde{w}^{\mathsf{T}}\tilde{x}),$$

a finite width network can be created by sampling $((s_j, \tilde{w}_j))_{j=1}^m$, and then using the transport mapping to construct the NTK.

To see how the scaling naturally arises, note simply via sampling that

$$\begin{aligned}\mathbb{E}_{s,\tilde w} s\mathcal{T}(\tilde w)^{\intercal}\tilde x\sigma_{\mathrm{r}}'(\tilde w^{\intercal}\tilde x) &\approx \boldsymbol{\frac{1}{m}}\sum_{j=1}^{m} s_j\mathcal{T}(\tilde w_j)^{\intercal}\tilde x\sigma_{\mathrm{r}}'(\tilde w_j^{\intercal}\tilde x)\\ &= \boldsymbol{\left[\frac{\epsilon}{\sqrt{m}}\right]\cdot\left[\frac{1}{\epsilon\sqrt{m}}\right]}\sum_{j=1}^{m} s_j\mathcal{T}(\tilde w_j)^{\intercal}\tilde x\sigma_{\mathrm{r}}'(\tilde w_j^{\intercal}\tilde x)\\ &= \frac{\epsilon}{\sqrt{m}}\sum_{j=1}^{m} s_j\left(\frac{\mathcal{T}(\tilde w_j)}{\epsilon\sqrt{m}}\right)^{\intercal}\tilde x\sigma_{\mathrm{r}}'(\tilde w_j^{\intercal}\tilde x);\end{aligned}$$

the expression in bold immediately implies that, after sampling, the weight increments move $1/\sqrt{m}$ as far!

This only forces the new parameters to be small, but it does not cause them to be near initialization; to achieve this, they can simply be added in, noting that $\mathbb{E}_s s = 0$:

$$\frac{\epsilon}{\sqrt{m}}\sum_{j=1}^{m} s_j\left(\frac{\mathcal{T}(\tilde w_j)}{\epsilon\sqrt{m}}\right)^{\intercal}\tilde x\sigma_{\mathrm{r}}'(\tilde w_j^{\intercal}\tilde x); = \frac{\epsilon}{\sqrt{m}}\sum_{j=1}^{m} s_j\left(\frac{\mathcal{T}(\tilde w_j)}{\epsilon\sqrt{m}} + \tilde w_j\right)^{\intercal}\tilde x\sigma_{\mathrm{r}}'(\tilde w_j^{\intercal}\tilde x).$$

By construction, the new weights are now very close to the initial weights.

Lemma 2.1. *Define* $B := \sup_{\tilde w}\|\mathcal{T}(\tilde w)\|_2$*, and* $R := \sqrt{d+1} + \sqrt{2\ln(m/\delta)}$*, and*

$$\mathcal{T}_{m,\epsilon}(\tilde w) := \frac{\mathcal{T}(\tilde w)}{\epsilon\sqrt{m}} + \tilde w, \qquad f(x) := \mathbb{E}_{s,\tilde w} s\mathcal{T}(\tilde w)^{\intercal}\tilde x\mathbb{1}[\tilde w^{\intercal}\tilde x \ge 0].$$

With probability at least $1-2\delta$ *over* $((s_j,\tilde w_j))_{j=1}^{m}$*,*

$$\left\| f(x) - \frac{1}{\sqrt{m}}\sum_j s_j\mathcal{T}(\tilde w_j)^{\intercal}\tilde x\mathbb{1}[\tilde w_j^{\intercal}\tilde x \ge 0]\right\|_{L_2(P)} \le \left(\frac{B}{\sqrt{m}} + \epsilon R\right)\cdot\left(1 + \sqrt{\ln(1/\delta)}\right).$$

and$\left\|\mathcal{T}_{m,\epsilon}(\tilde w) - \tilde w\right\|_2 \le \frac{B}{\epsilon\sqrt{m}}$*, and* $\|\tilde w_j\| \le R$*.*

The key step of the proof is to apply McDiarmid's inequality to the *entire* $L_2(P)$ norm, with the randomness coming from $((s_j,\tilde w_j))_{j=1}^{m}$. In order to control the expected value, the classical Maurey sampling lemma is used (Pisier, 1980), which is also the sampling tool invoked by Barron (1993).

Lemma 2.2 (Maurey)**.** *Let basis functions* $v \mapsto g(\cdot;v)$ *be given, along with a random sample* $((s_j,v_j))_{j=1}^{m}$ *from a product measure* ν *over* $s \in \{-1,+1\}$ *and* $v \in \mathbb{R}^p$ *for some* p*. For convenience define*

$$f(x) := \int sg(x;v)\,\mathrm{d}\mu(s,v) \qquad \text{and} \qquad g_j(x) := g(x;v_j).$$

Then

$$\mathbb{E}_{((s_j,v_j))_{j=1}^{m}}\left\| f - \frac{1}{m}\sum_{j=1}^{m} s_j g_j\right\|_{L^2(P)}^2 \le \frac{\mathbb{E}_v\left\|g(\cdot;v)\right\|_{L_2(P)}^2}{m}.$$

Once again using McDiarmid's inequality, it is possible to show that the sampled NTK is close to the sampled network. As in the optimization literature, this proof relies upon an anti-concentration property, namely that the initialization is much larger than the adjustment, which requires $\|x\|$ to not be too small. This gives a constant r which is positive whenever the distribution is not simply the point mass at the origin, and thus was left in the constants of Theorem 1.3.

Lemma 2.3. *Let probability measure* P *over* $\{x \in \mathbb{R}^d : \|x\| \le 1\}$ *be given, along with a constant* r *so that* $\Pr[\|x\| \ge r] \ge 1/2$*. Let* $\mathcal{T}$ *be a transport map with* $\max_{\tilde w}\|\mathcal{T}(\tilde w)\| \le B_1$ *and* $\max_{\tilde w}\|\mathcal{T}(\tilde w) - \tilde w\|_2 \le B_2$ *. Then, with probability at least* $1-2\delta$*,*

$$\left\|\frac{\epsilon}{\sqrt{m}}\sum_j s_j\mathcal{T}(\tilde w)^{\intercal}\tilde x\mathbb{1}[\tilde w^{\intercal}\tilde x \ge 0] - \frac{\epsilon}{\sqrt{m}}\sum_j s_j\sigma_{\mathrm{r}}(\mathcal{T}(\tilde w)^{\intercal}\tilde x)\right\|_{L_2(P)} \le \epsilon B_2\left(\sqrt{\frac{B_1^2}{2\pi r^2}} + \sqrt{2\ln(1/\delta)}\right).$$

3 Constructing a few transport mappings

The previous section should how to satisfy the usual NTK setting *given* a way to write a function as a transport map from initialization; this section will construct the corresponding transport maps.

3.1 Transport maps via Fourier transforms

The first approach uses Fourier transforms. The interested reader is directed to standard analysis textbooks for an overview of Fourier transforms (Folland, 1999). The technical details are not essential to the present discussion, however; the key is that the Fourier transform gives an immediate way to write down a function as an infinite-width network! Specifically, the Fourier inversion theorem gives (for well-behaved f) the formula

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

where $\hat{f}$ is the Fourier transform of f; this is an infinite-width network with complex exponential activations! A key insight of Barron (1993) was that if the left hand side is real, then right hand side can be forced to be real: recalling that $\hat{f}$ may be written as a radial component $|\theta(w)| \leq 1$ and a magnitude component $|\hat{f}(w)|$,

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

After this step, the proof here goes a separate way, though interestingly still relies on many of the same quantities. In order to introduce a Gaussian distribution on w, the entire integrand may be scaled by ${}^{G(w)}\!/_{G(w)}$; further introducing the bias b may be accomplished via a variety of integration tricks, most notably writing

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

where the last expression now has the indicators which appear in the NTK! After some changes of variable and algebra, an initial transport map is explicitly given as follows.

Lemma 3.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be given, and define*

$$\mathcal{T}(w,b)_{d+1} = b' := 2\left[f(0) + \int |\hat{f}(w)|\,\mathrm{d}w\right] + \frac{|\hat{f}(w)|}{G(w)}\sin(2\pi z)e^{b^2/2}\mathbb{1}[\theta(w) - b \geq 0],$$

and

$$g(x) := \int \mathcal{T}(\tilde{w})^\top \tilde{x}\mathbb{1}[\tilde{w}^\top \tilde{x} \geq 0].$$

If this g exists and is well-defined, then $f = g$ everywhere.

While it is nice that this transport mapping gives an equality, it is difficult to use with Section 2 due to certain large factors, for instance $e^{b^2/2}$. Instead, all the mapping in the present work will use some sort of truncation of the transport mapping. Truncating the above transport gives the following lemma.

Lemma 3.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be given, and define*

$$\mathcal{T}(w,b)_{d+1} = b' := 2\left[f(0) + \int |\hat{f}(w)|\,\mathrm{d}w\right] + \frac{|\hat{f}(w)|}{G(w)}\sin(2\pi z)e^{b^2/2}\mathbb{1}[\theta(w) - b \geq 0]\mathbb{1}[\theta \geq -B]\mathbb{1}[\|w\| \leq B],$$

and

$$g(x) := \int \mathcal{T}(\tilde{w})^\top \tilde{x}\mathbb{1}[\tilde{w}^\top \tilde{x} \geq 0].$$

Then

$$\sup_{\tilde{w}} \|\mathcal{T}(w)\|_2 \leq 2|f(0)| + 2\int |\hat{f}(w)|\,\mathrm{d}w + \sup_{\substack{\|w\|\leq B\\ |b|\leq B}} \frac{|\hat{f}(w)|}{G(w)} e^{b^2/2}.$$

$$\sup_{\|x\|\leq 1} |f(x) - g(x)| \leq \int_{\|w\|\geq B} |\hat{f}(w)|\,\mathrm{d}w + \int_{\|w\|\geq B} |\hat{f}(w)| \cdot \|w\|\,\mathrm{d}w.$$

The quantities in this bound are quite complicated; evaluating them on some easy cases gives the following estimates.

Lemma 3.3. *Let G_σ be a Gaussian density with covariance $\sigma^2 I$ where $\sigma \leq 1/2\pi$, and $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function whose Fourier transform $\hat{f}$ exists, and let $\epsilon > 0$ be given.*

1. *Let $\mathcal{T}_1$ denote the truncated transported map for G_σ as defined in Lemma 3.2, and let g_1 denote the mapping defined by $\mathcal{T}_1$. If $B = \mathcal{O}(\ln(1/\epsilon)/\sigma)$, then*

$$\sup_{\|x\|\leq 1} |G_\sigma(x) - g_1(x)| \leq \epsilon \qquad \text{and} \qquad \sup_{\tilde{w}} \|\mathcal{T}_1(\tilde{w})\| = \mathcal{O}\left((2\pi\sigma^2)^{d/2}\right).$$

2. *Let $\mathcal{T}_2$ denote the truncated transported map for $f * G_\sigma$ as defined in Lemma 3.2, and let g_2 denote the mapping defined by $\mathcal{T}_2$. If $B = \mathcal{O}(\ln(\|f\|_{L_2}/\epsilon)/\sigma)$, then*

$$\sup_{\|x\|\leq 1} |(f * G_\sigma)(x) - g_2(x)| \leq \epsilon \qquad \text{and} \qquad \sup_{\tilde{w}} \|\mathcal{T}_2(\tilde{w})\| = \mathcal{O}\left(\|f\|_{L_2}(2\pi\sigma^2)^{d/2}\right).$$

3.2 Transport maps for continuous functions

Approximation of continuous functions now proceeds by the apocryphal technique of *randomized smoothing*, or rather convolution with a Gaussian (Weierstrass, 1885; Wendland, 2004). The precise bound will rely upon a quantity which converts between an output tolerance ϵ of a function and an input tolerance δ.

Definition 3.4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be given, and define *modulus of continuity* ω_f as

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

◊

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 Lemma 3.3 only required an L_2 bound on the pre-convolution function f!

Lemma 3.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$ be given, and define with $M := \sup_{\|x\|\leq 1+\delta} |f(x)|$. Let G_σ denote a Gaussian with variance*

$$\sigma^2 := \delta^2/(d + \sqrt{8d\ln(4M/\omega_f(\delta))}$$

Then

$$\sup_{\|x\|\leq 1} \|f - f * G_\sigma\|_{L_2(P)} \leq 2\omega_f(\delta).$$

Moreover, if f is continuous, then $\lim_{\delta\to 0} \omega_f(\delta) \to 0$.

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 due not matter due to the Gaussian convolution. The full details are in the appendix.

3.3 A natural Reproducing Kernel Hilbert Space

The natural Hilbert space is directly on transport mappings:

$$\|\mathcal{T}\|_{\mathcal{H}}^2 = \int \|\mathcal{T}(\tilde{w})\|_{\mathcal{H}}^2\,\mathrm{d}G(\tilde{w}).$$

Defining a feature mapping $\Phi_x(\tilde{w}) = \tilde{x}\mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0]$ gives function representation

$$\langle \mathcal{T}, \Phi_x \rangle_{\mathcal{H}} = \int \mathcal{T}(\tilde{w})^\intercal \tilde{x}\mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \, \mathrm{d}G(\tilde{w})$$

as desired, and kernel

$$k(x, x') = \tilde{x}^\intercal \tilde{x}' \frac{\pi - \arccos(\tilde{x}^\intercal \tilde{x}' / (\|\tilde{x}\| \cdot \|\tilde{x}'\|))}{2\pi}.$$

Since the Taylor series of this dot product kernel has infinitely many even terms, the Kernel is universal (Steinwart & Christmann, 2008, Lemma 4.55). (Indeed, it is even universal if biases and the point $x = 0$ are excluded!)

As the space is directly over transport mappings, all that remains is to prove that truncation still yields a good transport mapping; after this, the tools of Section 2 may be applied. Meanwhile, controlling this truncation is nothing more than an application of Cauchy-Schwarz and Gaussian concentration.

Lemma 3.6. *Let $\mathcal{T}_{\leq B}(w) := \mathcal{T}(w)\mathbb{1}[\|w\| \leq B]$ denote the truncated transport, and P a probability measure on $\|x\| \leq 1$. Then*

$$\left\| \langle \mathcal{T}, \Phi_\cdot \rangle - \left\langle \mathcal{T}_{\leq B}, \Phi_\cdot \right\rangle \right\|_{L_2(P)}^2 \leq \|\mathcal{T}\|_{\mathcal{H}}^2 \exp(-(B-d)^2/(10d^2)).$$

4 Open problems

The optimization community has also investigated phenomena learned by deep networks which go beyond NTK (Allen-Zhu & Li, 2019); is there some way to exhibit this for the present approximation setting; namely an approximation-theoretic statement which is possible with deep networks but impossible with the NTK?

More on the technical side, there are many possible improvements. The transport maps plugged into Section 2 are all truncated; is there some way to avoid the use of truncation? As another technical point, Section 3.3 pointed out that universal approximation does not need biases; is there an explicit construction of transport maps, perhaps even with the techniques here, which does not include biases? Lastly, the work here only explicitly discusses the ReLU; other activations are left out as, on the one hand, one can always swap in other activations with the penalty of some constant factors, and perhaps a worse dependence on $1/\epsilon$; on the other hand, are there some general approximation-theoretic claims which work much better with certain activations than others?

References

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *CoRR*, abs/1905.10337, 2019. URL `http://arxiv.org/abs/1905.10337`.

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

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 preprint arXiv:1901.08584*, 2019.

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

Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. *arXiv e-prints*, art. arXiv:1805.09545, May 2018.

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

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

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

K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Netw.*, 2(3):183–192, May 1989. ISSN 0893-6080.

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. URL `http://dblp.uni-trier.de/db/journals/nn/nn6.html#LeshnoLPS93`.

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. *arXiv e-prints*, art. arXiv:1804.06561, Apr 2018.

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

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

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

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

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. doi: 10.1017/CBO9780511617539.

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

A Deferred proofs

Proof of Lemma 2.1. First, by construction,

$$\left\|\mathcal{T}_{m,\epsilon}(\tilde{w}) - \tilde{w}\right\|_2 \le \left\|\frac{\mathcal{T}(\tilde{w})}{\epsilon\sqrt{m}}\right\|_2 = \frac{B}{\epsilon\sqrt{m}}.$$

To control the function approximation error, note by Gaussian concentration and a union bound, with probability at least $1-\delta$,

$$\max_j \|\tilde{w}_j\| \le \sqrt{d+1} + \sqrt{2\ln(m/\delta)} = R;$$

henceforth condition away this failure event, meaning subsequently $\|\tilde{w}\|_2 \le R$.

Let $\left((s_j, \tilde{w}_j)\right)_{j=1}^m$ be drawn IID from μ, define for convenience $g_j := s_j \mathcal{T}_{m,\epsilon}(\tilde{w}_j)^\intercal \tilde{x} \mathbb{1}[\tilde{w}_j^\intercal \tilde{x} \ge 0]$ and $f(x) := \mathbb{E}_{s,\tilde{w}} g_1(x)$, and first note the mapping

$$\left((s_j, w_j)\right)_{j=1}^m \mapsto \left\| f - \frac{\epsilon}{\sqrt{m}} \sum_j s_j g_j \right\|_{L_2(P)}$$

satisfies the bounded differences condition with constant $C := \frac{2}{\sqrt{m}}(R\epsilon + B/\sqrt{m})$: using the general inequality $\big|\|p-q\| - \|r-q\|\big| \le \|p-r\|$, given two parameter collections $((s_j, \tilde{w}_j))_{j=1}^m$ and $((s_j', \tilde{w}_j'))_{j=1}^m$ and corresponding g_j and g_j' only differing in one index k,

$$\begin{aligned}
\left\| f - \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^m s_j g_j \right\|_{L^2(P)} - \left\| f - \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^m s_j' g_j' \right\|_{L^2(P)} &\le \left\| \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^m s_j g_j - \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^m s_j' g_j' \right\|_{L^2(P)} \\
&= \frac{\epsilon}{\sqrt{m}} \big\| s_k g_k - s_k' g_k' \big\|_{L^2(P)} \\
&\le \frac{2\epsilon}{\sqrt{m}} \sup_{\tilde{w}} \big\| g(\cdot; \tilde{w}) \big\|_{L^2(P)} \\
&\le \frac{2\epsilon}{\sqrt{m}} \|\mathcal{T}_{m,\epsilon}(\tilde{w})\|_2 \\
&\le \frac{2\epsilon}{\sqrt{m}} \left(\|\mathcal{T}_{m,\epsilon}(\tilde{w}) - \tilde{w}\|_2 + \|\tilde{w}\| \right) \le C.
\end{aligned}$$

By McDiarmid's inequality, with probability at least $1 - \delta$,

$$\left\| f - \frac{1}{m} \sum_{j=1}^m s_j g_j \right\|_{L^2(P)} \le \mathbb{E}_{((s_j, v_j))_{j=1}^m} \left\| f - \frac{1}{m} \sum_{j=1}^m s_j g_j \right\|_{L^2(P)} + C\sqrt{m \ln(1/\delta)/2}.$$

To finish, note since $\mathbb{E}_{s,\tilde{w}} s = 0$ that

$$\begin{aligned}
f(x) &= \mathbb{E}_{s,\tilde{w}} s \mathcal{T}(\tilde{w})^\intercal \tilde{x} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \ge 0] \\
&= \epsilon \mathbb{E}_{s,\tilde{w}} s \left(\frac{\mathcal{T}(\tilde{w})}{\epsilon} + \tilde{w}\sqrt{m} \right)^\intercal \tilde{x} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \ge 0] \\
&= \mathbb{E} \frac{\epsilon}{\sqrt{m}} \sum_j s_j g_j,
\end{aligned}$$

whereby Lemma 2.2 grants

$$\begin{aligned}
\mathbb{E}_{((s_j, v_j))_{j=1}^m} \left\| f - \frac{\epsilon}{\sqrt{m}} \sum_{j=1}^m s_j g_j \right\|_{L^2(P)} &\le \sqrt{ \mathbb{E}_{((s_j, v_j))_{j=1}^m} \left\| f - \frac{1}{m} \sum_{j=1}^m s_j g_j \right\|_{L^2(P)}^2 } \\
&\le \sqrt{ \frac{1}{m} \sup_{\|\tilde{w}\| \le R} \|\epsilon \sqrt{m} \mathcal{T}_{m,\epsilon}\|_2^2 } \\
&\le \frac{B}{\sqrt{m}} + \epsilon R = \frac{C\sqrt{m}}{2}.
\end{aligned}$$

□

Proof of Lemma 2.2. Following the usual Maurey scheme (Pisier, 1980),

$$\begin{aligned}
\mathbb{E}_{((s_j,v_j))_{j=1}^m}\left\|f - m^{-1}\sum_j s_j g_j\right\|_{L_2(P)}^2 &= \frac{1}{m^2}\mathbb{E}_{((s_j,v_j))_{j=1}^m}\left\|\sum_j \left(f - s_j g_j\right)\right\|_{L_2(P)}^2 \\
&= \frac{1}{m^2}\mathbb{E}_{((s_j,v_j))_{j=1}^m}\sum_j \left\|f - s_j g_j\right\|_{L_2(P)}^2 \\
&= \frac{1}{m}\mathbb{E}_{((s_j,v_j))_{j=1}^m}\|f - s_1 g_1\|_{L_2(P)}^2 \\
&= \frac{1}{m}\mathbb{E}_{(s_1,v_1)}\|f - s_1 g_1\|_{L_2(P)}^2 \\
&= \frac{1}{m}\mathbb{E}_{v_1}\left(\|g_1\|_{L_2(P)}^2 - \|f\|_{L_2(P)}^2\right) \\
&\le \frac{1}{m}\mathbb{E}_{v_1}\|g_1\|_{L_2(P)}^2 .
\end{aligned}$$

□

Proof of Lemma 2.3. Define $S := ((s_j, \tilde{w}_j))_{j=1}^m$; the proof proceeds via the bounded differences property on the map

$$g(S) := \left\|\frac{1}{\sqrt{m}}\sum_j s_j \mathcal{T}(\tilde{w})^\intercal \tilde{x}\mathbb{1}[\tilde{w}^\intercal \tilde{x} \ge 0] - \frac{1}{\sqrt{m}}\sum_j s_j \sigma_{\mathrm{r}}(\mathcal{T}(\tilde{w})^\intercal \tilde{x})\right\|_{L_2(P)},$$

which can be simplified via $\sigma_{\mathrm{r}}(z) = z\mathbb{1}[z \ge 0]$ into

$$\begin{aligned}
g(((s_j, \tilde{w}_j))_{j=1}^m) &= \left\|\frac{1}{\sqrt{m}}\sum_j s_j \mathcal{T}(\tilde{w})^\intercal \tilde{x}\mathbb{1}[\tilde{w}^\intercal \tilde{x} \ge 0] - \frac{1}{\sqrt{m}}\sum_j s_j \mathcal{T}(\tilde{w})^\intercal \tilde{x}\mathbb{1}[\mathcal{T}(\tilde{w})^\intercal \tilde{x} \ge 0]\right\|_{L_2(P)}, \\
&= \frac{1}{\sqrt{m}}\left\|\sum_j s_j \mathcal{T}(\tilde{w})^\intercal \tilde{x}\left(\mathbb{1}[\tilde{w}^\intercal \tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w})^\intercal \tilde{x} \ge 0]\right)\right\|_{L_2(P)}.
\end{aligned}$$

To verify the bounded differences property, note by the general inequality $\big|\|p\| - \|q\|\big| \le \|p - q\|$ with pairs $S := ((s_j, \tilde{w}_j))_{j=1}^m$ and $S' := ((s_j', \tilde{w}_j'))_{j=1}^m$ differing on a single element k that

$$\begin{aligned}
\left|g(S) - g(S')\right| &\le \frac{1}{\sqrt{m}}\Bigg\|\sum_j s_j \mathcal{T}(\tilde{w}_j)^\intercal \tilde{x}\left(\mathbb{1}[\tilde{w}_j^\intercal \tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w}_j)^\intercal \tilde{x} \ge 0]\right) \\
&\qquad - \sum_j s_j' \mathcal{T}(\tilde{w}_j')^\intercal \tilde{x}\left(\mathbb{1}[\tilde{w}_j'^\intercal \tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w}_j')^\intercal \tilde{x} \ge 0]\right)\Bigg\|_{L_2(P)}, \\
&\le \frac{1}{\sqrt{m}}\Big\|s_k \mathcal{T}(\tilde{w}_k)^\intercal \tilde{x}\left(\mathbb{1}[\tilde{w}_k^\intercal \tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w}_k)^\intercal \tilde{x} \ge 0]\right) \\
&\qquad - s_k' \mathcal{T}(\tilde{w}')^\intercal \tilde{x}\left(\mathbb{1}[\tilde{w}_k'^\intercal \tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w}_k')^\intercal \tilde{x} \ge 0]\right)\Big\|_{L_2(P)}, \\
&\le \frac{1}{\sqrt{m}}\left(\left\|\mathcal{T}(\tilde{w}_k)^\intercal \tilde{x}\right\|_{L_2(P)} + \left\|\mathcal{T}(\tilde{w}_k')^\intercal \tilde{x}\right\|_{L_2(P)}\right) \\
&\le \frac{2B_1}{\sqrt{m}}.
\end{aligned}$$

Thus, by McDiarmid's inequality, with probability at least $1 - \delta$,

$$g(S) \le \mathbb{E}g(S) + \sqrt{2B_1^2\ln(1/\delta)}.$$

To bound the expectation, since $\mathbb{E}g(S) = \mathbb{E}\sqrt{g(S)^2} \le \sqrt{\mathbb{E}g(S)^2}$, defining

$$V_j(x) := \Big(\mathbb{1}[\tilde{w}_j^{\intercal}\tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w}_j)^{\intercal}\tilde{x} \ge 0]\Big),$$

the quantity $\mathbb{E}g(S)^2$ may be bounded as

$$\begin{aligned}
m\mathbb{E}g(S)^2 &= \mathbb{E}_S\mathbb{E}_x\left[\sum_j s_j\mathcal{T}(\tilde{w})^{\intercal}\tilde{x}\left(\mathbb{1}[\tilde{w}^{\intercal}\tilde{x} \ge 0] - \mathbb{1}[\mathcal{T}(\tilde{w})^{\intercal}\tilde{x} \ge 0]\right)\right]^2 \\
&= \mathbb{E}_S\mathbb{E}_x\left[\sum_i \left(\mathcal{T}(\tilde{w}_i)^{\intercal}\tilde{x}V_i(x)\right)^2 + \sum_{i\ne j} s_is_j\mathcal{T}(\tilde{w}_i)^{\intercal}\tilde{x}\mathcal{T}(\tilde{w}_j)^{\intercal}\tilde{x}V_i(x)V_j(x)\right] \\
&\le \mathbb{E}_S\mathbb{E}_x\left[B_1^2\sum_i V_i(x)^2\right] \\
&= mB_1^2\mathbb{E}_{\tilde{w}_1}\mathbb{E}_xV_1(x)^2.
\end{aligned}$$

To analyze these further, note $|V_1(x)| = \mathbb{1}\left[\mathbb{1}[\tilde{w}_1^{\intercal}\tilde{x} \ge 0] \ne \mathbb{1}[\mathcal{T}(\tilde{w}_1)^{\intercal}\tilde{x} \ge 0]\right] =: A(\tilde{w}, x)$, so it suffices to upper bound $\mathbb{E}_{x,\tilde{w}}A(\tilde{w}, x)$. By rotational invariance, this is upper bounded by the definition of r and Gaussian concentration on a single-variate Gaussian Z, specifically by

$$\begin{aligned}
\Pr\left[|Z\|x\|| \le \|\mathcal{T}(\tilde{w}_1) - \tilde{w}_1\|\right] \ge \frac{1}{2}\Pr\left[|Zr| \le \|\mathcal{T}(\tilde{w}_1) - \tilde{w}_1\|\right] \\
\le \frac{1}{2\sqrt{2\pi}}\int_{-\|\mathcal{T}(\tilde{w}_1)-\tilde{w}_1\|/r}^{\|\mathcal{T}(\tilde{w}_1)-\tilde{w}_1\|/r} e^{-Z^2/2}\,\mathrm{d}Z \\
\le \frac{\|\mathcal{T}(\tilde{w}_1) - \tilde{w}_1\|}{r\sqrt{2\pi}} \le \frac{B_2}{r\sqrt{2\pi}},
\end{aligned}$$

which together gives

$$g(S) \le \sqrt{\mathbb{E}g(S)^2} + B_2\sqrt{2\ln(1/\delta)} \le B_2\left(\sqrt{\frac{B_1^2}{2\pi r^2}} + \sqrt{2\ln(1/\delta)}\right).$$

The final bounds come from multiplying back in the factor ϵ in the statement which was dropped from g. □

Proof of Lemma 3.1. Proceeding as in the initial steps of (Barron, 1993), and letting $\theta(w)$ be a function with so that $\hat{f}(w)$ may be written as a radial component $|\theta(w)| \le 1$ and a magnitude component $|\hat{f}(w)|$,

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

The next step is similar; Barron (1993) introduces a factor $\|w\|$ to control cos, which does not have compact support, whereas this proof will introduce the Gaussian density $G(w)$ in order to recover the random initialization:

$$f(x) - f(0) = \int \frac{|\hat{f}(w)|}{G(w)}\cos\left(2\pi(x^{\intercal}w + \theta(w))\right)G(w)\,\mathrm{d}w.$$

The next step is to replace cos and introduce b, where the proofs now differ. Focusing on the cos term (which is the only term with x), setting $h(z) := \cos(2\pi z)$ and $z := w^{\intercal}x + \theta(w)$ for convenience, assuming $z \geq 0$ (the other case is analogous)

$$\begin{aligned}
h(z) - h(0) &= \int_0^z h'(b)\,\mathrm{d}b \\
&= \int_{\mathbb{R}} h'(b) \cdot \mathbb{1}[z - b \geq 0] \cdot \mathbb{1}[b \geq 0]\,\mathrm{d}b \\
&= \int_{\mathbb{R}} h'(b) \cdot \mathbb{1}[w^\intercal x + \theta(w) - b \geq 0] \cdot \mathbb{1}[b \geq 0]\,\mathrm{d}b \\
&= -\int_{\mathbb{R}} h'(\theta(w) - b) \cdot \mathbb{1}[w^\intercal x + b \geq 0] \cdot \mathbb{1}[\theta(w) - b \geq 0]\,\mathrm{d}b \qquad \text{via } b \mapsto \theta(w) - b \\
&= -\int_{\mathbb{R}} h'(\theta(w) - b) e^{b^2/2} \cdot \mathbb{1}[w^\intercal x + b \geq 0] \cdot \mathbb{1}[\theta(w) - b \geq 0] e^{-b^2/2}\,\mathrm{d}b.
\end{aligned}$$

Plugging this back in,

$$\begin{aligned}
f(x) - f(0) &= \int \frac{|\hat{f}(w)|}{G(w)} \cos\left(2\pi(x^\intercal w + \theta(w))\right) G(w)\,\mathrm{d}w \\
&= h(0) \int |\hat{f}(w)|\,\mathrm{d}w + 2\pi \int \frac{|\hat{f}(w)|}{G(w)} \sin(2\pi z) e^{b^2/2} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \mathbb{1}[\theta(w) - b \geq 0]\,\mathrm{d}G(\tilde{w}).
\end{aligned}$$

Define $\mathcal{T}(w, b) = (0, b')$, where

$$\mathcal{T}(w, b)_{d+1} = b' := 2\left[f(0) + \int |\hat{f}(w)|\,\mathrm{d}w\right] + \frac{|\hat{f}(w)|}{G(w)} \sin(2\pi z) e^{b^2/2} \mathbb{1}[\theta(w) - b \geq 0],$$

whereby the fact $\mathbb{E}_{\tilde{w}} \mathbb{1}[\tilde{w}^\intercal \tilde{x}]\,\mathrm{d}\tilde{w} = {}^1\!/\!_2$ grants

$$\begin{aligned}
\int \mathcal{T}(\tilde{w})^\intercal \tilde{x} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0]\,\mathrm{d}G(\tilde{w}) &= \int \mathcal{T}(\tilde{w})_{d+1} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0]\,\mathrm{d}G(\tilde{w}) \\
&= 2 \int \left[f(0) + \int |\hat{f}(w)|\,\mathrm{d}w\right] \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0]\,\mathrm{d}G(\tilde{w}) \\
&\quad + \int \frac{|\hat{f}(w)|}{G(w)} \sin(2\pi z) e^{b^2/2} \mathbb{1}[\theta(w) - b \geq 0] \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0]\,\mathrm{d}G(\tilde{w}) \\
&= f(x).
\end{aligned}$$

□

Proof of Lemma 3.2. Directly by construction, since $|\theta(w)| \leq 1$,

$$\sup_{\tilde{w}} \|\mathcal{T}(\tilde{w})\|_2 \leq 2|f(0)| + 2 \int |\hat{f}(w)|\,\mathrm{d}w + \sup_{\substack{\|w\| \leq B \\ |b| \leq B}} \frac{|\hat{f}(w)|}{G(w)} e^{b^2/2}.$$

For the function approximation error, for any $\|x\| \leq 1$, using the form of the exact transport mapping for f from Lemma 3.1,

$$\begin{aligned}
&\left| f(x) - \int \mathcal{T}(\tilde{w})^\intercal \tilde{x} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \, \mathrm{d}G(\tilde{w}) \right| \\
&= \left| f(x) - \int \mathcal{T}(\tilde{w})_{d+1} \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \, \mathrm{d}G(\tilde{w}) \right| \\
&\leq \left| \int_{\|w\|>B} \int \frac{|\hat{f}(w)|}{G(w)} \sin(2\pi z) e^{b^2/2} \mathbb{1}[\theta(w) - b \geq 0] \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \, \mathrm{d}G(b) \, \mathrm{d}G(w) \right| \\
&\leq \int_{\|w\|\geq B} \int \left| \frac{|\hat{f}(w)|}{G(w)} \sin(2\pi z) e^{b^2/2} \mathbb{1}[\theta(w) - b \geq 0] \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \right| \mathrm{d}G(b) \, \mathrm{d}G(w) \\
&= \int_{\|w\|\geq B} \int \left| \frac{|\hat{f}(w)|}{G(w)} \sin(2\pi z) \mathbb{1}[\theta(w) - b \geq 0] \mathbb{1}[\tilde{w}^\intercal \tilde{x} \geq 0] \right| \mathrm{d}b \, \mathrm{d}G(w) \\
&\leq \int_{\|w\|\geq B} \frac{|\hat{f}(w)|}{G(w)} \int \Big| |\sin(2\pi z)| \mathbb{1}[b \leq 1] \mathbb{1}[-b \leq w^\intercal x] \Big| \, \mathrm{d}b \, \mathrm{d}G(w) \\
&\leq \int_{\|w\|\geq B} \frac{|\hat{f}(w)|}{G(w)} \max\{0, 1 - w^\intercal x\} \, \mathrm{d}G(w) \\
&\leq \int_{\|w\|\geq B} |\hat{f}(w)| \max\{0, 1 - w^\intercal x\} \, \mathrm{d}w \\
&\leq \int_{\|w\|\geq B} |\hat{f}(w)| \, \mathrm{d}w + \int_{\|w\|\geq B} |\hat{f}(w)| \cdot \|w\| \, \mathrm{d}w.
\end{aligned}$$

□

Proof of Lemma 3.3. The proof plugs the estimates from Lemma A.1 into the truncated transport bounds in Lemma 3.2. □

Lemma A.1. *Let G_σ be a Gaussian density with covariance $\sigma^2 I$ where $\sigma \leq 1/2\pi$, and $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function whose Fourier transform $\hat{f}$ exists, and let $B \geq 1$ be arbitrary.*

1. *G_σ satisfies*

$$\int |\hat{G}_\sigma(w)| \, \mathrm{d}w = \left(2\pi\sigma^2\right)^{-\frac{d}{2}},$$

$$\sup_{\substack{\|w\|\leq B \\ |b|\leq B}} \frac{|\hat{G}_\sigma(w)|}{G(w)} e^{b^2/2} = (2\pi)^{\frac{d}{2}} \exp\left((1 - (\sqrt{2}\pi\sigma)^2) B^2\right),$$

$$\int_{\|w\|>B} |\hat{G}_\sigma(w)| \cdot \|w\|_2 \, \mathrm{d}w \leq 4^d \sqrt{d} \left(2\pi\sigma^2\right)^{-\frac{d+1}{2}} B^d \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).$$

2. *The convolution $f_\sigma := f * G_\sigma$ satisfies*

$$\int |\hat{f}_\sigma(w)| \, \mathrm{d}w \leq \int |\hat{G}_\sigma(w)| \, \mathrm{d}w \cdot \sqrt{\int f^2(x) \, \mathrm{d}x},$$

$$\sup_{\substack{\|w\|\leq B \\ |b|\leq B}} \frac{|\hat{f}_\sigma(w)|}{G(w)} e^{b^2/2} \leq \sup_{\substack{\|w\|\leq B \\ |b|\leq B}} \frac{|\hat{G}_\sigma(w)|}{G(w)} e^{b^2/2} \cdot \sqrt{\int f^2(x) \, \mathrm{d}x},$$

$$\int_{\|w\|>B} |\hat{f}_\sigma(w)| \cdot \|w\|_2 \, \mathrm{d}w \leq \int_{\|w\|>B} |\hat{G}_\sigma(w)| \cdot \|w\|_2 \, \mathrm{d}w \cdot \sqrt{\int f^2(x) \, \mathrm{d}x}.$$

Proof of Lemma A.1. 1. The Gaussian density is separable, $G_\sigma(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x_i^2}{2\sigma^2}\right)$, so its Fourier transform can be written as

$$\hat{G}_\sigma(w) = \prod_{i=1}^{d} e^{-2\pi^2\sigma^2 w_i^2} = \left(2\pi\sigma^2\right)^{-\frac{d}{2}} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi(2\pi\sigma)^{-2}}} \exp\left(-\frac{w_i^2}{2(2\pi\sigma)^{-2}}\right),$$

and immediately $\int |\hat{G}_\sigma(w)|\,\mathrm{d}w = \left(2\pi\sigma^2\right)^{-\frac{d}{2}}$.

Next, note that

$$\frac{|\hat{G}_\sigma(w)|}{G(w)} e^{b^2/2} = (2\pi)^{\frac{d}{2}} \exp\left(\frac{\|w\|^2(1-(2\pi\sigma)^2)+b^2}{2}\right),$$

so

$$\sup_{\substack{\|w\|\le B\\ |b|\le B}} \frac{|\hat{G}_\sigma(w)|}{G(w)} e^{b^2/2} = \begin{cases} (2\pi)^{\frac{d}{2}} \exp\left(B^2(2-(2\pi\sigma)^2)/2\right) & \text{if } \sigma < 1/2\pi \\ (2\pi)^{-\frac{d}{2}} \exp\left(B^2/2\right) & \text{else.} \end{cases}$$

Lastly, it is shown via induction on d that

$$\int_{\|w\|>B} |\hat{G}_\sigma(w)| \cdot \|w\|_2 \,\mathrm{d}w \le 4^d\sqrt{d}\left(2\pi\sigma^2\right)^{-\frac{d+1}{2}} B^d \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).$$

When $d := k = 1$,

$$\int_{|w_1|>B} |\hat{G}_\sigma(w_1)| \cdot |w_1| \,\mathrm{d}w_1 = 2\int_{w_1>B} e^{-2\pi^2\sigma^2 w_1^2} \cdot w_1 \,\mathrm{d}w_1 = 2\left[-\frac{1}{4\pi^2\sigma^2} e^{-2\pi^2\sigma^2 w_1^2}\right]_{w_1=B}^{\infty}$$
$$= \frac{1}{\pi}\left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).$$

Now, let $d := k > 1$. By breaking down the area of integration into $\|w\| > B \iff |w_1| \le B, \|w_{2:d}\| > \sqrt{B^2 - w_1^2}$ or $|w_1| > B, \|w_{2:d}\| \ge 0$, the following fact can be shown by induction and the Chernoff bound on Gaussian tail, that

$$\int_{\|w\|>B} |G_\sigma(w)| \,\mathrm{d}w \le 2^{d-1} \exp\left(-\frac{B^2}{2\sigma^2}\right).$$

Because $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, the quantity of interest can be evaluated via

$$\int_{\|w\|>B} |\hat{G}_\sigma(w)| \cdot \|w\|_2 \,\mathrm{d}w \le \int_{\|w\|>B} |\hat{G}_\sigma(w)| \cdot (|w_1| + \|w_{2:k}\|_2) \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1$$
$$= \left(\iint_{|w_1|\le B,\ \|w_{2:k}\|>\sqrt{B^2-w_1^2}} + \iint_{|w_1|>B,\ \|w_{2:k}\|\ge 0}\right) |\hat{G}_\sigma(w)| \cdot (|w_1| + \|w_{2:k}\|_2) \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1,$$

and the four terms are computed separately below.

- By the Chernoff bound on Gaussian tail,

$$\iint_{|w_1|\le B,\ \|w_{2:k}\|>\sqrt{B^2-w_1^2}} |\hat{G}_\sigma(w)| \cdot |w_1| \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1$$
$$= \int_{|w_1|\le B} |\hat{G}_\sigma(w_1)| \cdot |w_1| \int_{\|w_{2:k}\|>\sqrt{B^2-w_1^2}} |\hat{G}_\sigma(w_{2:k})| \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1$$
$$\le \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}} 2^{k-1} \int_{|w_1|\le B} |\hat{G}_\sigma(w_1)| \exp\left(-\frac{B^2 - w_1^2}{2(2\pi\sigma)^{-2}}\right) \cdot |w_1| \,\mathrm{d}w_1$$
$$= \left(\pi\sigma^2/2\right)^{-\frac{k-1}{2}} \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right) \cdot \int_{|w_1|\le B} |w_1| \,\mathrm{d}w_1$$
$$= \left(\pi\sigma^2/2\right)^{-\frac{k-1}{2}} B^2 \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).$$

- By the inductive hypothesis,

$$\begin{aligned}
&\iint_{|w_1|\le B,\ \|w_{2:k}\|>\sqrt{B^2-w_1^2}} |\hat{G}_\sigma(w)| \cdot \|w_{2:k}\|_2 \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1 \\
&\quad\le 4^{k-1}\sqrt{k-1}\left(2\pi\sigma^2\right)^{-\frac{k+1}{2}} B^{k-1}\cdot \int_{|w_1|\le B} |\hat{G}_\sigma(w_1)| \exp\left(-\frac{B^2-w_1^2}{2(2\pi\sigma)^{-2}}\right)\mathrm{d}w_1 \\
&\quad= 4^{k-1}\sqrt{k-1}\left(2\pi\sigma^2\right)^{-\frac{k+1}{2}} B^{k-1} \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right)\cdot \int_{|w_1|\le B} |\hat{G}_\sigma(w_1)|\,\mathrm{d}w_1 \\
&\quad= 4^{k-1}\cdot 2\sqrt{k-1}\left(2\pi\sigma^2\right)^{-\frac{k+1}{2}} B^{k} \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).
\end{aligned}$$

- By the fundamental theorem of calculus,

$$\begin{aligned}
&\iint_{|w_1|> B,\ \|w_{2:k}\|\ge 0} |\hat{G}_\sigma(w)| \cdot |w_1| \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1 = \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}}\cdot \int_{|w_1|>B} |\hat{G}_\sigma(w_1)|\cdot |w_1|\,\mathrm{d}w_1 \\
&\quad= \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}}\cdot 2\int_{w_1>B} e^{-2\pi^2\sigma^2 w_1^2}\cdot w_1\,\mathrm{d}w_1 = \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}}\cdot 2\left[-\frac{1}{4\pi^2\sigma^2}e^{-2\pi^2\sigma^2 w_1^2}\right]_{w_1=B}^{\infty} \\
&\quad= \frac{1}{\pi}\left(2\pi\sigma^2\right)^{-\frac{k+1}{2}} \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).
\end{aligned}$$

- By the fact that for $X \sim G$, $\mathbb{E}\|X\| \le \sqrt{d}$,

$$\begin{aligned}
&\iint_{|w_1|> B,\ \|w_{2:k}\|\ge 0} |\hat{G}_\sigma(w)| \cdot \|w_{2:k}\|_2 \,\mathrm{d}w_{2:k}\,\mathrm{d}w_1 \\
&\quad\le \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}}\cdot (2\pi\sigma)^{-1}\sqrt{k-1}\cdot \int_{|w_1|>B} |\hat{G}_\sigma(w_1)|\,\mathrm{d}w_1 \\
&\quad= \sqrt{2\pi(k-1)}\left(2\pi\sigma^2\right)^{-\frac{k}{2}}\cdot \int_{|w_1|>B} |\hat{G}_\sigma(w_1)|\,\mathrm{d}w_1 \\
&\quad\le \sqrt{2\pi(k-1)}\left(2\pi\sigma^2\right)^{-\frac{k}{2}} \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).
\end{aligned}$$

Collecting the terms, since it is assumed that $B \ge 1$ and $\sigma \le 1$,

$$\begin{aligned}
&\int_{\|w\|>B} |\hat{G}_\sigma(w)| \cdot \|w\|_2\,\mathrm{d}w \\
&\quad\le \left(2^{k-1} + 2^{2k-1} + \frac{1}{\pi} + \sqrt{2\pi}\right)\sqrt{k}\left(2\pi\sigma^2\right)^{-\frac{k+1}{2}} B^k \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right) \\
&\quad\le 4^k\sqrt{k}\left(2\pi\sigma^2\right)^{-\frac{k+1}{2}} B^k \exp\left(-\frac{B^2}{2(2\pi\sigma)^{-2}}\right).
\end{aligned}$$

2. By Jensen's inequality and Parseval's theorem, $\int |\hat{f}(w)|\,\mathrm{d}w \le (\int f^2(x)\,\mathrm{d}x)^{1/2}$, so the claim follows from part 1, the nonnegativity of the function $|\hat{f}|$, and the property that $|\hat{f}_\sigma(w)| = |\hat{f}(w)\hat{G}_\sigma(w)| = |\hat{f}(w)|\cdot|\hat{G}_\sigma(w)|$.

□

Proof of Lemma 3.5. Splitting the integral into two terms, for any $\|x\| \le 1$,

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

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

$$\int_{\|z\|\le\delta} \left| f(x) - \int f(x-z)\right| G_\sigma(z)\,\mathrm{d}z \le \int_{\|z\|\le\delta} \omega_f(\delta) G_\sigma(z)\,\mathrm{d}z = \omega_f(\delta),$$

whereas Gaussian concentration gives

$$\int_{\|z\|>\delta} \left| f(x) - \int f(x-z)\right| G_\sigma(z)\,\mathrm{d}z \le 2M\Pr[\|z\| > \delta] \le \omega_f(\delta).$$

□

Proof of Lemma 3.6. By Cauchy-Schwarz,

$$\begin{aligned}
\left\|\langle \mathcal{T}, \Phi.\rangle - \langle \mathcal{T}_{\le B}, \Phi.\rangle\right\|_{L_2(P)}^2 &= \left\|\langle \mathcal{T} - \mathcal{T}_{\le B}, \Phi.\rangle\right\|_{L_2(P)}^2 \\
&= \int \left(\int \mathcal{T}(w)\Phi_x(w)\mathbb{1}[\|w\| > B]\,\mathrm{d}G(w)\right)^2 \mathrm{d}x \\
&\le \int \|\mathcal{T}\|_{\mathcal{H}}^2 \|\Phi_x \mathbb{1}[\|w\| \le B]\|_{\mathcal{H}}^2\,\mathrm{d}x.
\end{aligned}$$

To finish, note by Gaussian concentration that

$$\|\Phi_x \mathbb{1}[\|w\| \le B]\|_{\mathcal{H}}^2 = \mathop{\mathbb{E}}_{\|w\|>B} x^\intercal x \mathbb{1}[w^\intercal x \ge 0] \le \frac{1}{2}\Pr[\|w\| > B] \le \exp\left(-(B-d)^2/(10d^2)\right).$$

□

Proof of Theorem 1.3. Let $\mathcal{T}_0$ denote the exact transport mapping and set $B := \sup_{\tilde{w}} \|\mathcal{T}_0\|_2$, as given by one of the following three lemmas:

- For $f * G_\sigma$ has a nice Fourier transform, then Lemma 3.3 gives $B = \widetilde{\mathcal{O}}(\|f\|_{L_2}\sigma^{-d})$.
- If f is continuous, then the preceding combined with Lemma 3.5 gives $B = \widetilde{\mathcal{O}}(\|f\|_{L_2}\delta^{-d})$ where $\epsilon = \omega_f(\delta)$.
- If $\|f\|_{\mathcal{H}}$ is small, then Lemma 3.6 gives $B = \ln(\|f\|_{\mathcal{H}}/\epsilon)$.

Now let $\mathcal{T}_1$ be the finite-width mapping provided by Lemma 2.1. Inspecting Lemmas 2.1 and 2.3, all upper bounds in Theorem 1.3 by plugging in the above estimates. □