A CLOSER LOOK AT THE APPROXIMATION CAPABILITIES OF NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The *universal approximation theorem*, in one of its most general versions, says that if we consider only continuous activation functions σ, then a standard feedforward neural network with one hidden layer is able to approximate any continuous multivariate function f to any given approximation threshold ε, if and only if σ is non-polynomial. In this paper, we give a direct algebraic proof of the theorem. Furthermore we shall explicitly quantify the number of hidden units required for approximation. Specifically, if $X \subseteq \mathbb{R}^n$ is compact, then a neural network with n input units, m output units, and a single hidden layer with $\binom{n+d}{d}$ hidden units (independent of m and ε), can uniformly approximate any polynomial function $f : X \to \mathbb{R}^m$ whose total degree is at most d for each of its m coordinate functions. In the general case that f is any continuous function, we show there exists some $N \in \mathcal{O}(\varepsilon^{-n})$ (independent of m), such that N hidden units would suffice to approximate f. We also show that this uniform approximation property (UAP) still holds even under seemingly strong conditions imposed on the weights. We highlight several consequences: (i) For any $\delta > 0$, the UAP still holds if we restrict all non-bias weights w in the last layer to satisfy $|w| < \delta$. (ii) There exists some $\lambda > 0$ (depending only on f and σ), such that the UAP still holds if we restrict all non-bias weights w in the first layer to satisfy $|w| > \lambda$. (iii) If the non-bias weights in the first layer are *fixed* and randomly chosen from a suitable range, then the UAP holds with probability 1.

1 INTRODUCTION AND OVERVIEW

A standard (feedforward) neural network with n input units, m output units, and with one or more hidden layers, refers to a computational model $\mathcal{N}$ that can compute a certain class of functions $\rho : \mathbb{R}^n \to \mathbb{R}^m$, where $\rho = \rho_W$ is parametrized by W (called the *weights* of $\mathcal{N}$). Implicitly, the definition of ρ depends on a choice of some fixed function $\sigma : \mathbb{R} \to \mathbb{R}$, called the *activation function* of $\mathcal{N}$. Typically, σ is assumed to be continuous, and historically, the earliest commonly used activation functions were sigmoidal.

A key fundamental result justifying the use of sigmoidal activation functions was due to Cybenko (1989), Hornik et al. (1989), and Funahashi (1989), who independently proved the first version of what is now famously called the *universal approximation theorem*. This first version says that if σ is sigmoidal, then a standard neural network with one hidden layer would be able to uniformly approximate any continuous function $f : X \to \mathbb{R}^m$ whose domain $X \subseteq \mathbb{R}^n$ is compact. Hornik (1991) extended the theorem to the case when σ is any continuous bounded non-constant activation function. Subsequently, Leshno et al. (1993) proved that for the class of continuous activation functions, a standard neural network with one hidden layer is able to uniformly approximate any continuous function $f : X \to \mathbb{R}^m$ on any compact $X \subseteq \mathbb{R}^n$, if and only if σ is non-polynomial.

Although a single hidden layer is sufficient for the uniform approximation property (UAP) to hold, the number of hidden units required could be arbitrarily large. Given a subclass $\mathcal{F}$ of real-valued continuous functions on a compact set $X \subseteq \mathbb{R}^n$, a fixed activation function σ, and some $\varepsilon > 0$, let $N = N(\mathcal{F}, \sigma, \varepsilon)$ be the minimum number of hidden units required for a single-hidden-layer neural network to be able to uniformly approximate *every* $f \in \mathcal{F}$ within an approximation error threshold of ε. If σ is the rectified linear unit (ReLU) $x \mapsto \max(0, x)$, then N is at least $\Omega(\frac{1}{\sqrt{\varepsilon}})$ when $\mathcal{F}$ is the class of C^2 non-linear functions (Yarotsky, 2017), or the class of strongly convex

differentiable functions (Liang & Srikant, 2016); see also (Arora et al., 2018). If σ is any smooth non-polynomial function, then N is at most $\mathcal{O}(\varepsilon^{-n})$ for the class of C^1 functions with bounded Sobolev norm (Mhaskar, 1996); cf. (Pinkus, 1999, Thm. 6.8), (Maiorov & Pinkus, 1999). As a key highlight of this paper, we show that if σ is an *arbitrary* continuous non-polynomial function, then N is at most $\mathcal{O}(\varepsilon^{-n})$ for the entire class of continuous functions. In fact, we give an explicit upper bound for N in terms of ε and the modulus of continuity of f, so better bounds could be obtained for certain subclasses $\mathcal{F}$, which we discuss further in Section 4. Furthermore, even for the wider class $\mathcal{F}$ of all continuous functions $f : X \to \mathbb{R}^m$, the bound is still $\mathcal{O}(\varepsilon^{-n})$, independent of m.

To prove this bound, we shall give a direct algebraic proof of the universal approximation theorem, in its general version as stated by Leshno et al. (1993) (i.e. σ is continuous and non-polynomial). An important advantage of our algebraic approach is that we are able to glean additional information on sufficient conditions that would imply the UAP. Another key highlight we have is that if $\mathcal{F}$ is the subclass of polynomial functions $f : X \to \mathbb{R}^m$ with total degree at most d for each coordinate function, then $\binom{n+d}{d}$ hidden units would suffice. In particular, notice that our bound $N \leq \binom{n+d}{d}$ does not depend on the approximation error threshold ε or the output dimension m.

We shall also show that the UAP holds even under strong conditions on the weights. Given any $\delta > 0$, we can always choose the non-bias weights in the last layer to have small magnitudes no larger than δ. Furthermore, we show that there exists some $\lambda > 0$ (depending only on σ and the function f to be approximated), such that the non-bias weights in the first layer can always be chosen to have magnitudes greater than λ. Even with these seemingly strong restrictions on the weights, we show that the UAP still holds. Thus, our main results can be collectively interpreted as a quantitative refinement of the universal approximation theorem, with extensions to restricted weight values.

Outline: Section 2 covers the preliminaries, including relevant details on arguments involving dense sets. Section 3 gives precise statements of our results, while Section 4 discusses the consequences of our results. Section 5 introduces our algebraic approach and includes most details of the proofs of our results; details omitted from Section 5 can be found in the appendix. Finally, Section 6 concludes our paper with further remarks.

2 Preliminaries

2.1 Notation and Definitions

Let $\mathbb{N}$ be the set of non-negative integers, let $\mathbf{0}_n$ be the zero vector in $\mathbb{R}^n$, and let $\mathrm{Mat}(k, \ell)$ be the vector space of all k-by-ℓ matrices with real entries. For any function $f : \mathbb{R}^n \to \mathbb{R}^m$, let $f^{[t]}$ denote the t-th coordinate function of f (for each $1 \leq t \leq m$). Given $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ and any n-tuple $x = (x_1, \ldots, x_n)$, we write x^α to mean $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. If $x \in \mathbb{R}^n$, then x^α is a real number, while if x is a sequence of variables, then x^α is a *monomial*, i.e. an n-variate polynomial with a single term. Define $\mathcal{W}_N^{n,m} := \{W \in \mathrm{Mat}(n+1, N) \times \mathrm{Mat}(N+1, m)\}$ for each $N \geq 1$, and define $\mathcal{W}^{n,m} = \bigcup_{N \geq 1} \mathcal{W}_N^{n,m}$. If the context is clear, we supress the superscripts n, m in $\mathcal{W}_N^{n,m}$ and $\mathcal{W}^{n,m}$.

Given any $X \subseteq \mathbb{R}^n$, let $\mathcal{C}(X)$ be the vector space of all continuous functions $f : X \to \mathbb{R}$. We use the convention that every $f \in \mathcal{C}(X)$ is a function $f(x_1, \ldots, x_n)$ in terms of the variables $x_1, \ldots, x_n$, unless $n = 1$, in which case f is in terms of a single variable x (or y). We say that f is *non-zero* if f is not identically the zero function on X. Let $\mathcal{P}(X)$ be the subspace of all polynomial functions in $\mathcal{C}(X)$. For each $d \in \mathbb{N}$, let $\mathcal{P}_{\leq d}(X)$ (resp. $\mathcal{P}_d(X)$) be the subspace consisting of all polynomial functions of total degree $\leq d$ (resp. exactly d). More generally, let $\mathcal{C}(X, \mathbb{R}^m)$ be the vector space of all continuous functions $f : X \to \mathbb{R}^m$, and define $\mathcal{P}(X, \mathbb{R}^m)$, $\mathcal{P}_{\leq d}(X, \mathbb{R}^m)$, $\mathcal{P}_d(X, \mathbb{R}^m)$ analogously.

Throughout, we assume that $\sigma \in \mathcal{C}(\mathbb{R})$. For every $W = (W^{(1)}, W^{(2)}) \in \mathcal{W}$, let $\mathbf{w}_j^{(k)}$ be the j-th column vector of $W^{(k)}$, and let $w_{i,j}^{(k)}$ be the (i, j)-th entry of $W^{(k)}$ (for $k = 1, 2$). The index i begins at $i = 0$, while the indices j, k begin at $j = 1$, $k = 1$ respectively. For convenience, let $\widehat{\mathbf{w}}_j^{(k)}$ denote the truncation of $\mathbf{w}_j^{(k)}$ obtained by removing the first entry $w_{0,j}^{(k)}$. Define the function

$\rho_W^\sigma : \mathbb{R}^n \to \mathbb{R}^m$ so that for each $1 \leq j \leq m$, the j-th coordinate function $\rho_W^{\sigma\,[j]}$ is given by the map

$$x \mapsto w_{0,j}^{(2)} + \sum_{i=1}^N w_{i,j}^{(2)} \sigma(\mathbf{w}_i^{(1)} \cdot (1, x)),$$

where "$\cdot$" denotes dot product, and $(1, x)$ denotes a column vector in $\mathbb{R}^{n+1}$ formed by concatenating 1 before x. The class of functions that neural networks $\mathcal{N}$ with one hidden layer can compute is precisely $\{\rho_W^\sigma : W \in \mathcal{W}\}$, where σ is called the *activation function* of $\mathcal{N}$ (or of ρ_W^σ). Functions ρ_W^σ satisfying $W \in \mathcal{W}_N$ correspond to neural networks with N hidden units (in its single hidden layer). Every $w_{i,j}^{(k)}$ is called a *weight* in the k-th layer, where $w_{i,j}^{[k]}$ is called a *bias weight* (resp. *non-bias weight*) if $i = 0$ (resp. $i \neq 0$).

Notice that we do not apply the activation function σ to the output layer. This is consistent with previous approximation results for neural networks. The reason is simple: $\sigma \circ \rho_W^{\sigma\,[j]}$ (restricted to domain $X \subseteq \mathbb{R}^n$) cannot possibly approximate $f : X \to \mathbb{R}$ if there exists some $x_0 \in X$ such that $\sigma(X)$ is bounded away from $f(x_0)$. If instead $f(X)$ is contained in the closure of $\sigma(X)$, then applying σ to $\rho_W^{\sigma\,[j]}$ has essentially the same effect as allowing for bias weights $w_{0,j}^{(2)}$.

Although some authors, e.g. (Leshno et al., 1993), do not explicitly include bias weights in the output layer, the reader should check that if σ is not identically zero, say $\sigma(y_0) \neq 0$, then having a bias weight $w_{0,j}^{(2)} = c$ is equivalent to setting $w_{0,j}^{(2)} = 0$ (i.e. no bias weight in the output layer) and introducing an $(N+1)$-th hidden unit, with corresponding weights $w_{0,N+1}^{(1)} = y_0$, $w_{i,N+1}^{(1)} = 0$ for all $1 \leq i \leq n$, and $w_{N+1,j}^{(2)} = \frac{c}{\sigma(y_0)}$; this means our results also apply to neural networks without bias weights in the output layer (but with one additional hidden unit).

2.2 Arguments involving dense subsets

A key theme in this paper is the use of dense subsets of metric spaces. We shall consider several notions of "dense". First, recall that a *metric* on a set S is any function $\mathfrak{d} : S \times S \to \mathbb{R}$ such that for all $x, y, z \in S$, the following conditions hold:

(i) $\mathfrak{d}(x, y) \geq 0$, with equality holding if and only if $x = y$;
(ii) $\mathfrak{d}(x, y) = \mathfrak{d}(y, x)$;
(iii) $\mathfrak{d}(x, z) \leq \mathfrak{d}(x, y) + \mathfrak{d}(y, z)$.

The set S, together with a metric on S, is called a *metric space*. For example, the usual Euclidean norm for vectors in $\mathbb{R}^n$ gives the *Euclidean metric* $(u, v) \mapsto \|u - v\|_2$, hence $\mathbb{R}^n$ is a metric space. In particular, every pair in $\mathcal{W}_N$ can be identified with a vector in $\mathbb{R}^{(m+n+1)N}$, so $\mathcal{W}_N$, together with the Euclidean metric, is a metric space.

Given a metric space X (with metric $\mathfrak{d}$), and some subset $U \subseteq X$, we say that U is *dense* in X (w.r.t. $\mathfrak{d}$) if for all $\varepsilon > 0$ and all $x \in X$, there exists some $u \in U$ such that $\mathfrak{d}(x, u) < \varepsilon$. Arbitrary unions of dense subsets are dense. If $U \subseteq U' \subseteq X$ and U is dense in X, then U' must also be dense in X.

A basic result in algebraic geometry says that if $p \in \mathcal{P}(\mathbb{R}^n)$ is non-zero, then $\{x \in \mathbb{R}^n : p(x) \neq 0\}$ is a dense subset of $\mathbb{R}^n$ (w.r.t. the Euclidean metric). This subset is in fact an open set in the Zariski topology, hence any finite intersection of such *Zariski-dense open sets* is dense; see (Eisenbud, 1995). More generally, the following is true: Let $p_1, \ldots, p_k \in \mathcal{P}(\mathbb{R}^n)$, and suppose that $X := \{x \in \mathbb{R}^n : p_i(x) = 0 \text{ for all } 1 \leq i \leq k\}$. If $p \in \mathcal{P}(X)$ is non-zero, then $\{x \in X : p(x) \neq 0\}$ is a dense subset of X (w.r.t. the Euclidean metric). In subsequent sections, we shall frequently use these facts.

Let $X \subseteq \mathbb{R}^n$ be a compact set. (Recall that X is *compact* if it is bounded and contains all of its limit points.) For any real-valued function f whose domain contains X, the *uniform norm* of f on X is $\|f\|_{\infty,X} := \sup\{|f(x)| : x \in X\}$. More generally, if $f : X \to \mathbb{R}^m$, then we define $\|f\|_{\infty,X} := \max\{\|f^{[j]}\|_{\infty,X} : 1 \leq j \leq m\}$. The uniform norm of functions on X gives the *uniform metric* $(f, g) \mapsto \|f - g\|_{\infty,X}$, hence $\mathcal{C}(X)$ is a metric space.

2.3 Background on approximation theory

Theorem 2.1 (Stone–Weirstrass theorem)**.** *Let $X \subseteq \mathbb{R}^n$ be compact. For any $f \in \mathcal{C}(X)$, there exists a sequence $\{p_k\}_{k\in\mathbb{N}}$ of polynomial functions in $\mathcal{P}(X)$ such that* $\lim_{k\to\infty} \|f - p_k\|_{\infty,X} = 0$.

Let $X \subseteq \mathbb{R}$ be compact. For all $d \in \mathbb{N}$ and $f \in \mathcal{C}(X)$, define

$$E_d(f) := \inf\{\|f - p\|_{\infty,X} : p \in \mathcal{P}_{\leq d}(X)\}. \tag{1}$$

A central result in approximation theory, due to Chebyshev, says that for fixed d, f, the infimum in (1) is attained by some unique $p^* \in \mathcal{P}_{\leq d}(X)$; see (Rivlin, 1981, Chap. 1). This unique polynomial p^* is called the *best polynomial approximant* to f of degree d.

Given a metric space X with metric $\mathfrak{d}$, and any uniformly continuous function $f : X \to \mathbb{R}$, the *modulus of continuity* of f is a function $\omega_f : [0, \infty] \to [0, \infty]$ defined by

$$\omega_f(\delta) := \sup\{|f(x) - f(y)| : x, y \in X, \mathfrak{d}(x, y) \leq \delta\}.$$

By the Heine–Cantor theorem, any continuous f with a compact domain is uniformly continuous.

Theorem 2.2 (Jackson's theorem; see (Rivlin, 1981, Cor. 1.4.1))**.** *Let $d \geq 1$ be an integer, and let $Y \subseteq \mathbb{R}$ be a closed interval of length $r \geq 0$. Suppose $f \in \mathcal{C}(Y)$, and let p^* be the best polynomial approximant to f of degree d. Then $\|f - p^*\|_{\infty,Y} = E_d(f) \leq 6\omega_f(\frac{r}{2d})$.*

3 Main Results

Throughout this section, let $X \subseteq \mathbb{R}^n$ be a compact set.

Theorem 3.1. *Let $d \geq 1$ be an integer, and let $f \in \mathcal{P}_{\leq d}(X, \mathbb{R}^m)$ (i.e. each coordinate function $f^{[t]}$ has total degree $\leq d$). If $\sigma \in \mathcal{C}(\mathbb{R})\backslash\mathcal{P}_{\leq d-1}(\mathbb{R})$, then for every $\varepsilon > 0$, there exists some $W \in \mathcal{W}_{\binom{n+d}{d}}$ such that $\|f - \rho^\sigma_W\|_{\infty,X} < \varepsilon$. Furthermore, the following holds:*

- *(i) Given any $\lambda > 0$, we can choose this W to satisfy the condition that $|w^{(2)}_{i,j}| < \lambda$ for all non-bias weights $w^{(2)}_{i,j}$ (i.e. $i \neq 0$) in the second layer.*
- *(ii) There exists some $\lambda' > 0$, depending only on f and σ, such that we could choose the weights of W in the first layer to satisfy the condition that $\|\widehat{\mathbf{w}}^{(1)}_j\|_2 > \lambda'$ for all j.*

Theorem 3.2. *Let $f \in \mathcal{C}(X, \mathbb{R}^m)$, and suppose $\sigma \in \mathcal{C}(\mathbb{R})\backslash\mathcal{P}(\mathbb{R})$. Then for every $\varepsilon > 0$, there exists an integer $N \in \mathcal{O}(\varepsilon^{-n})$ (independent of m), and some $W \in \mathcal{W}_N$, such that $\|f - \rho^\sigma_W\|_{\infty,X} < \varepsilon$. In particular, if we let $D := \sup\{\|x - y\|_2 : x, y \in X\}$ be the diameter of X, then we can set $N = \binom{n+d_\varepsilon}{d_\varepsilon}$, where $d_\varepsilon := \min\{d \in \mathbb{Z} : d \geq 1, \omega_{f^{[t]}}(\frac{D}{2d}) < \frac{\varepsilon}{6}$ for all $1 \leq t \leq m\}$. (Note that d_ε is well-defined, since $\lim_{\delta\to 0^+} \omega_{f^{[t]}}(\delta) = 0$ for each t.) Furthermore, we could choose this W to satisfy either (i) or (ii), where (i), (ii) are conditions on W as described in Theorem 3.1.*

Theorem 3.3. *Let $f \in \mathcal{C}(X, \mathbb{R}^m)$, and suppose that $\sigma \in \mathcal{C}(\mathbb{R})\backslash\mathcal{P}(\mathbb{R})$. Then there exists $\lambda > 0$ (which depends only on f and σ) such that for every $\varepsilon > 0$, there exists an integer N (independent of m) such that the following holds:*

> *Let $W \in \mathcal{W}_N$ such that each $\widehat{\mathbf{w}}^{(1)}_j \in \mathbb{R}^n$ (for $1 \leq j \leq N$) is chosen uniformly at random from the set $\{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u}\|_2 > \lambda\}$. Then, with probability 1, there exist choices for the bias weights $w^{(1)}_{0,j}$ (for $1 \leq j \leq N$) in the first layer, and (both bias and non-bias) weights $w^{(2)}_{i,j}$ in the second layer, such that $\|f - \rho^\sigma_W\|_{\infty,X} < \varepsilon$.*

Moreover, $N \in \mathcal{O}(\varepsilon^{-n})$ for general $f \in \mathcal{C}(X, \mathbb{R}^m)$, and we can let $N = \binom{n+d}{d}$ if $f \in \mathcal{P}_{\leq d}(X, \mathbb{R}^m)$.

4 Discussion

The universal approximation theorem (version of Leshno et al. (1993)) is an immediate consequence of Theorem 3.2 and the observation that σ must be non-polynomial for the UAP to hold, which follows from the fact that the uniform closure of $\mathcal{P}_{\leq d}(X)$ is $\mathcal{P}_{\leq d}(X)$ itself, for every integer $d \geq 1$. Alternatively, we could infer the universal approximation theorem by applying the Stone–Weirstrass theorem (Theorem 2.1) to Theorem 3.1.

Given fixed n, m, d, a compact set $X \subseteq \mathbb{R}^n$, and $\sigma \in \mathcal{C}(\mathbb{R})\backslash\mathcal{P}_{\leq d-1}(\mathbb{R})$, Theorem 3.1 says that we could use a *fixed* number N of hidden units (independent of ε) and still be able to approximate any function $f \in \mathcal{P}_{\leq d}(X, \mathbb{R}^m)$ to any desired approximation error threshold ε. Our ε-free bound,

although possibly surprising to some readers, is not the first instance of an ε-free bound: Neural networks with *two* hidden layers of sizes $2n+1$ and $4n+3$ respectively are able to uniformly approximate any $f \in \mathcal{C}(X)$, provided that we use a (somewhat pathological) activation function (Maiorov & Pinkus, 1999); cf. (Pinkus, 1999). Lin et al. (2017) showed that for fixed n, d, and a fixed smooth non-linear σ, there is a fixed N (i.e. ε-free), such that a neural network with N hidden units is able to approximate any $f \in \mathcal{P}_{\leq d}(X)$. An explicit expression for N is not given, but we were able to infer from their constructive proof that $N = 4\binom{n+d+1}{d} - 4$ hidden units are required, over $d-1$ hidden layers (for $d \geq 2$). In comparison, we require less hidden units and a single hidden layer.

Our proof of Theorem 3.2 is an application of Jackson's theorem (Theorem 2.2) to Theorem 3.1, which gives an explicit bound in terms of the values of the modulus of continuity ω_f of the function f to be approximated. The moduli of continuity of several classes of continuous functions have explicit characterizations. For example, given constants $k > 0$ and $0 < \alpha \leq 1$, recall that a continuous function $f : X \to \mathbb{R}$ (for compact $X \subseteq \mathbb{R}^n$) is called *k-Lipschitz* if $|f(x) - f(y)| \leq k\|x-y\|$ for all $x, y \in X$, and it is called α-Hölder if there is some constant c such that $|f(x)-f(y)| \leq c\|x-y\|^\alpha$ for all $x, y \in X$. The modulus of continuity of a k-Lipschitz (resp. α-Hölder) continuous function f is $\omega_f(t) = kt$ (resp. $\omega_f(t) = ct^\alpha$), hence Theorem 3.2 implies the following corollary.

Corollary 4.1. *Suppse $\sigma \in \mathcal{C}(\mathbb{R}) \backslash \mathcal{P}(\mathbb{R})$.*

(i) If $f : [0,1]^n \to \mathbb{R}$ is k-Lipschitz continuous, then for every $\varepsilon > 0$, there exists some $W \in \mathcal{W}_N$ that satisfies $\|f - \rho_W^\sigma\|_{\infty,X} < \varepsilon$, where $N = \binom{n+\lceil \frac{3k}{\varepsilon} \rceil}{n}$.

(ii) If $f : [0,1]^n \to \mathbb{R}$ is α-Hölder continuous, then there is a constant k such that for every $\varepsilon > 0$, there exists some $W \in \mathcal{W}_N$ that satisfies $\|f - \rho_W^\sigma\|_{\infty,X} < \varepsilon$, where $N = \binom{n+d}{d}$, and $d = \lceil \frac{1}{2}(\frac{k}{\varepsilon})^{1/\alpha} \rceil$.

An interesting consequence of Theorem 3.3 is the following: The freezing of lower layers of a neural network, even in the extreme case that all frozen layers are randomly initialized and the last layer is the only "non-frozen" layer, does not necessarily reduce the representability of the resulting model. Specifically, in the single-hidden-layer case, we have shown that if the non-bias weights in the first layer are *fixed* and randomly chosen from some suitable fixed range, then the UAP holds with probability 1, provided that there are sufficiently many hidden units. Of course, this representability does not reveal anything about the learnability of such a model. In practice, layers are already pre-trained before being frozen. It would be interesting to understand quantitatively the difference between having pre-trained frozen layers and having randomly initialized frozen layers.

Theorem 3.3 can be viewed as a result on random features, which were formally studied in relation to kernel methods (Rahimi & Recht, 2007). In the case of ReLU activation functions, Sun et al. (2019) proved an analog of Theorem 3.3 for the approximation of functions in a reproducing kernel Hilbert space; cf. (Rahimi & Recht, 2008). For a good discussion on the role of random features in the representability of neural networks, see (Yehudai & Shamir, 2019).

The UAP is also studied in other contexts, most notably in relation to the depth and width of neural networks. Lu et al. (2017) proved the UAP for neural networks with hidden layers of bounded width, under the assumption that ReLU is used as the activation function. Soon after, Hanin (2017) strengthened the bounded-width UAP result by considering the approximation of continuous convex functions. Recently, the role of depth in the expressive power of neural networks has gathered much interest (Delalleau & Bengio, 2011; Eldan & Shamir, 2016; Mhaskar et al., 2017; Montúfar et al., 2014; Telgarsky, 2016). We do not address depth in this paper, but we believe it is possible that our results can be applied iteratively to deeper neural networks, perhaps in particular for the approximation of compositional functions; cf. (Mhaskar et al., 2017).

5 An Algebraic Approach for Proving UAP

We begin with a "warm-up" result. Subsequent results, even if they seem complicated, are actually multivariate extensions of this "warm-up" result, using very similar ideas.

Theorem 5.1. *Let $p(x)$ be a real polynomial of degree d with non-zero coefficients, and let $a_1, \ldots, a_{d+1}$ be real numbers. For each $1 \leq j \leq d+1$, define $f_j : \mathbb{R} \to \mathbb{R}$ by $x \mapsto p(a_j x)$. Then $f_1, \ldots, f_{d+1}$ are linearly independent if and only if $a_1, \ldots, a_{d+1}$ are distinct.*

Proof. For each $0 \leq i, k \leq d$ and each $1 \leq j \leq d+1$, let $f_j^{(i)}$ (resp. $p^{(i)}$) be the i-th derivative of f_j (resp. p), and let $\alpha_k^{(i)}$ be the coefficient of x^k in $p^{(i)}(x)$. Recall that the *Wronskian* of $(f_1, \ldots, f_{d+1})$ is defined to be the determinant of the matrix $M(x) := [f_j^{(i-1)}(x)]_{1 \leq i,j \leq d+1}$. Since $f_1, \ldots, f_{d+1}$ are polynomial functions, it follows that $(f_1, \ldots, f_{d+1})$ is a sequence of linearly independent functions if and only if its Wronskian is not the zero function (LeVeque, 1956, Thm. 4.7(a)). Clearly, if $a_i = a_j$ for $i \neq j$, then $\det M(x)$ is identically zero. Thus, it suffices to show that if $a_1, \ldots, a_{d+1}$ are distinct, then the evaluation $\det M(1)$ of this Wronskian at $x = 1$ gives a non-zero value.

Now, the (i, j)-th entry of $M(1)$ equals $a_j^{i-1} p^{(i-1)}(a_j)$, so $M(1) = M'M''$, where M' is an upper triangular matrix whose (i, j)-th entry equals $\alpha_{j-i}^{(i-1)}$, and $M'' = [a_j^{i-1}]_{1 \leq i,j \leq d+1}$ is the transpose of a Vandermonde matrix, whose determinant is

$$\det(M'') = \prod_{1 \leq i < j \leq d+1} (a_j - a_i).$$

Note that the k-th diagonal entry of M' is $\alpha_0^{(k-1)} = (k-1)!\alpha_{k-1}^{(0)}$, which is non-zero by assumption, so $\det(M') \neq 0$. Thus, if $a_1, \ldots, a_{d+1}$ are distinct, then $\det M(1) = \det(M') \det(M'') \neq 0$. $\square$

Definition 5.2. Given $N \geq 1$, $W \in \mathcal{W}_N^{n,m}$, $x_0 \in \mathbb{R}^n$, and any function $g : \mathbb{R} \to \mathbb{R}$, let $\mathcal{F}_{g,x_0}(W)$ denote the sequence of functions $(f_1, \ldots, f_N)$, such that each $f_j : \mathbb{R}^n \to \mathbb{R}$ is defined by the map $x \mapsto g(\widehat{\mathbf{w}}_j^{(1)} \cdot (x - x_0))$. Also, define the set

$${}^g\mathcal{W}_{n,N;x_0}^{\text{ind}} := \{W \in \mathcal{W}_N^{n,m} : \mathcal{F}_{g,x_0}(W) \text{ is linearly independent}\}.$$

Note that the value of m is irrelevant for defining ${}^g\mathcal{W}_{n,N;x_0}^{\text{ind}}$.

Remark 5.3. Given $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, consider the ring automorphism $\varphi : \mathcal{P}(\mathbb{R}^n) \to \mathcal{P}(\mathbb{R}^n)$ induced by $x_i \mapsto x_i - a_i$ for all $1 \leq i \leq n$. For any $f_1, \ldots, f_k \in \mathcal{P}(\mathbb{R}^n)$ and scalars $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, note that $\alpha_1 f_1 + \cdots + \alpha_k f_k = 0$ if and only if $\alpha_1 \varphi(f_1) + \cdots + \alpha_k \varphi(f_k) = 0$, thus linear independence is preserved under φ. Consequently, if the function g in Definition 5.2 is polynomial, then ${}^g\mathcal{W}_{n,N;x_0}^{\text{ind}} = {}^g\mathcal{W}_{n,N;\mathbf{0}_n}^{\text{ind}}$ for all $x_0 \in \mathbb{R}^n$.

Corollary 5.4. *Let m be arbitrary. If $p \in \mathcal{P}_d(\mathbb{R})$ has non-zero coefficients, then ${}^p\mathcal{W}_{1,d+1;0}^{\text{ind}}$ is a dense subset of $\mathcal{W}_{d+1}^{1,m}$ (in the Euclidean metric).*

Proof. For all $1 \leq j < j' \leq N$, let $\mathcal{A}_{j,j'} := \{W \in \mathcal{W}_{d+1}^{1,m} : w_{1,j'}^{(1)} - w_{1,j}^{(1)} \neq 0\}$, and note that $\mathcal{A}_{j,j'}$ is dense in $\mathcal{W}_{d+1}^{1,m}$. So by Theorem 5.1, ${}^p\mathcal{W}_{1,d+1;0}^{\text{ind}} = \bigcap_{1 \leq j < j' \leq N} \mathcal{A}_{j,j'}$ is also dense in $\mathcal{W}_{d+1}^{1,m}$. $\square$

As we have seen in the proof of Theorem 5.1, Vandermonde matrices play an important role. To extend this theorem (and Corollary 5.4) to the multivariate case, we need a generalization of the Vandermonde matrix as described in (D'Andrea & Tabera, 2009). (Other generalizations of the Vandermonde matrix exist in the literature.) First, define the sets

$$\Lambda_{\leq d}^n := \{(\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n : \alpha_1 + \cdots + \alpha_n \leq d\};$$
$$\mathcal{M}_{\leq d}^n := \{(x \mapsto x^\alpha) \in \mathcal{P}(\mathbb{R}^n) : \alpha \in \Lambda_{\leq d}^n\}.$$

It is easy to show that $|\Lambda_{\leq d}^n| = \binom{n+d}{d}$, and that the set $\mathcal{M}_{\leq d}^n$ of monomial functions forms a basis for $\mathcal{P}_{\leq d}(\mathbb{R}^n)$. Sort the n-tuples in $\Lambda_{\leq d}^n$ in colexicographic order, i.e. $(\alpha_1, \ldots, \alpha_n) < (\alpha'_1, \ldots, \alpha'_n)$ if and only if $\alpha_i < \alpha'_i$ for the largest index i such that $\alpha_i \neq \alpha'_i$. Let $\lambda_1 < \cdots < \lambda_{\binom{n+d}{d}}$ denote all the $\binom{n+d}{d}$ n-tuples in $\Lambda_{\leq d}^n$ after sorting. Analogously, let $q_1, \ldots q_{\binom{n+d}{d}}$ be all the monomial functions in $\mathcal{M}_{\leq d}^n$ in this order, i.e. each $q_k : \mathbb{R}^n \to \mathbb{R}$ is given by the map $x \mapsto x^{\lambda_k}$. Given any sequence $(v_1, \ldots, v_{\binom{n+d}{d}})$ of vectors in $\mathbb{R}^n$, the *generalized Vandermonde matrix* associated to it is

$$Q = Q[v_1, \ldots, v_{\binom{n+d}{d}}] := [q_i(v_j)]_{1 \leq i,j \leq \binom{n+d}{d}}. \tag{2}$$

Definition 5.5. Given any $W \in \mathcal{W}_{\binom{n+d}{d}}^{n,m}$, we define the *non-bias Vandermonde matrix* of W to be the generalized Vandermonde matrix $Q[W] := [q_i(\widehat{\mathbf{w}}_j^{(1)})]_{1 \leq i,j \leq \binom{n+d}{d}}$ associated to $(\widehat{\mathbf{w}}_1^{(1)}, \ldots, \widehat{\mathbf{w}}_{\binom{n+d}{d}}^{(1)})$.

Theorem 5.6. *Let m be arbitrary, let $p \in \mathcal{P}_d(\mathbb{R}^n)$, and suppose that p has all-non-zero coefficients. Also, suppose that $p_1, \ldots, p_k \in \mathcal{P}(\mathcal{W}^{n,m}_{\binom{n+d}{d}})$ are fixed polynomial functions on the non-bias weights of the first layer. Define the following sets:*

$$\mathcal{U} := \{W \in \mathcal{W}^{n,m}_{\binom{n+d}{d}} : p_i(W) = 0 \textit{ for all } 1 \leq i \leq k\};$$
$${}^{p}\mathcal{U}^{\mathrm{ind}} := \{W \in \mathcal{U} : \mathcal{F}_{p,\mathbf{0}_n}(W) \textit{ is linearly independent}\}.$$

If there exists $W \in \mathcal{U}$ such that the non-bias Vandermonde matrix of W is non-singular, then ${}^{p}\mathcal{U}^{\mathrm{ind}}$ is dense in $\mathcal{U}$ (w.r.t. the Euclidean metric).

Proof. We essentially extend the ideas in the proofs of Theorem 5.1 and Corollary 5.4, using the notion of generalized Wronskians; see Appendix A for details. □

Corollary 5.7. *Let m be arbitrary. If $p \in \mathcal{P}(\mathbb{R})$ is a fixed polynomial function of degree d with non-zero coefficients, then ${}^{p}\mathcal{W}^{\mathrm{ind}}_{n,\binom{n+d}{d}}$ is a dense subset of $\mathcal{W}^{n,m}_{\binom{n+d}{d}}$.*

Proof. Let $N := \binom{n+d}{d}$. By Theorem 5.6, it suffices to show that there exists some $W \in \mathcal{W}^{n,m}_N$ such that the non-bias Vandermonde matrix of W is non-singular. Consider $W \in \mathcal{W}^{n,m}_N$ such that $w^{(1)}_{i,j} = (w^{(1)}_{1,j})^{(d+1)^i}$. Recall that the monomials in $\mathcal{M}^n_{\leq d}$ are arranged in colexicographic order, i.e.

$$1, x_1, x_1^2, \ldots, x_1^d, x_2, x_1x_2, x_1^2x_2, \ldots, x_2^2, x_1x_2^2, \ldots, x_n^d.$$

Thus, there are fixed integers $0 = \beta_1 < \beta_2 < \cdots < \beta_N$, such that the (i,j)-th entry of $Q[W]$ is $(w^{(1)}_{1,j})^{\beta_i}$. Such matrices are well-studied in algebraic combinatorics, and the determinant of $Q[W]$ is a *Schur polynomial*; see (Stanley, 1999). In particular, if we choose positive pairwise distinct values for $w^{(1)}_{1,j}$ (for $1 \leq j \leq N$), then $Q[W]$ is non-singular, since a Schur polynomial can be expressed as a (non-negative) sum of certain monomials; see (Stanley, 1999, Sec. 7.10) for details. □

Remark 5.8. The proof of Corollary 5.7 still holds even if we restrict every non-bias weight $w^{(1)}_{i,j}$ in the first layer to satisfy $|w^{(1)}_{i,j}| < \lambda$ for some fixed constant $\lambda > 0$.

Lemma 5.9. *Let $d \geq 1$ be an integer, and suppose $\sigma \in \mathcal{C}(\mathbb{R})$. For every $r > 0$, let $Y_r \subseteq \mathbb{R}$ be a closed interval of length r, such that $Y_r \subseteq Y_{r'}$ whenever $r \leq r'$, and let σ_r be the best polynomial approximant to $\sigma|_{Y_r}$ (i.e. σ restricted to domain Y_r) of degree d. Then*

$$\lim_{r\to\infty} \frac{\|\sigma_r - \sigma\|_{\infty,Y_r}}{r} = 0.$$

Proof. Theorem 2.2 gives $\|\sigma_r - \sigma\|_{\infty,Y_r} = E_d(\sigma|_{Y_r}) \leq 6\omega_{\sigma|_{Y_r}}(\frac{r}{2d})$. Recall that any modulus of continuity ω_f is subadditive (i.e. $f(x+y) \leq f(x) + f(y)$ for all x, y); see (Rivlin, 1981, Chap. 1). Thus for fixed d, we have $\omega_{\sigma|_{Y_r}}(\frac{r}{2d}) \in \mathcal{O}(r)$, which implies $\left(r \mapsto \frac{\|\sigma_r - \sigma\|_{\infty,Y_r}}{r}\right) \in o(r)$. □

Outline of strategy for proving Theorem 3.1. The first crucial insight is that $\mathcal{P}_{\leq d}(\mathbb{R}^n)$, as a real vector space, has dimension $\binom{n+d}{d}$. Our strategy is to consider $N = \binom{n+d}{d}$ hidden units. Every hidden unit represents a continuous function $g_j : X \to \mathbb{R}$ determined by its weights W and the activation function σ. If $g_1, \ldots, g_N$ can be well-approximated (on X) by linearly independent polynomial functions in $\mathcal{P}_{\leq d}(\mathbb{R}^n)$, then we can choose suitable linear combinations of these N functions to approximate all coordinate functions $f^{[t]}$ (independent of how large m is). To approximate each g_j, we consider a suitable sequence $\{\sigma_{\lambda_k}\}_{k=1}^{\infty}$ of degree d polynomial approximations to σ, so that g_j is approximated by a sequence of degree d polynomial functions $\{\widehat{g}^W_{j,k}\}_{k=1}^{\infty}$. We shall also vary W concurrently with k, so that $\|\widehat{\mathbf{w}}^{(1)}_j\|_2$ increases together with k. By Corollary 5.7, the weights can always be chosen so that $\widehat{g}^W_{1,k}, \ldots, \widehat{g}^W_{N,k}$ are linearly independent.

The second crucial insight is that every function in $\mathcal{P}_{\leq d}(\mathbb{R}^n)$ can be identified geometrically as a point in Euclidean $\binom{n+d}{d}$-space. We shall choose the bias weights so that $\widehat{g}^W_{1,k}, \ldots, \widehat{g}^W_{N,k}$ correspond to points on a hyperplane, and we shall consider the barycentric coordinates of the projections of both $f^{[t]}$ and the constant function onto this hyperplane, with respect to $\widehat{g}^W_{1,k}, \ldots, \widehat{g}^W_{N,k}$. As the

values of k and $\|\widehat{\mathbf{w}}_j^{(1)}\|_2$ increase, both projection points have barycentric coordinates that approach $(\frac{1}{N}, \dots, \frac{1}{N})$, and their difference approach $\mathbf{0}$. This last observation, in particular, when combined with Lemma 5.9, is a key reason why the minimum number N of hidden units required for the UAP to hold is independent of the approximation error threshold ε.

Proof of Theorem 3.1. Fix some $\varepsilon > 0$, and for brevity, let $N = \binom{n+d}{d}$. The case that f is a constant function is trivial, so assume f is non-constant. Fix a point $x_0 \in X$, and define $f_{\mathbf{0}} \in \mathcal{C}(X, \mathbb{R}^m)$ by $f_{\mathbf{0}}^{[t]} := f^{[t]} - f^{[t]}(x_0)$ for all $1 \leq t \leq m$. For every $r \geq 0$, let $Y_r \subseteq \mathbb{R}$ be a closed interval of length r, such that $\sigma|_{Y_r}$ is not identically zero; Y_r is well-defined, since $\sigma \notin \mathcal{P}_0(\mathbb{R})$ by assumption. Without loss of generality, assume that $Y_{r'} \subseteq Y_r$ whenever $r' \leq r$. Also, let σ_r be the best polynomial approximant to $\sigma|_{Y_r}$ of degree d. Note that for any $W \in \mathcal{W}_N^{n,m}$, we could choose $\sigma' \in \mathcal{C}(X)$ to be arbitrarily close to σ in the uniform metric, such that $\|\rho_W^{\sigma} - \rho_W^{\sigma'}\|_{\infty,X}$ is arbitrarily small. Since $\sigma \in \mathcal{C}(\mathbb{R}) \backslash \mathcal{P}_{\leq d-1}(\mathbb{R})$ by assumption, we may perturb σ if necessary, and assume without loss of generality that $\sigma_r \in \mathcal{P}_d(Y_r)$, and that σ_r has all-non-zero coefficients.

For every $h \in \mathcal{P}_{\leq d}(\mathbb{R}^n)$, let $\nu(h) \in \mathbb{R}^N$ be the vector of coefficients with respect to the basis $\{q_1(x - x_0), \dots, q_N(x - x_0)\}$ (i.e. if $\nu(h) = (\nu_1, \dots, \nu_N)$, then $h(x) = \sum_{1 \leq i \leq N} \nu_i q_i(x - x_0)$), and let $\widehat{\nu}(h) \in \mathbb{R}^{N-1}$ be the truncation of $\nu(h)$ by removing the first coordinate. Note that $q_1(x)$ is the constant monomial, so this first coordinate is the coefficient of the constant term. Define

$$c_f := \max\{\|\widehat{\nu}(f^{[t]})\|_2 : 1 \leq t \leq m\} > 0.$$

In particular, $c_f \neq 0$, since f is non-constant. Next, define $C := 8N(N-1)c_f > 0$.

For every $r > 0$, let $\mathcal{W}_r' := \{W \in \mathcal{W}_N^{n,m} : \|\widehat{\mathbf{w}}_j^{(1)}\|_2 < r \text{ for all } 1 \leq j \leq N\}$. Let $\{\lambda_k\}_{k \in \mathbb{N}}$ be a divergent strictly increasing sequence of positive real numbers, such that $\|\sigma_{\lambda_k} - \sigma\|_{\infty, Y_{\lambda_k}} < \frac{\varepsilon \lambda_k}{C}$ for all $k \in \mathbb{N}$. This sequence exists, since $\lim_{r \to \infty} \frac{\|\sigma_r - \sigma\|_{\infty, Y_r}}{r} = 0$ by Lemma 5.9. For each $k \in \mathbb{N}$, let y_k be any element of Y_{λ_k} such that $\sigma(y_k) = \sigma_{\lambda_k}(y_k)$; clearly y_k exists, since if instead $\sigma_{\lambda_k}(y) \neq \sigma(y)$ for all $y \in Y_{\lambda_k}$, then σ_{λ_k} cannot possibly be the best polynomial approximant to $\sigma|_{Y_{\lambda_k}}$. By perturbing $\{\lambda_k\}_{k \in \mathbb{N}}$ if necessary, we shall assume without loss of generality that $\sigma(\lambda_k) \neq 0$ for all $k \in \mathbb{N}$. Also, for every $k \in \mathbb{N}$, define

$$\lambda_k' := \sup\Big\{r > 0 : \{y_k + \widehat{\mathbf{w}}_j^{(1)} \cdot (x - x_0) \in \mathbb{R} : x \in X, W \in \mathcal{W}_r'\} \subseteq Y_{\lambda_k} \text{ for all } 1 \leq j \leq N\Big\}.$$

Each λ_k' is well-defined, since the compactness of X implies that $\|x - x_0\|_2$ is bounded. Let $r_X(x_0) := \sup\{\|x - x_0\|_2 : x \in X\}$, and note that $2\lambda_k' \cdot r_X(x_0) \leq \lambda_k$ by the definition of Y_{λ_k}. Also, for every $k \in \mathbb{N}$, $W \in \mathcal{W}_{\lambda_k'}'$ and $1 \leq j \leq N$, define the functions $g_{j,k}^W, \widehat{g}_{j,k}^W$ in $\mathcal{C}(X)$ by $x \mapsto \sigma(\mathbf{w}_j^{(1)} \cdot (1, x))$ and $x \mapsto \sigma_{\lambda_k}(\mathbf{w}_j^{(1)} \cdot (1, x))$ respectively.

Next, consider an arbitrary $k \in \mathbb{N}$, define the "shifted" function $\sigma_{\lambda_k}' : Y_{\lambda_k} \to \mathbb{R}$ by $y \mapsto \sigma_{\lambda_k}(y + y_k)$, and suppose that $W \in {}^{\sigma_{\lambda_k}'}\mathcal{W}_{n,N;x_0}^{\text{ind}} \cap \mathcal{W}_{\lambda_k'}'$. Note that in the definition of ${}^{\sigma_{\lambda_k}'}\mathcal{W}_{n,N;x_0}^{\text{ind}} \cap \mathcal{W}_{\lambda_k'}'$, we do not impose any restrictions on the bias weights. Thus, given any such W, we could choose the bias weights of $W^{(1)}$ to be $w_{j,0}^{(1)} = y_k - \widehat{\mathbf{w}}_j^{(1)} \cdot x_0$ for all $1 \leq j \leq N$. This implies that each $\widehat{g}_{j,k}^W$ represents the map $x \mapsto \sigma_{\lambda_k}(\widehat{\mathbf{w}}_j^{(1)} \cdot (x - x_0) + y_k)$, hence $\widehat{g}_{j,k}^W(x_0) = \sigma_{\lambda_k}(y_k) = \sigma(y_k)$. Consequently, by the definition of Y_{λ_k} and $\mathcal{W}_{\lambda_k'}'$, we infer that

$$\|g_{j,k}^W - \widehat{g}_{j,k}^W\|_{\infty,X} < \frac{\varepsilon \lambda_k}{C}. \tag{3}$$

By Corollary 5.7 and Remark 5.3, ${}^{\sigma_{\lambda_k}'}\mathcal{W}_{n,N;x_0}^{\text{ind}} \cap \mathcal{W}_{\lambda_k'}'$ is dense in $\mathcal{W}_{\lambda_k'}'$, so such a W exists (with its bias weights given as above); we can furthermore choose W so that $\frac{1}{2}\lambda_k' < \|\widehat{\mathbf{w}}_j^{(1)}\|_2 < \lambda_k'$ for all $1 \leq j \leq N$. By the definition of ${}^{\sigma_{\lambda_k}'}\mathcal{W}_{n,N;x_0}^{\text{ind}}$, we infer that $\{\widehat{g}_{1,k}^W, \dots, \widehat{g}_{N,k}^W\}$ is linearly independent and hence spans $\mathcal{P}_{\leq d}(X)$. Thus, for every $1 \leq t \leq m$, there exist $a_{1,k}^{[t]}, \dots, a_{N,k}^{[t]} \in \mathbb{R}$, which are uniquely determined once k is fixed, such that $f_{\mathbf{0}}^{[t]} = a_{1,k}^{[t]} \widehat{g}_{1,k}^W + \dots + a_{N,k}^{[t]} \widehat{g}_{N,k}^W$. Evaluating both

sides of this equation at $x = x_0$, we then get

$$a_{1,k}^{[t]} + \cdots + a_{N,k}^{[t]} = 0. \tag{4}$$

For each $\ell \in \mathbb{R}$, define the hyperplane $\mathcal{H}_\ell := \{(u_1, \ldots, u_N) \in \mathbb{R}^N : u_1 = \ell\}$. Recall that $q_1(x)$ is the constant monomial, so the first coordinate of each $\nu(\widehat{g}_{j,k}^W)$ equals $\sigma(y_k)$, which implies that $\nu(\widehat{g}_{1,k}^W), \ldots, \nu(\widehat{g}_{N,k}^W)$ are N points on $\mathcal{H}_{\sigma(y_k)}$. Since $\nu(f_{\mathbf{0}}^{[t]}) = a_{1,k}^{[t]}\nu(\widehat{g}_{1,k}^W) + \cdots + a_{N,k}^{[t]}\nu(\widehat{g}_{N,k}^W)$, it then follows from (4) that $\nu(f_{\mathbf{0}}^{[t]})$ lies on $\mathcal{H}_0$. Note also that $\widehat{\nu}(f_{\mathbf{0}}^{[t]}) = \widehat{\nu}(f^{[t]})$ for all $1 \le t \le m$.

Observe that $\lambda_k \to \infty$ (and hence $\lambda_k' \to \infty$) as $k \to \infty$. Thus, when k is sufficiently large, we can choose $W \in {}^{\sigma'_{\lambda_k}}\mathcal{W}_{n,N;x_0}^{\text{ind}} \cap \mathcal{W}_{\lambda_k'}'$, with each $\|\widehat{\nu}(\widehat{g}_{j,k}^W)\|_2$ sufficiently large, such that the convex hull of $\{\widehat{\nu}(\widehat{g}_{1,k}^W), \ldots, \widehat{\nu}(\widehat{g}_{N,k}^W)\}$ contains the ball $B_r := \{u \in \mathbb{R}^{N-1} : \|u\|_2 \le r\}$ (centered at $\mathbf{0}_{N-1}$) of any desired large radius r; we used here the fact that W could be chosen so that $\frac{1}{2}\lambda_k' < \|\widehat{\mathbf{w}}_j^{(1)}\|_2 < \lambda_k'$ for all $1 \le j \le N$. In particular, for sufficiently large k, we can choose W so that this convex hull contains both points $\widehat{\nu}(f^{[t]})$ and $\mathbf{0}_{N-1}$. This implies that for each $1 \le t \le m$, we have $\widehat{\nu}(f^{[t]}) = b_{1,k}^{[t]}\widehat{\nu}(\widehat{g}_{1,k}^W) + \cdots + b_{N,k}^{[t]}\widehat{\nu}(\widehat{g}_{N,k}^W)$ and $\mathbf{0}_{N-1} = b_{1,k}'\widehat{\nu}(\widehat{g}_{1,k}^W) + \cdots + b_{N,k}'\widehat{\nu}(\widehat{g}_{N,k}^W)$ for some $b_{1,k}^{[t]}, \ldots, b_{N,k}^{[t]}, b_{1,k}', \ldots, b_{N,k}' \ge 0$ satisfying $b_{1,k}^{[t]} + \cdots + b_{N,k}^{[t]} = b_{1,k}' + \cdots + b_{N,k}' = 1$.

Note that $(b_{1,k}^{[t]}, \ldots, b_{N,k}^{[t]})$ and $(b_{1,k}', \ldots, b_{N,k}')$ are barycentric coordinate vectors of the points $\widehat{\nu}(f^{[t]})$ and $\mathbf{0}_{N-1}$ respectively, w.r.t. $\{\widehat{\nu}(\widehat{g}_{1,k}^W), \ldots, \widehat{\nu}(\widehat{g}_{N,k}^W)\}$. Both points are in the ball $B_{\|\widehat{\nu}(f^{[t]})\|_2}$. Since W is chosen so that $\frac{1}{2}\lambda_k' < \|\widehat{\mathbf{w}}_j^{(1)}\|_2 < \lambda_k'$ for all $1 \le j \le N$, it follows that as $\lambda_k' \to \infty$, the barycentric coordinate vector of every point in $B_{\|\widehat{\nu}(f^{[t]})\|_2}$ (w.r.t. $\{\widehat{\nu}(\widehat{g}_{1,k}^W), \ldots, \widehat{\nu}(\widehat{g}_{N,k}^W)\}$) would approach the barycenter with barycentric coordinate vector $(\frac{1}{N}, \ldots, \frac{1}{N})$. Since $\lambda_k \to \infty$ as $k \to \infty$, we infer that for every $\delta > 0$, there exists some sufficiently large k, such that $b_{j,k}^{[t]}, b_{j,k}' \in (\frac{1}{N} - \delta, \frac{1}{N} + \delta)$ for all $1 \le j \le N$, $1 \le t \le m$.

Note that

$$\nu(f_{\mathbf{0}}^{[t]} + \sigma(y_k)) = b_{1,k}^{[t]}\nu(\widehat{g}_{1,k}^W) + \cdots + b_{N,k}^{[t]}\nu(\widehat{g}_{N,k}^W);$$
$$(\mathbf{0}_{N-1}, \sigma(y_k)) = b_{1,k}'\nu(\widehat{g}_{1,k}^W) + \cdots + b_{N,k}'\nu(\widehat{g}_{N,k}^W);$$

hence

$$f_{\mathbf{0}}^{[t]} = \big(b_{1,k}^{[t]} - b_{1,k}'\big)\widehat{g}_{1,k}^W + \cdots + \big(b_{N,k}^{[t]} - b_{N,k}'\big)\widehat{g}_{N,k}^W. \tag{5}$$

Since $a_{1,k}^{[t]}, \ldots, a_{N,k}^{[t]}$ are unique (for fixed k), we infer that $a_{j,k}^{[t]} = b_{j,k}^{[t]} - b_{j,k}'$ for each $1 \le j \le N$. Thus, for sufficiently large k, it follows from $b_{j,k}^{[t]}, b_{j,k}' \ge 0$ and $b_{j,k}^{[t]}, b_{j,k}' \in (\frac{1}{N} - \delta, \frac{1}{N} + \delta)$ that

$$a_{j,k}^{[t]} \ge (\tfrac{1}{N} - \delta) - (\tfrac{1}{N} + \delta) \ge -2\delta. \tag{6}$$

Now, δ is an upper bound of the normalized distance of each barycentric coordinate of the two points $\widehat{\nu}(f^{[t]})$ and $\mathbf{0}_{N-1}$ (contained in $B_{\|\widehat{\nu}(f^{[t]})\|_2}$), which satisfies $\delta \le \frac{\|\widehat{\nu}(f^{[t]})\|_2}{r}$ for any $r > 0$ such that B_r is contained in the convex hull of $\{\widehat{\nu}(\widehat{g}_{1,k}^W), \ldots, \widehat{\nu}(\widehat{g}_{N,k}^W)\}$. Recall the well-known fact that the inradius of a regular $(N-1)$-simplex circumscribed by a unit $(N-1)$-sphere is $\frac{1}{N-1}$ (Klamkin & Tsintsifas, 1979). Since $\frac{1}{2}\lambda_k' < \|\widehat{\mathbf{w}}_j^{(1)}\|_2 < \lambda_k'$ for all $1 \le j \le N$, we could set $r = \frac{\lambda_k'}{2(N-1)}$. Consequently, $\delta \le \frac{2(N-1)\cdot\|\widehat{\nu}(f^{[t]})\|_2}{\lambda_k'}$, so it follows from (4) and (6) that for sufficiently large k,

$$|a_{1,k}^{[t]}| + \cdots + |a_{N,k}^{[t]}| \le a_{1,k}^{[t]} + \cdots + a_{N,k}^{[t]} + 4N\delta \le \frac{8N(N-1)\cdot\|\widehat{\nu}(f^{[t]})\|_2}{\lambda_k'} \le \frac{C}{\lambda_k'} \tag{7}$$

Now, for this sufficiently large k, define $g : X \to \mathbb{R}^m$ by $g^{[t]} = a_{1,k}^{[t]}g_{1,k}^W + \cdots + a_{N,k}^{[t]}g_{N,k}^W$ for all $1 \le t \le m$. Using (3) and (7), it follows that

$$\begin{aligned}\|f_{\mathbf{0}}^{[t]} - g^{[t]}\|_{\infty,X} &= \|a_{1,k}^{[t]}(g_{1,k}^W - \widehat{g}_{1,k}^W) + \cdots + a_{N,k}^{[t]}(g_{N,k}^W - \widehat{g}_{N,k}^W)\|_{\infty,X} \\ &\le |a_{1,k}^{[t]}| \cdot \|g_{1,k}^W - \widehat{g}_{1,k}^W\|_{\infty,X} + \cdots + |a_{N,k}^{[t]}| \cdot \|g_{N,k}^W - \widehat{g}_{N,k}^W\|_{\infty,X} \\ &< \varepsilon.\end{aligned}$$

Finally, for all $1 \leq t \leq m$, let $w_{j,t}^{(2)} = a_{j,k}^{[t]}$ for each $1 \leq j \leq N$, and let $w_{0,t}^{(2)} = f^{[t]}(x_0)$. This gives $\rho_W^{\sigma\,[t]} = g + f^{[t]}(x_0)$. Therefore, the identity $f^{[t]} = f_{\mathbf{0}}^{[t]} + f^{[t]}(x_0)$ implies $\|f - \rho_W^{\sigma}\|_{\infty,X} < \varepsilon$.

Notice that for all $\delta > 0$, we showed in (6) that there is a sufficiently large k such that $a_{j,k}^{[t]} \geq -2\delta$. A symmetric argument yields $a_{j,k}^{[t]} \leq 2\delta$. Thus, for all $\lambda > 0$, we can choose W so that all non-bias weights in $W^{(2)}$ are contained in the interval $(-\lambda, \lambda)$; this proves assertion (i) of the theorem.

Note also that we do not actually require $\delta > 0$ to be arbitrarily small. Suppose instead that we choose $k \in \mathbb{N}$ sufficiently large, so that the convex hull of $\{\widehat{\nu}(\widehat{g}_{1,k}^{W}), \ldots, \widehat{\nu}(\widehat{g}_{N,k}^{W})\}$ contains both points $\widehat{\nu}(f^{[t]})$ and $\mathbf{0}_{N-1}$. In this case, observe that our choice of k depends only on f (via $\widehat{\nu}(f^{[t]})$) and σ (via the definition of $\{\lambda_k\}_{k\in\mathbb{N}}$). The inequality $\delta \leq \frac{2(N-1)\cdot\|\widehat{\nu}(f^{[t]})\|_2}{\lambda'_k}$ still holds for any δ satisfying $b_{j,k}^{[t]}, b'_{j,k} \in (\frac{1}{N} - \delta, \frac{1}{N} + \delta)$ for all $1 \leq j \leq N, 1 \leq t \leq m$. Consequently, our argument to show that $\|f - \rho_W^{\sigma}\|_{\infty,X} < \varepsilon$ holds verbatim, which proves assertion (ii) of the theorem. □

Proof of Theorem 3.2. Fix some $\varepsilon > 0$, and consider an arbitrary $t \in \{1, \ldots, m\}$. For each integer $d \geq 1$, let $p_d^{[t]}$ be the best polynomial approximant to $f^{[t]}$ of degree d. By Theorem 2.2, we have $\|f^{[t]} - p_d^{[t]}\|_{\infty,X} \leq 6\omega_{f^{[t]}}(\frac{D}{2d})$ for all $d \geq 1$, hence it follows from the definition of d_ε that

$$\|f^{[t]} - p_{d_\varepsilon}^{[t]}\|_{\infty,X} \leq 6\omega_{f^{[t]}}(\tfrac{D}{2d_\varepsilon}) < \varepsilon.$$

Define $\varepsilon' := \varepsilon - \max\{6\omega_{f^{[t]}}(\frac{D}{2d_\varepsilon}) : 1 \leq t \leq m\}$. Note that $\varepsilon' > 0$, and $\|f^{[t]} - p_{d_\varepsilon}^{[t]}\|_{\infty,X} \leq \varepsilon - \varepsilon'$ (for all $1 \leq t \leq m$). By Theorem 3.1, there exists some $W \in \mathcal{W}_{\binom{n+d_\varepsilon}{d_\varepsilon}}$ satisfying $\|p_{d_\varepsilon}^{[t]} - \rho_W^{\sigma\,[t]}\|_{\infty,X} < \varepsilon'$ for all $1 \leq t \leq m$, which implies

$$\|f^{[t]} - \rho_W^{\sigma\,[t]}\|_{\infty,X} \leq \|f^{[t]} - p_{d_\varepsilon}^{[t]}\|_{\infty,X} + \|p_{d_\varepsilon}^{[t]} - \rho_W^{\sigma\,[t]}\|_{\infty,X} < (\varepsilon - \varepsilon') + \varepsilon' = \varepsilon,$$

therefore $\|f - \rho_W^{\sigma}\|_{\infty,X} < \varepsilon$. Conditions (i) and (ii) follow from Theorem 3.1. □

Proof of Theorem 3.3. Most of the work has already been done earlier in the proofs of Theorem 3.1 and Theorem 3.2. The key observation is that $\det(Q[W])$ is a non-zero polynomial in terms of the weights W, hence $\{\det(Q[W]) \neq 0 : W \in \mathcal{W}_{\binom{n+d}{d}}\}$ is dense in $\mathcal{W}_{\binom{n+d}{d}}$, or equivalently, its complement has Lebesgue measure zero. □

6 Conclusion and Further Remarks

Theorem 5.6 is rather general, and could potentially be used to prove analogs of the universal approximation theorem for other classes of neural networks, such as convolutional neural networks and recurrent neural networks. In particular, finding a *single* suitable set of weights (as a representative of the infinitely many possible sets of weights in the given class of neural networks), with the property that its corresponding "non-bias Vandermonde matrix" (see Definition 5.5) is non-singular, would serve as a straightforward criterion for showing that the UAP holds for the given class of neural networks (with certain weight constraints). We formulated this criterion to be as general as we could, with the hope that it would applicable to future classes of "neural-like" networks.

We believe our algebraic approach could be emulated to eventually yield a unified understanding of how depth, width, constraints on weights, and other architectural choices, would influence the approximation capabilities of arbitrary neural networks.

Finally, we end our paper with an open-ended question. The proofs of our results in Section 5 seem to suggest that non-bias weights and bias weights play very different roles. We could impose very strong restrictions on the non-bias weights and still have the UAP. What about the bias weights?

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1J_rgWRW.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989. doi: 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

Carlos D'Andrea and Luis Felipe Tabera. Tropicalization and irreducibility of generalized Vandermonde determinants. *Proc. Amer. Math. Soc.*, 137(11):3647–3656, 2009. ISSN 0002-9939. doi: 10.1090/S0002-9939-09-09951-1. URL https://doi-org.library.sutd.edu.sg:2443/10.1090/S0002-9939-09-09951-1.

Olivier Delalleau and Yoshua Bengio. Shallow vs. deep sum-product networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 666–674. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4350-shallow-vs-deep-sum-product-networks.pdf.

Ronald A. DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta Math.*, 63(4):469–478, 1989. ISSN 0025-2611. doi: 10.1007/BF01171759. URL https://doi-org.library.sutd.edu.sg:2443/10.1007/BF01171759.

David Eisenbud. *Commutative algebra*, volume 150 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. ISBN 0-387-94268-8; 0-387-94269-6. doi: 10.1007/978-1-4612-5350-1. URL http://dx.doi.org/10.1007/978-1-4612-5350-1. With a view toward algebraic geometry.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 907–940, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL http://proceedings.mlr.press/v49/eldan16.html.

K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Netw.*, 2(3):183–192, May 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90003-8. URL http://dx.doi.org/10.1016/0893-6080(89)90003-8.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. 08 2017. Preprint arXiv:1708.02691 [stat.ML].

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8. URL http://dx.doi.org/10.1016/0893-6080(89)90020-8.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

Murray S. Klamkin and George A. Tsintsifas. The circumradius-inradius inequality for a simplex. *Math. Mag.*, 52(1):20–22, 1979. ISSN 0025-570X. doi: 10.2307/2689968. URL https://doi-org.library.sutd.edu.sg:2443/10.2307/2689968.

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. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80131-5. URL http://www.sciencedirect.com/science/article/pii/S0893608005801315.

William Judson LeVeque. *Topics in number theory. Vols. 1 and 2*. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1956.

Shiyu Liang and R. Srikant. Why deep neural networks? *CoRR*, abs/1610.04161, 2016. URL http://arxiv.org/abs/1610.04161.

Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, Sep 2017.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6231–6239. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width.pdf.

V. E. Maiorov and R. Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Adv. Comput. Math.*, 13(1):79–103, 2000. ISSN 1019-7168. doi: 10.1023/A:1018993908478. URL https://doi-org.library.sutd.edu.sg:2443/10.1023/A:1018993908478.

Vitaly Maiorov and Allan Pinkus. Lower bounds for approximation by mlp neural networks. *Neurocomputing*, 25(1):81 – 91, 1999. ISSN 0925-2312. doi: https://doi.org/10.1016/S0925-2312(98)00111-8. URL http://www.sciencedirect.com/science/article/pii/S0925231298001118.

H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1):164–177, Jan 1996. doi: 10.1162/neco.1996.8.1.164.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones?, 2017. URL https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14849.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2924–2932, Cambridge, MA, USA, 2014. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969033.2969153.

Allan Pinkus. Approximation theory of the mlp model in neural networks. *ACTA NUMERICA*, 8: 143–195, 1999.

A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555–561, Sep. 2008. doi: 10.1109/ALLERTON.2008.4797607.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pp. 1177–1184, USA, 2007. Curran Associates Inc. ISBN 978-1-60560-352-0. URL http://dl.acm.org/citation.cfm?id=2981562.2981710.

Theodore J. Rivlin. *An introduction to the approximation of functions*. Dover Publications, Inc., New York, 1981. ISBN 0-486-64069-8. Corrected reprint of the 1969 original, Dover Books on Advanced Mathematics.

Richard P. Stanley. *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-56069-1; 0-521-78987-7. doi: 10.1017/CBO9780511609589. URL https://doi-org.library.sutd.edu.sg:2443/10.1017/CBO9780511609589. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

Yitong Sun, Anna Gilbert, and Ambuj Tewari. On the approximation properties of random ReLU features. Preprint arXiv:1810.04374v3 [stat.ML], August 2019.

Matus Telgarsky. benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL http://proceedings.mlr.press/v49/telgarsky16.html.

K. Wolsson. Linear dependence of a function set of m variables with vanishing generalized Wronskians. *Linear Algebra Appl.*, 117:73–80, 1989. ISSN 0024-3795. doi: 10.1016/0024-3795(89)90548-X. URL https://doi-org.library.sutd.edu.sg:2443/10.1016/0024-3795(89)90548-X.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103 – 114, 2017. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002. URL http://www.sciencedirect.com/science/article/pii/S0893608017301545.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. Preprint arXiv:1904.00687v2 [stat.ML], June 2019.

A Generalized Wronskians and the proof of Theorem 5.6

First, we recall the notion of generalized Wronskians as given in (LeVeque, 1956, Chap. 4.3). Let $\Delta_0, \ldots, \Delta_{N-1}$ be any N differential operators of the form

$$\Delta_k = \left(\tfrac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\tfrac{\partial}{\partial x_n}\right)^{\alpha_n}, \text{ where } \alpha_1 + \cdots + \alpha_n \leq k.$$

Let $f_1, \ldots, f_N \in \mathcal{P}(\mathbb{R}^n)$. The *generalized Wronskian* of $(f_1, \ldots, f_N)$ associated to $\Delta_0, \ldots, \Delta_{N-1}$ is defined as the determinant of the matrix $M = [\Delta_{i-1} f_j(x)]_{1 \leq i,j \leq N}$. In general, $(f_1, \ldots, f_N)$ has multiple generalized Wronskians, corresponding to multiple choices for $\Delta_0, \ldots, \Delta_{N-1}$.

A.1 Proof of Theorem 5.6

For brevity, let $N = \binom{n+d}{d}$ and let $\mathbf{x} = (x_1, \ldots, x_n)$. Recall that $\lambda_1 < \cdots < \lambda_N$ are all the n-tuples in $\Lambda^n_{\leq d}$ in the colexicographic order. For each $1 \leq i, k \leq N$, write $\lambda_k = (\lambda_{k,1}, \ldots, \lambda_{k,n})$, define the differential operator $\Delta_{\lambda_k} = \left(\frac{\partial}{\partial x_1}\right)^{\lambda_{k,1}} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\lambda_{k,n}}$, and let $\alpha^{(i)}_{\lambda_k}$ be the coefficient of the monomial $q_k(\mathbf{x})$ in $\Delta_{\lambda_i} p(\mathbf{x})$. Consider an arbitrary $W \in \mathcal{U}$, and for each $1 \leq j \leq N$, define $f_j \in \mathcal{P}_{\leq d}(\mathbb{R}^n)$ by the map $\mathbf{x} \mapsto p(w^{(1)}_{1,j} x_1, \ldots, w^{(1)}_{n,j} x_n)$. Note that $\mathcal{F}_{p,\mathbf{0}_n}(W) = (f_1, \ldots, f_N)$ by definition. Next, define the matrix $M_W(\mathbf{x}) := [\Delta_i f_j(x)]_{1 \leq i,j \leq N}$, and note that $\det M_W(\mathbf{x})$ is the generalized Wronskian of $(f_1, \ldots, f_N)$ associated to $\Delta_1, \ldots, \Delta_N$. In particular, this generalized Wronskian is well-defined, since the definition of the colexicographic order implies that $\lambda_{k,1} + \cdots + \lambda_{k,n} \leq k$ for all possible k. Similar to the univariate case, $(f_1, \ldots, f_N)$ is linearly independent if (and only if) its generalized Wronskian is not the zero function (Wolsson, 1989). Thus, to show that $W \in {}^p\mathcal{U}^{\text{ind}}$, it suffices to show that the evaluation $\det M_W(\mathbf{1}_n)$ of this generalized Wronskian at $\mathbf{x} = \mathbf{1}_n$ gives a non-zero value, where $\mathbf{1}_n$ denotes the all-ones vector in $\mathbb{R}^n$.

Observe that the (i,j)-th entry of $M_W(\mathbf{1}_n)$ equals $(\widehat{\mathbf{w}}^{(1)}_j)^{\lambda_i} (\Delta_{\lambda_i} p)(\widehat{\mathbf{w}}^{(1)}_j)$, hence we can check that $M_W(\mathbf{1}_n) = M' M''$, where M' is an N-by-N matrix whose (i,j)-th entry is given by

$$M'_{i,j} = \begin{cases} \alpha^{(i)}_{\lambda_j - \lambda_i}, & \text{if } \lambda_j - \lambda_i \in \Lambda^n_{\leq d}; \\ 0, & \text{if } \lambda_j - \lambda_i \notin \Lambda^n_{\leq d}; \end{cases}$$

and where $M'' = Q[W]$ is the non-bias Vandermonde matrix of W.

It follows from the definition of the colexicographic order that $\lambda_j - \lambda_i$ necessarily contains at least one strictly negative entry whenever $j < i$, hence we infer that M' is upper triangular. The diagonal entries of M' are $\alpha^{(1)}_{\mathbf{0}_n}, \alpha^{(2)}_{\mathbf{0}_n}, \ldots, \alpha^{(N)}_{\mathbf{0}_n}$, and note that $\alpha^{(i)}_{\mathbf{0}_n} = (\lambda_{i,1}! \cdots \lambda_{i,n}!) \alpha^{(1)}_{\lambda_i}$ for each $1 \leq i \leq N$, where $\lambda_{i,1}! \cdots \lambda_{i,n}!$ denotes the product of the factorials of the entries of the n-tuple λ_i. In particular, $\lambda_{i,1}! \cdots \lambda_{i,n}! \neq 0$, and $\alpha^{(1)}_{\lambda_i}$, which is the coefficient of the monomial $q_i(\mathbf{x})$ in $p(\mathbf{x})$, is non-zero. Thus, $\det(M') \neq 0$.

We have come to the crucial step of our proof. If we can show that $\det(M'') = \det(Q[W]) \neq 0$, then $\det(M_W(\mathbf{1}_n)) = \det(M') \det(M'') \neq 0$, and hence we can infer that $W \in {}^p\mathcal{U}^{\text{ind}}$. This means that ${}^p\mathcal{U}^{\text{ind}}$ contains the subset $\mathcal{U}' \subseteq \mathcal{U}$ consisting of all W such that $Q[W]$ is non-singular. Note that $\det(Q[W])$ is a polynomial in terms of the non-bias weights in $W^{(1)}$ as its variables, so we could write this polynomial as $r = r(W)$. Consequently, if we can find a single $W \in \mathcal{U}$ such that $Q[W]$ is non-singular, then $r(W)$ is not identically zero on $\mathcal{U}$, which then implies that $\mathcal{U}' = \{W \in \mathcal{U} : r(W) \neq 0\}$ is dense in $\mathcal{U}$ (w.r.t. the Euclidean metric). $\square$

B Conjectured optimality of upper bound $\mathcal{O}(\varepsilon^{-n})$ in Theorem 3.2

It was conjectured by Mhaskar (1996) that there exists some smooth non-polynomial function σ, such that at least $\Omega(\varepsilon^{-n})$ hidden units is required to uniformly approximate every function in the

class $\mathfrak{S}$ of C^1 functions with bounded Sobolev norm. As evidence that this conjecture is true, a heuristic argument was provided in (Mhaskar, 1996), which uses a result by DeVore et al. (1989); cf. (Pinkus, 1999, Thm. 6.5). To the best of our knowledge, this conjecture remains open. If this conjecture is indeed true, then our upper bound $\mathcal{O}(\varepsilon^{-n})$ in Theorem 3.2 is optimal for general continuous non-polynomial activation functions.

For specific activation functions, such as the logistic sigmoid function, or any polynomial spline function of fixed degree with finitely many knots (e.g. the ReLU function), it is known that the minimum number N of hidden units required to uniformly approximate every function in $\mathfrak{S}$ must satisfy $(N \log N) \in \Omega(\varepsilon^{-n})$ (Maiorov & Meir, 2000); cf. (Pinkus, 1999, Thm. 6.7). Hence there is still a gap between the lower and upper bounds for N in these specific cases. It would be interesting to find optimal bounds for these cases.