

Universal Approximation of Neural Networks

Uri Shaham

May 31, 2026

1 Overview and Motivation

Before asking whether a neural network can be *learned efficiently*, we must ask whether the target function lies in the class at all — that is, whether neural networks are **expressive enough**. The results in this lecture answer this question in the affirmative, in two complementary ways.

- **Cybenko's theorem (1989)** is an existential result: any continuous function on a compact domain can be approximated arbitrarily well by a single hidden layer network with sigmoidal activations. It says nothing about *how many* neurons are required.
- **Barron's theorem (1993)** is a quantitative result: for functions whose Fourier spectrum is sufficiently concentrated (the Barron class), n neurons suffice to achieve L_2 error $O(1/n)$, *independent of the input dimension*. This is a striking escape from the curse of dimensionality.

Together, these results form the theoretical foundation of the expressive power of shallow neural networks.

2 Setup: Shallow Neural Networks

Throughout this lecture, we consider **single hidden layer** (shallow) neural networks. Given an input $x \in \mathbb{R}^d$, such a network computes

$$f_\theta(x) = \sum_{k=1}^n c_k \sigma(w_k^\top x + b_k),$$

where $w_k \in \mathbb{R}^d$ are the input weights, $b_k \in \mathbb{R}$ are biases, $c_k \in \mathbb{R}$ are the output weights, $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is an activation function, and n is the number of hidden units. We write $\Sigma_n(\sigma)$ for the class of all such networks with n hidden units.

Definition 2.1 (Sigmoidal Function). *A function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is **sigmoidal** if*

$$\sigma(t) \rightarrow \begin{cases} 1 & \text{as } t \rightarrow +\infty \\ 0 & \text{as } t \rightarrow -\infty. \end{cases}$$

The standard logistic function $\sigma(t) = (1 + e^{-t})^{-1}$ is sigmoidal, as is any shifted and scaled version. The hyperbolic tangent $\tanh$ is also sigmoidal (after rescaling). The definition does not require continuity or monotonicity.

3 Cybenko's Theorem

3.1 Statement

Theorem 3.1 (Cybenko, 1989). *Let σ be a continuous sigmoidal function. Then the class $\Sigma(\sigma)$ of single hidden layer (of finite width) networks is dense in $C([0, 1]^d)$ in the sup-norm: for every $f \in C([0, 1]^d)$ and every $\varepsilon > 0$, there exists n and parameters θ such that*

$$\sup_{x \in [0, 1]^d} |f(x) - f_\theta(x)| < \varepsilon.$$

In other words, shallow sigmoidal networks are *universal approximators* for continuous functions. The theorem is purely existential: it guarantees the existence of a good approximation but gives no bound on n .

3.2 Discriminatory Functions

The proof rests on a key property of the activation function.

Definition 3.2 (Discriminatory Function). *A bounded measurable function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is **discriminatory** if the only finite signed Borel measure¹ μ on $[0, 1]^d$ satisfying*

$$\int_{[0, 1]^d} \sigma(w^\top x + b) d\mu(x) = 0 \quad \text{for all } w \in \mathbb{R}^d, b \in \mathbb{R}$$

is the zero measure $\mu = 0$.

Intuitively, if σ is discriminatory, then one cannot find a nonzero signed measure that is “invisible” to all ridge functions $x \mapsto \sigma(w^\top x + b)$. This will be exactly the property needed to force density.

Lemma 3.3 (Sigmoidal Functions are Discriminatory). *Every continuous sigmoidal function is discriminatory.*

Proof. Let μ be a signed Borel measure on $[0, 1]^d$ satisfying $\int \sigma(w^\top x + b) d\mu(x) = 0$ for all w, b . We want to show $\mu = 0$.

Fix $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $\varphi \in \mathbb{R}$. Define

$$\sigma_\lambda(x) = \sigma(\lambda(w^\top x + b) + \varphi), \quad \lambda > 0.$$

By assumption, $\int \sigma_\lambda(x) d\mu(x) = 0$ for all $\lambda > 0$. As $\lambda \rightarrow +\infty$, pointwise:

$$\sigma_\lambda(x) \rightarrow \begin{cases} \sigma(\varphi + \infty) = 1 & \text{if } w^\top x + b > 0 \\ \sigma(\varphi) & \text{if } w^\top x + b = 0 \\ \sigma(\varphi - \infty) = 0 & \text{if } w^\top x + b < 0. \end{cases}$$

Since $|\sigma_\lambda| \leq 1$ and μ is a finite measure, **dominated convergence**² gives

$$0 = \lim_{\lambda \rightarrow \infty} \int_{[0, 1]^d} \sigma_\lambda(x) d\mu(x) = 1 \cdot \mu(\Pi_{w,b}^+) + \sigma(\varphi) \cdot \mu(\Pi_{w,b}^0) + 0 \cdot \mu(\Pi_{w,b}^-),$$

¹Without being too formal, say that a finite signed Borel measure on $[0, 1]^d$ is a function $\mu : [0, 1]^d \rightarrow \mathbb{R}$ such that $\int_{[0, 1]^d} d\mu(x) \leq \infty$.

²limit of the integral equals integral of the limit.

where the equality stems from our assumption, and $\Pi_{w,b}^+ = \{x : w^\top x + b > 0\}$, $\Pi_{w,b}^0 = \{x : w^\top x + b = 0\}$, $\Pi_{w,b}^- = \{x : w^\top x + b < 0\}$.

This holds for all $\varphi \in \mathbb{R}$. Taking $\varphi \rightarrow +\infty$: $\sigma(\varphi) \rightarrow 1$, so

$$\mu(\Pi_{w,b}^+) + \mu(\Pi_{w,b}^0) = 0.$$

Taking $\varphi \rightarrow -\infty$: $\sigma(\varphi) \rightarrow 0$, so

$$\mu(\Pi_{w,b}^+) = 0.$$

Combining: $\mu(\Pi_{w,b}^0) = 0$ as well. Therefore $\mu(\Pi_{w,b}^+) = 0$, and also $\mu(\Pi_{w,b}^- \cup \Pi_{w,b}^0) = \mu(\overline{\Pi_{w,b}^+}) = 0$.

Since every open set in $[0, 1]^d$ can be approximated via unions and intersections of half spaces, the argument above implies that μ is the zero measure. $\square$

3.3 Proof of Cybenko's Theorem

Proof of Theorem 3.1. Let $S = \overline{\text{span}}\{\sigma(w^\top \cdot + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\} \subseteq C([0, 1]^d)$ be the closure (in sup-norm) of the set of all finite linear combinations of ridge functions. This is the set of functions representable by shallow networks with arbitrarily many neurons. We want to show $S = C([0, 1]^d)$.

Suppose for contradiction that $S \neq C([0, 1]^d)$. Then S is a proper closed subspace of $C([0, 1]^d)$. By the **Hahn-Banach theorem**, there exists a bounded linear functional $L \in C([0, 1]^d)^*$ such that $L \neq 0$ but $L(g) = 0$ for all $g \in S$. By the **Riesz representation theorem**, every bounded linear functional on $C([0, 1]^d)$ is represented by a unique finite signed Borel measure μ :

$$L(g) = \int_{[0,1]^d} g(x) d\mu(x) \quad \text{for all } g \in C([0, 1]^d).$$

Since $L \neq 0$, the measure μ is nonzero.

But $L(g) = 0$ for all $g \in S$ means

$$\int_{[0,1]^d} \sigma(w^\top x + b) d\mu(x) = 0 \quad \text{for all } w \in \mathbb{R}^d, b \in \mathbb{R}.$$

By Lemma 3.3, σ discriminatory forces $\mu = 0$. This contradicts $\mu \neq 0$.

Therefore $S = C([0, 1]^d)$, which is exactly the statement that shallow sigmoidal networks are dense in $C([0, 1]^d)$. $\square$

3.4 Discussion

Remark 3.4 (What Cybenko does not say). *Cybenko's theorem is purely existential. It tells us that for every $\varepsilon > 0$ a good approximation exists, but gives no bound on how many neurons n are required as a function of ε , the dimension d , or the smoothness of f . In the worst case, n can be exponentially large.*

Remark 3.5 (Universality is not special to sigmoids). *The proof only used that σ is discriminatory. Later work (Hornik, 1991; Leshno et al., 1993) showed that a continuous function σ is a universal approximator if and only if it is not a polynomial. This includes ReLU, tanh, and essentially all activations used in practice.*

Remark 3.6 (Depth). *Cybenko's theorem is about width: it says one hidden layer with enough neurons suffices. A separate line of work shows that depth also buys expressive power in a qualitatively different way: some functions require exponentially many neurons in a shallow network but only polynomially many in a deep one.*

4 Barron's Theorem

Before stating Barron's theorem, it is important to understand what it is escaping from.

4.1 Classical Approximation Theory

In classical nonparametric approximation theory, the accuracy of approximating a function f depends critically on the input dimension d . The dependence on d , typically exponentially, is the **curse of dimensionality**: for large d , even a mildly accurate approximation requires exponentially many parameters. Any approximation scheme that inherits such a rate is said to be *cursed*. The question is whether neural networks can do better for certain function classes.

4.2 Barron's Insight

Barron (1993) identified a class of functions — those with a bounded first absolute moment of the Fourier transform — for which shallow networks achieve error $O(1/n)$, completely *free of the dimension d* . The key idea is that such functions can be written as expectations over random neurons, and sampling those neurons gives a good approximation.

4.3 Barron's Theorem

We now state and prove a classical Fourier version of Barron's theorem for two-layer neural networks with sigmoidal activation functions.

Let $\Omega \subseteq \mathbb{R}^d$ be contained in the Euclidean ball of radius R :

$$\Omega \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}.$$

Let ρ be any probability distribution on Ω , and define

$$\|g\|_{L^2(\rho)} = \left(\int_{\Omega} |g(x)|^2 d\rho(x) \right)^{1/2}.$$

Theorem 4.1 (Barron's theorem). *Let $f : \Omega \rightarrow \mathbb{R}$ admit the Fourier representation*

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega^\top x} \widehat{f}(\omega) d\omega,$$

where $\widehat{f}$ is integrable and

$$C_f = \int_{\mathbb{R}^d} \|\omega\|_2 |\widehat{f}(\omega)| d\omega < \infty.$$

Then, for every integer $m \geq 1$ and every $\eta > 0$, there exists a two-layer sigmoidal neural network

$$f_m(x) = c + \sum_{r=1}^m a_r \sigma(w_r^\top x + b_r),$$

with $a_r, c \in \mathbb{R}$, $w_r \in \mathbb{R}^d$, and $b_r \in \mathbb{R}$, such that

$$\|f - f_m\|_{L^2(\rho)} \leq \frac{2RC_f}{\sqrt{m}} + \eta.$$

Equivalently,

$$\inf_{f_m} \|f - f_m\|_{L^2(\rho)} \leq \frac{2RC_f}{\sqrt{m}},$$

where the infimum is over all m -unit two-layer networks with activation σ .

Proof. We prove the theorem in four steps.

Step 1: Reduce to approximating $f - f(0)$. Define

$$g(x) = f(x) - f(0).$$

Since

$$f(0) = \int_{\mathbb{R}^d} \hat{f}(\omega) d\omega,$$

we have

$$g(x) = \int_{\mathbb{R}^d} \left(e^{i\omega^\top x} - 1 \right) \hat{f}(\omega) d\omega.$$

At the end of the proof, we will add back the constant $f(0)$.

Step 2: Represent each Fourier mode by ridge threshold functions. This step is only an intermediate analytic device. The final approximating network will use sigmoidal units, not threshold units.

Fix $\omega \in \mathbb{R}^d$ and set

$$A_\omega = R\|\omega\|_2.$$

For every $x \in \Omega$,

$$|\omega^\top x| \leq \|\omega\|_2 \|x\|_2 \leq R\|\omega\|_2 = A_\omega.$$

Thus $t = \omega^\top x$ lies in $[-A_\omega, A_\omega]$.

For every $t \in [-A_\omega, A_\omega]$, the following identity holds:

$$e^{it} - 1 = \int_0^{A_\omega} ie^{ib} \mathbf{1}\{t \geq b\} db - \int_{-A_\omega}^0 ie^{ib} \mathbf{1}\{t \leq b\} db.$$

Indeed, if $t \geq 0$, then the second integral is zero and

$$\int_0^{A_\omega} ie^{ib} \mathbf{1}\{t \geq b\} db = \int_0^t ie^{ib} db = e^{it} - 1.$$

If $t < 0$, then the first integral is zero and

$$-\int_{-A_\omega}^0 ie^{ib} \mathbf{1}\{t \leq b\} db = -\int_t^0 ie^{ib} db = e^{it} - 1.$$

Substituting $t = \omega^\top x$, we get

$$e^{i\omega^\top x} - 1 = \int_0^{A_\omega} ie^{ib} \mathbf{1}\{\omega^\top x \geq b\} db - \int_{-A_\omega}^0 ie^{ib} \mathbf{1}\{\omega^\top x \leq b\} db.$$

Therefore

$$g(x) = \operatorname{Re} \int_{\mathbb{R}^d} \widehat{f}(\omega) \int_0^{A_\omega} ie^{ib} \mathbf{1}\{\omega^\top x \geq b\} db d\omega \\ - \operatorname{Re} \int_{\mathbb{R}^d} \widehat{f}(\omega) \int_{-A_\omega}^0 ie^{ib} \mathbf{1}\{\omega^\top x \leq b\} db d\omega.$$

Thus g is a continuous superposition of ridge threshold functions of the form

$$x \mapsto \mathbf{1}\{\omega^\top x \geq b\} \quad \text{and} \quad x \mapsto \mathbf{1}\{\omega^\top x \leq b\}.$$

The total absolute coefficient size in this representation is at most

$$B = \int_{\mathbb{R}^d} |\widehat{f}(\omega)| \left(\int_0^{A_\omega} |ie^{ib}| db + \int_{-A_\omega}^0 |ie^{ib}| db \right) d\omega \\ = \int_{\mathbb{R}^d} |\widehat{f}(\omega)| (A_\omega + A_\omega) d\omega \\ = 2R \int_{\mathbb{R}^d} \|\omega\|_2 |\widehat{f}(\omega)| d\omega \\ = 2RC_f.$$

Step 3: Approximate the threshold superposition by m sampled terms. The previous step says that g can be viewed as a signed weighted average of bounded ridge functions, with total coefficient budget at most

$$B = 2RC_f.$$

Equivalently, we can write

$$g(x) = B \mathbb{E}_\Theta [\varepsilon_\Theta h_\Theta(x)],$$

where $\Theta = (\omega, b)$, and each h_Θ is a ridge threshold function of the form $h_\Theta(x) = \mathbf{1}\{\omega^\top x \leq b\}$ or $h_\Theta(x) = \mathbf{1}\{\omega^\top x \geq b\}$. $\Theta = (\omega, b)$ has a density of $c_+(\omega, b) := |\operatorname{Re}(ie^{ib}\widehat{f}(\omega))|/B^*$ if $b \in [0, A_\omega]$, and $c_-(\omega, b) := |-\operatorname{Re}(ie^{ib}\widehat{f}(\omega))|/B^*$ if $b \in [-A_\omega, 0]$, and where $\varepsilon_\Theta \in \{-1, +1\}$ is a sign. The normalization constant is $B^* \leq B$, so that the density integrates to one. I.e., the notation $\mathbb{E}_\Theta$ simply denotes averaging with respect to the normalized absolute coefficients in the representation from Step 2.

Draw independent samples $\Theta_1, \dots, \Theta_m$ from this normalized coefficient distribution, and define

$$\widetilde{g}_m(x) = \frac{B}{m} \sum_{r=1}^m \varepsilon_{\Theta_r} h_{\Theta_r}(x).$$

For every fixed x ,

$$\mathbb{E}[\widetilde{g}_m(x)] = g(x).$$

Moreover, since

$$|\varepsilon_{\Theta_r} h_{\Theta_r}(x)| \leq 1,$$

we have

$$\operatorname{Var}(\varepsilon_\Theta h_\Theta(x)) \leq 1.$$

Therefore,

$$\mathbb{E} \left[|g(x) - \widetilde{g}_m(x)|^2 \right] \leq \frac{B^2}{m}.$$

Integrating over $x \sim \rho$ gives

$$\mathbb{E} \left[\|g - \tilde{g}_m\|_{L^2(\rho)}^2 \right] \leq \frac{B^2}{m}.$$

Hence there exists at least one deterministic choice of $\Theta_1, \dots, \Theta_m$ such that

$$\|g - \tilde{g}_m\|_{L^2(\rho)} \leq \frac{B}{\sqrt{m}} = \frac{2RC_f}{\sqrt{m}}.$$

Thus we have found an m -term ridge-threshold approximation

$$\tilde{g}_m(x) = \sum_{r=1}^m \alpha_r h_r(x)$$

such that

$$\|g - \tilde{g}_m\|_{L^2(\rho)} \leq \frac{2RC_f}{\sqrt{m}}.$$

Step 4: Replace the threshold ridge functions by sigmoidal units. Each selected threshold ridge function is of one of the two forms

$$h_r(x) = \mathbf{1}\{u_r^\top x + \beta_r \geq 0\}$$

or

$$h_r(x) = \mathbf{1}\{u_r^\top x + \beta_r \leq 0\}.$$

The second type can be written as

$$\mathbf{1}\{-u_r^\top x - \beta_r \geq 0\}.$$

So it is enough to consider thresholds of the form

$$h_r(x) = \mathbf{1}\{u_r^\top x + \beta_r \geq 0\}.$$

For $\lambda > 0$, define the steep sigmoidal ridge function

$$h_{r,\lambda}(x) = \sigma(\lambda(u_r^\top x + \beta_r)).$$

Because σ is sigmoidal,

$$h_{r,\lambda}(x) \rightarrow h_r(x)$$

for every x such that

$$u_r^\top x + \beta_r \neq 0.$$

Since the thresholds in Step 3 are drawn from intervals with ordinary one-dimensional densities, these exceptional thresholds are avoided with probability one.

Therefore, for each selected unit,

$$h_{r,\lambda} \rightarrow h_r \quad \text{in } L^2(\rho).$$

Since there are only finitely many selected units, we may choose λ large enough for each unit so that

$$\left\| \sum_{r=1}^m \alpha_r h_r - \sum_{r=1}^m \alpha_r h_{r,\lambda_r} \right\|_{L^2(\rho)} \leq \eta.$$

Thus there exists a sigmoidal network

$$g_m(x) = \sum_{r=1}^m a_r \sigma(w_r^\top x + b_r)$$

such that

$$\|g - g_m\|_{L^2(\rho)} \leq \frac{2RC_f}{\sqrt{m}} + \eta.$$

Finally, recalling that $g = f - f(0)$, define

$$f_m(x) = f(0) + g_m(x).$$

Then

$$f_m(x) = c + \sum_{r=1}^m a_r \sigma(w_r^\top x + b_r)$$

with $c = f(0)$, and

$$\|f - f_m\|_{L^2(\rho)} \leq \frac{2RC_f}{\sqrt{m}} + \eta.$$

This proves the theorem. □

Interpretation. The Fourier quantity

$$C_f = \int_{\mathbb{R}^d} \|\omega\|_2 |\hat{f}(\omega)| d\omega$$

measures the amount of high-frequency content in f . If C_f is finite, it means that f cannot have too strong high-frequency terms. In such cases, we say that f belongs to the Barron class. Barron's theorem says that such functions can be approximated by two-layer sigmoidal networks at rate

$$\|f - f_m\|_{L^2(\rho)} \lesssim \frac{C_f}{\sqrt{m}}.$$

The important point is that the approximation rate depends on the input dimension only through the Barron quantity C_f , rather than through an explicit exponentially bad factor in d .