Depth Separation for Neural Networks

Amit Daniely*

March 1, 2017

Abstract

Let \( f : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R} \) be a function of the form \( f(x, x') = g(\langle x, x' \rangle) \) for \( g : [-1, 1] \rightarrow \mathbb{R} \). We give a simple proof that shows that poly-size depth two neural networks with (exponentially) bounded weights cannot approximate \( f \) whenever \( g \) cannot be approximated by a low degree polynomial. Moreover, for many \( g \)'s, such as \( g(x) = \sin(\pi d^3 x) \), the number of neurons must be \( 2^{\Omega(d \log(d))} \). Furthermore, the result holds w.r.t. the uniform distribution on \( S^{d-1} \times S^{d-1} \). As many functions of the above form can be well approximated by poly-size depth three networks with poly-bounded weights, this establishes a separation between depth two and depth three networks w.r.t. the uniform distribution on \( S^{d-1} \times S^{d-1} \).

1 Introduction and main result

Many aspects of the expressive power of neural networks has been studied over the years. In particular, separation for deep networks [11, 10], expressive power of depth two networks [4, 8, 7, 2], and more [5, 3]. We focus on the basic setting of depth 2 versus depth 3 networks. We ask what functions are expressible (or well approximated) by poly-sized depth-3 networks, but cannot be approximated by an exponential size depth-2 network.

Two recent papers [9, 6] addressed this issue. Both papers presented a specific function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) and a distribution \( \mathcal{D} \) on \( \mathbb{R}^d \) such that \( f \) can be approximated w.r.t. \( \mathcal{D} \) by a poly(\( d \))-size depth 3 network, but not by a poly(\( d \))-size depth 2 network. In Martens et al. [9] this was shown for \( f \) being the inner product mod 2 and \( \mathcal{D} \) being the uniform distribution on \( \{0, 1\}^d \times \{0, 1\}^d \). In Eldan and Shamir [6] it was shown for a different (radial) function and some (unbounded) distribution.

We extend the above results and prove a similar result for an explicit and rich family of functions, and w.r.t. the uniform distribution on \( S^{d-1} \times S^{d-1} \). In addition, our lower bound on the number of required neurons is stronger: while previous papers showed that the number of neurons has to be exponential in \( d \), we show exponential dependency on \( d \log(d) \). Last, our proof is short, direct and is based only on basic Harmonic analysis over the sphere. In contrast, Eldan and Shamir [6]’s proof is rather lengthy and requires advanced technical tools such as tempered distributions, while Martens et al. [9] relied on the discrepancy of the inner product function mod 2. On the other hand, Eldan and Shamir [6] do not put any restriction on the magnitude of the weights, while we and Martens et al. [9] do require a mild (exponential) bound.

*Google Brain
Let us fix an activation function \( \sigma : \mathbb{R} \to \mathbb{R} \). For \( x \in \mathbb{R}^n \) we denote \( \sigma(x) = (\sigma(x_1), \ldots, \sigma(x_n)) \). We say that \( F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R} \) can be implemented by a depth-2 \( \sigma \)-network of width \( r \) and weights bounded by \( B \) if

\[
F(x, x') = w_3^T \sigma(W_1 x + W'_1 x' + b_1) + b_2 ,
\]

where \( W_1, W'_1 \in [-B, B]^{r \times d}, w_2 \in [-B, B]^r, b_1 \in [-B, B]^r \) and \( b_2 \in [-B, B] \). Similarly, \( F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R} \) can be implemented by a depth-3 \( \sigma \)-network of width \( r \) and weights bounded by \( B \) if

\[
F(x, x') = w_3^T \sigma(W_2 \sigma(W_1 x + W'_1 x' + b_1) + b_2) + b_3
\]

for \( W_1, W'_1 \in [-B, B]^{r \times d}, W_2 \in [-B, B]^{r \times r}, w_3 \in [-B, B]^r, b_1, b_2 \in [-B, B]^r \) and \( b_3 \in [-B, B] \). Denote

\[
N_{d,n} = \binom{d + n - 1}{d - 1} - \binom{d + n - 3}{d - 1} = \frac{(2n + d - 2)(n + d - 3)!}{n!(d-2)!}.
\]

Let \( \mu_d \) be the probability measure on \([-1, 1]\) given by \( d\mu_d(x) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi} \Gamma\left(d/2\right)} (1 - x^2)^{d/2 - 3/2} dx \) and define

\[
A_{n,d}(f) = \min_{p \text{ is degree } n-1 \text{ polynomial}} \|f - p\|_{L^2(\mu_d)}
\]

Our main theorem shows that if \( A_{n,d}(f) \) is large then \((x, x') \mapsto f((x, x'))\) cannot be approximated by a small depth-2 network.

**Theorem 1** (main). Let \( N : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R} \) be any function implemented by a depth-2 \( \sigma \)-network of width \( r \), with weights bounded by \( B \). Let \( f : [-1,1] \to \mathbb{R} \) and define \( F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R} \) by \( F(x, x') = f((x, x')) \). Then, for all \( n \),

\[
\|N - F\|_{L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})} \geq A_{n,d}(f) \left( A_{n,d}(f) - \frac{2rB \max_{|x| \leq \sqrt{4dB+B}} |\sigma(x)| + 2B}{\sqrt{N_{d,n}}} \right)
\]

**Example 2.** Let us consider the case that \( \sigma(x) = \max(0, x) \) is the ReLU function, \( f(x) = \sin(\pi d^3 x), n = d^2 \) and \( B = 2d \). In this case, lemma 5 implies that \( A_{n,d}(f) \geq \frac{1}{5\pi} \). Hence, to have \( \frac{1}{5\pi \pi^2} \)-approximation of \( F \), the number of hidden neurons has to be at least,

\[
\sqrt{N_{d,n}} = \frac{1}{20c \pi 2^d (1 + \sqrt{4d}) + 2d^4} \leq 2^\Omega(d \log(d))
\]

On the other hand, corollary 7 implies that \( F \) can be \( \epsilon \)-approximated by a ReLU network of depth 3, width \( \frac{16d^9}{\epsilon} \) and weights bounded by \( 2\pi d^5 \).

## 2 Proofs

Throughout, we fix a dimension \( d \). All functions \( f : \mathbb{S}^{d-1} \to \mathbb{R} \) and \( f : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R} \) will be assumed to be square integrable w.r.t. the uniform measure. Likewise, functions \( f : [-1,1] \to \mathbb{R} \) and \( f : [-1,1] \times [-1,1] \to \mathbb{R} \) will be assumed to be square integrable w.r.t. \( \mu_d \) or \( \mu_d \times \mu_d \). Norms and inner products of such functions are of the corresponding \( L^2 \) spaces. We will use the fact that \( \mu_d \) is the probability measure on \([-1,1]\) that is obtained by pushing forward the uniform measure on \( \mathbb{S}^{d-1} \) via the function \( x \mapsto x_1 \). We denote by \( \mathcal{P}_n : L^2(\mu_d) \to L^2(\mu_d) \) the projection on the complement of the space of degree \( \leq n - 1 \) polynomials. Note that \( A_{n,d}(f) = \|\mathcal{P}_n df\|_{L^2(\mu_d)} \).
2.1 Some Harmonic Analysis on the Sphere

The \( d \) dimensional Legendre polynomials are the sequence of polynomials over \([-1,1]\) defined by the recursion formula

\[
P_n(x) = \frac{2n+d-4}{n+d-3}xP_{n-1}(x) - \frac{n-1}{n+d-3}P_{n-2}(x)
\]

\[P_0 = 1, \quad P_1(x) = x\]

We also define \( h_n : S^{d-1} \times S^{d-1} \to \mathbb{R} \) by \( h_n(x, x') = \sqrt{N_{d,n}}P_n(\langle x, x' \rangle) \), and for \( x \in S^{d-1} \) we denote \( L_n^x(x') = h_n(x, x') \). We will make use of the following properties of the Legendre polynomials.

Proposition 3 (e.g. [1] chapters 1 and 2).

1. For every \( d \geq 2 \), the sequence \( \{\sqrt{N_{d,n}}P_n\} \) is orthonormal basis of the Hilbert space \( L^2(\mu_d) \).
2. For every \( n \), \( \|P_n\|_\infty = 1 \) and \( P_n(1) = 1 \).
3. \( \langle L_n^x, L_j^x' \rangle = \delta_{ij} \).

2.2 Main Result

We say that \( f : S^{d-1} \times S^{d-1} \to \mathbb{R} \) is an inner product function if it has the form \( f(x, x') = \phi(\langle x, x' \rangle) \) for some function \( \phi : [-1,1] \to \mathbb{R} \). Let \( \mathcal{H}_d \subset L^2(S^{d-1} \times S^{d-1}) \) be the space of inner product functions. We note that

\[
\|f\|^2 = \mathbb{E}_{x,x'} \phi^2(\langle x, x' \rangle) = \mathbb{E}_x \|\phi\|^2 = \|\phi\|^2
\]

Hence, the correspondence \( \phi \leftrightarrow f \) defines an isomorphism of Hilbert spaces between \( L^2(\mu_d) \) and \( \mathcal{H}_d \). In particular, the orthonormal basis \( \{\sqrt{N_{d,n}}P_n\}_n \) is mapped to \( \{h_n\}_n \). In particular,

\[
\mathcal{P}_n \left( \sum_{i=0}^{\infty} \alpha_i h_i \right) = \sum_{i=n}^{\infty} \alpha_i h_i
\]

Let \( v, v' \in S^{d-1} \). We say that \( f : S^{d-1} \times S^{d-1} \to \mathbb{R} \) is \((v, v')\)-separable if it has the form \( f(x, x') = \psi(\langle v, x \rangle, \langle v', x' \rangle) \) for some \( \psi : [-1,1]^2 \to \mathbb{R} \). We note that each neuron implements a separable function. Let \( \mathcal{H}_{v,v'} \subset L^2(S^{d-1} \times S^{d-1}) \) be the space of \((v, v')\)-separable functions. We note that

\[
\|f\|^2 = \mathbb{E}_{x,x'} \psi^2(\langle v, x \rangle, \langle v', x' \rangle) = \|\psi\|^2
\]

Hence, the correspondence \( \psi \leftrightarrow f \) defines an isomorphism of Hilbert spaces between \( L^2(\mu_d \times \mu_d) \) and \( \mathcal{H}_{v,v'} \). In particular, the orthonormal basis \( \{\sqrt{N_{d,n}}P_n \otimes \sqrt{N_{d,m}}P_m\}_{n,m} \) is mapped to \( \{L_n^v \otimes L_n^{v'}\}_{n,m} \).

The following theorem implies theorem 1, as under the conditions of theorem 1, any hidden neuron implement a separable function with norm at most \( B \max_{|x| \leq \sqrt{4dB} + B} |\sigma(x)| \), and the bias term is a separable function with norm at most \( B \).

Theorem 4. Let \( f : S^{d-1} \times S^{d-1} \to \mathbb{R} \) be an inner product function and let \( g_1, \ldots, g_r : S^{d-1} \times S^{d-1} \to \mathbb{R} \) be separable functions. Then

\[
\left\| f - \sum_{i=1}^{r} g_i \right\|^2 \geq \|P_n f\|^2 \left( \|P_n f\| - \frac{2\sum_{i=1}^{r} \|g_i\|}{\sqrt{N_{d,n}}} \right)
\]
Proof. We note that

\[
\mathbb{E} h_n(x, x') L^j_l(x) L^j_l(x') = \mathbb{E} L^j_l(x) \mathbb{E} h_n(x, x') L^j_l(x')
\]

\[
= \mathbb{E} L^j_l(x) \mathbb{E} L^j_l(x') \mathbb{E} h_n(x, x')
\]

\[
= \delta_{nj} \mathbb{E} L^j_l(x) P_n(\langle x, v' \rangle)
\]

\[
= \frac{\delta_{nj} \delta_{ni} P_n(\langle v, v' \rangle)}{\sqrt{N_{d,n}}}
\]

(2)

Suppose now that \( f = \sum_{i=n}^{\infty} \alpha_i h_i \) and suppose that \( g = \sum_{j=1}^{r} g_j \) where each \( g_j \) depends only on \( \langle v_j, x \rangle, \langle v'_j, x \rangle \) for some \( v_j, v'_j \in S^{d-1} \). Write \( g_j(x, x') = \sum_{k,l=0}^{\infty} \beta_{k,l}^j L^j_k(x) L^j_l(x') \). By equation (2), \( L^j_k(x) L^j_l(x') \) is orthogonal to \( f \) whenever \( k \neq l \). Hence, if we replace each \( g_j \) with \( \sum_{k=0}^{\infty} \beta_{k,k}^j L^j_k(x) L^j_k(x') \), the l.h.s. of (1) does not increase. Likewise, the r.h.s. does not decrease. Hence, we can assume w.l.o.g. that each \( g_j \) is of the form \( g_j(x, x') = \sum_{i=0}^{\infty} \beta_{i}^j L^j_i(x) L^j_i(x') \). Now, using (2) again, we have that

\[
\| f - g \|^2 = \sum_{i=0}^{\infty} \left\| \alpha_i h_i - \sum_{j=1}^{r} \beta_{i}^j L^j_i \otimes L^j_i \right\|^2
\]

\[
\geq \sum_{i=0}^{\infty} \left\| \alpha_i h_i - \sum_{j=1}^{r} \beta_{i}^j L^j_i \otimes L^j_i \right\|^2
\]

\[
\geq \sum_{i=0}^{\infty} \alpha_i^2 - 2 \sum_{i=0}^{\infty} \sum_{j=1}^{r} \langle \alpha_i h_i, \beta_{i}^j L^j_i \otimes L^j_i \rangle
\]

\[
= \| P_n f \|^2 - 2 \sum_{i=0}^{\infty} \sum_{j=1}^{r} \beta_{i}^j \alpha_i P_n(\langle v_j, v'_j \rangle) \sqrt{N_{d,k}}
\]

\[
\geq \| P_n f \|^2 - 2 \sum_{j=1}^{r} \sum_{i=0}^{\infty} \frac{\beta_{i}^j \left| \alpha_i \right| \sqrt{N_{d,n}}}{\sqrt{N_{d,n}}}
\]

\[
\geq \| P_n f \|^2 - 2 \sum_{j=1}^{r} \frac{1}{\sqrt{N_{d,n}}} \sqrt{\sum_{i=0}^{\infty} \left| \beta_{i}^j \right|^2} \sqrt{\sum_{i=0}^{\infty} \left| \alpha_i \right|^2}
\]

\[
\geq \| P_n f \|^2 - 2 \| P_n f \| \sum_{j=1}^{r} \| g_j \| \sqrt{N_{d,n}}
\]

\[
\square
\]
Corollary 7. \( s \) satisfies \( S \) least layer weights bounded by 2, and prediction layer weights bounded by 4.

Lemma 5. Define \( g_{d,m}(x) = \sin \left( \pi \sqrt{dx} \right) \). Then, for any \( d \geq d_0 \), for a universal constant \( d_0 > 0 \), and for any degree \( k \) polynomial \( p \) we have

\[
\int_{-1}^{1} (g_{d,m}(x) - p(x))^2 d\mu_d(x) \geq \frac{m - k}{4m}.
\]

Proof. We have that (e.g. [1]) \( d\mu_d(x) = \frac{\Gamma(d)}{\sqrt{\pi \Gamma(d + 1)}} (1 - x^2)^{d-\frac{3}{2}} dx \). Likewise, for large enough \( d \) and \( |x| < \frac{1}{\sqrt{d}} \) we have \( 1 - x^2 \geq e^{-2x^2} \geq e^{-\frac{d}{4}} \) and hence \( (1 - x^2)^{d-\frac{3}{2}} \geq e^{-d^3} \geq e^{-1} \). Likewise, since \( \frac{\Gamma(d)}{\Gamma(d + 1)} \approx \sqrt{\frac{d}{2}} \), we have that for large enough \( d \) and \( |x| \leq \frac{1}{\sqrt{d}} \), \( d\mu_d(x) \geq \frac{\sqrt{d}}{2\pi} \). Hence, for \( f \geq 0 \) we have

\[
\int_{-1}^{1} f(x) d\mu_d(x) \geq \int_{-\frac{1}{\sqrt{d}}}^{\frac{1}{\sqrt{d}}} f(x) d\mu_d(x) \geq \frac{\sqrt{d}}{2\pi} \int_{-\frac{1}{\sqrt{d}}}^{\frac{1}{\sqrt{d}}} f(x) dx = \frac{1}{2\pi} \int_{-1}^{1} f \left( \frac{t}{\sqrt{d}} \right) dt.
\]

Applying this equation for \( f = g_{d,m} - p \) we get that

\[
\int_{-1}^{1} (g_{d,m}(x) - p(x))^2 d\mu_d(x) \geq \frac{1}{2\pi} \int_{-1}^{1} (\sin(\pi mx) - q(x))^2 dx
\]

Where \( q(x) := p \left( \frac{x}{\sqrt{d}} \right) \). Now, in the \( 2m \) segments \( I_i = (-1 + \frac{i-1}{m}, -1 + \frac{i}{m}) \), \( i \in \{2m\} \) we have at least \( m - k \) segments on which \( x \mapsto \sin(\pi mx) \) and \( q \) do not change signs and have opposite signs.

On each of these intervals we have \( \int_I (\sin(\pi mx) - q(x))^2 dx \geq \int_0^\pi \sin^2(\pi mx) dx = \frac{1}{2m} \).

Lemma 6 (e.g. [6]). Let \( \sigma(x) = \max(x,0) \) be the ReLU activation, \( f : [-R,R] \to \mathbb{R} \) an L-Lipschitz function, and \( \epsilon > 0 \). There is a function

\[
g(x) = f(0) + \sum_{i=1}^{m} \alpha_i \sigma(\gamma_i x - \beta_i)
\]

for which \( ||g - f||_\infty \leq \epsilon \). Furthermore, \( m \leq \frac{2RL}{\epsilon} \), \( |\beta_i| \leq R \), \( |\alpha_i| \leq 2L \), \( \gamma_i \in \{-1,1\} \), and \( g \) is L-Lipschitz on all \( \mathbb{R} \).

Corollary 7. Let \( f : [-1,1] \to [-1,1] \) be an L-Lipschitz function and let \( \epsilon > 0 \). Define \( F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [-1,1] \) by \( F(x,x') = f(\langle x, x' \rangle) \). There is a function \( G : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [-1,1] \) that satisfies \( ||F - G||_\infty \leq \epsilon \) and furthermore \( G \) can be implemented by a depth-3 ReLU network of width \( \frac{16dL}{\epsilon} \) and weights bounded by \( \max(4,2L) \).

Proof. By Lemma 6 there is a depth-2 network \( N_{\text{square}} \) that calculates \( \frac{x^2}{2} \) in \([-2,2] \), with an error of \( \frac{\epsilon}{2L} \) and has width at most \( \frac{16dL}{\epsilon} \) and hidden layer weights bounded by 2, and prediction layer weights bounded by 4. For each \( i \in \{d\} \) we can compose the linear function \( \langle x, x' \rangle \mapsto x_i + x_i' \) with \( N_{\text{square}} \) to get a depth-2 network \( N_i \) that calculates \( \frac{(x_i + x_i')^2}{2} \) with an error of \( \frac{\epsilon}{2dL} \) and has the same width and weight bound as \( N_{\text{square}} \). Summing the networks \( N_i \) and subtracting 1 results with a depth-2 network \( N_{\text{inner}} \) that calculates \( \langle x, x' \rangle \) with an error of \( \frac{\epsilon}{2dL} \) and has width \( \frac{16d^2L}{\epsilon} \) and hidden layer weights bounded by 2, and prediction layer weights bounded by 4.
Now, again by lemma 6 there is a depth-2 network $N_f$ that calculates $f$ in $[-1, 1]$, with an error of $\frac{\epsilon}{2}$, has width at most $\frac{2L}{\epsilon}$, hidden layer weights bounded by 1 and prediction layer weights bounded by $2L$, and is $L$-Lipschitz. Finally, consider the depth-3 network $N_F$ that is the composition of $N_{\text{inner}}$ and $N_f$. $N_F$ has width at most $\frac{16dL}{\epsilon}$ weight bound of $\max(4, 2L)$, and it satisfies

$$|N_F(x, x') - F(x, x')| = |N_f(N_{\text{inner}}(x, x')) - f(\langle x, x' \rangle)|$$
$$\leq |N_f(N_{\text{inner}}(x, x')) - N_f(\langle x, x' \rangle)| + |N_f(\langle x, x' \rangle) - f(\langle x, x' \rangle)|$$
$$\leq L|N_{\text{inner}}(x, x') - \langle x, x' \rangle| + \frac{\epsilon}{2}$$
$$\leq L \frac{\epsilon}{2L} + \frac{\epsilon}{2} = \epsilon$$

References

[1] K. Atkinson and W. Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044. Springer, 2012.

[2] Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

[3] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *29th Annual Conference on Learning Theory*, pages 698–728, 2016.

[4] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 1989.

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

[6] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, pages 907–940, 2016.

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

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

[9] James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2877–2885, 2013.

[10] Itay Safran and Ohad Shamir. Depth separation in relu networks for approximating smooth non-linear functions. *arXiv preprint arXiv:1610.09887*, 2016.

[11] Matus Telgarsky. Representation benefits of deep feedforward networks. In *COLT*, 2016.