Depth separation for neural networks

Paper by Amit Daniely

Tomáš Kocák

May 10, 2019

Depth separation

Is there a function $F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ such that

- There exist a poly(d)-sized depth-3 network N_3 s.t.

$$\|N_3 - F\|_{L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})} \le \varepsilon$$

- For every poly(d)-sized depth-2 neural network N_2

$$\|N_2 - F\|_{L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})} > \varepsilon$$

Depth separation

Is there a function $F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ such that

- There exist a poly(d)-sized depth-3 network N_3 s.t.

$$\|N_3 - F\|_{L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})} \leq \varepsilon$$

- For every poly(d)-sized depth-2 neural network N_2

$$\|N_2 - F\|_{L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})} > \varepsilon$$

We restrict our search to the case of **inner product functions** i.e.

$$\boxed{F(\mathbf{x}, \mathbf{x}') = f(\langle \mathbf{x}, \mathbf{x}' \rangle)}$$

where $f : [-1, 1] \to \mathbb{R}$.

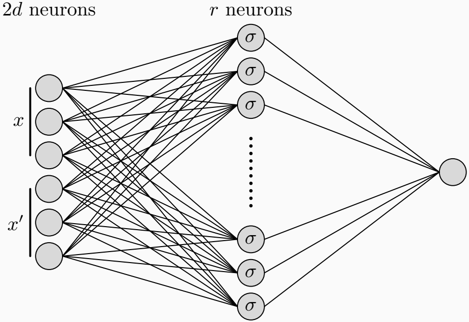
[Martens et al. 2013] [Eldam and Shamir 2016] - similar results

Depth-2 σ-network

Function $N : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ is implementing a depth-2 σ-network of width r and weights bounded by B if

$$N(\mathbf{x}, \mathbf{x}') = w_2^\intercal \sigma(W_1 \mathbf{x} + W_1' \mathbf{x}' + b1) + b_2$$

$W_1,\ W_1' \in [-B,\ B]^{r \times d}$, $w_2,\ b_1 \in [-B,\ B]^r$, $b_2 \in [-B,\ B]$.

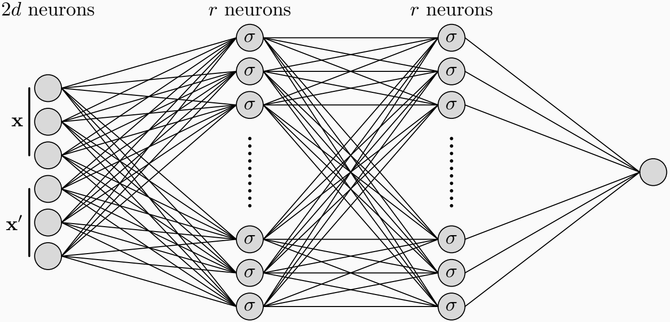


Depth-3 σ-network

Function $N : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ is implementing a depth-3 σ-network of width r and weights bounded by B if

$$N(\mathbf{x}, \mathbf{x}') = w_3^{\mathsf{T}} \sigma(W_2 \sigma(W_1 \mathbf{x} + W_1' \mathbf{x}' + b1) + b_2) + b_3$$

$W_1,\ W_1' \in [-B,\ B]^{r \times d}$, $W_2 \in [-B,\ B]^{r \times r}$, $b_1,\ b_2 \in [-B,\ B]^r$, $b_3 \in [-B,\ B]$.



Lipschitz function approximation

Every Lipschitz function can be ε-approximated by a poly-sized depth-2 NN:

- $\sigma(x) = \max\{0, x\}$ is the ReLU activation function
- $f : [-R, R] \to \mathbb{R}$ is an L-Lipschitz function
- There is a function (implemented by a depth-2 neural network)

$$\boxed{N_2(x) = f(0) + \sum_{i=1}^{m} \gamma_i \sigma(\alpha_i x + \beta_i)}$$

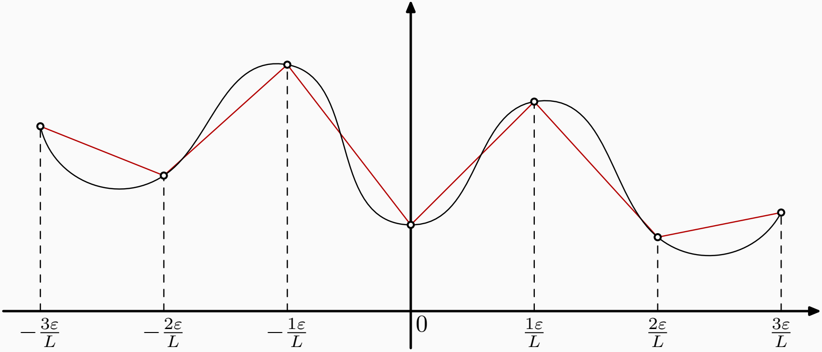
- $\|f - N_2\|_\infty \leq \varepsilon$
- N_2 is L-Lipschitz on all $\mathbb{R}$
- Width bounded as $m \leq \frac{2RL}{\varepsilon}$

- $\alpha_i \in \{-1, 1\}$
- $|\beta_i| \leq R$
- $|\gamma_i| \leq 2L$

[Cybenko 1989] [Hornik et al. 1989] [Funahashi 1989] - N_2 is a universal approximator

Lipschitz function approximation - proof

$$N_2(x) = f(0) + \sum_{i=1}^{m} \gamma_i \sigma(\alpha_i x + \beta_i)$$



Lipschitz function approximation - proof

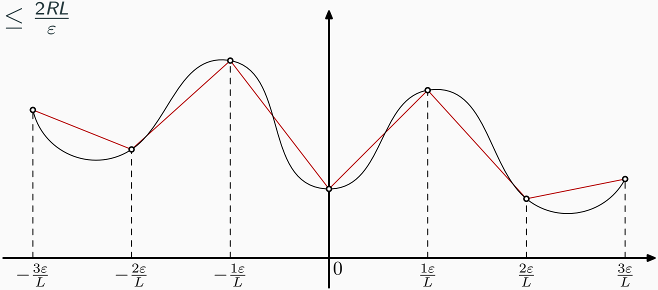
For every $x,\, x_1,\, x_2 \in \left\langle \frac{i\varepsilon}{L},\, \frac{(i+1)\varepsilon}{L} \right\rangle$

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2| \leq L\frac{\varepsilon}{L} = \varepsilon$$

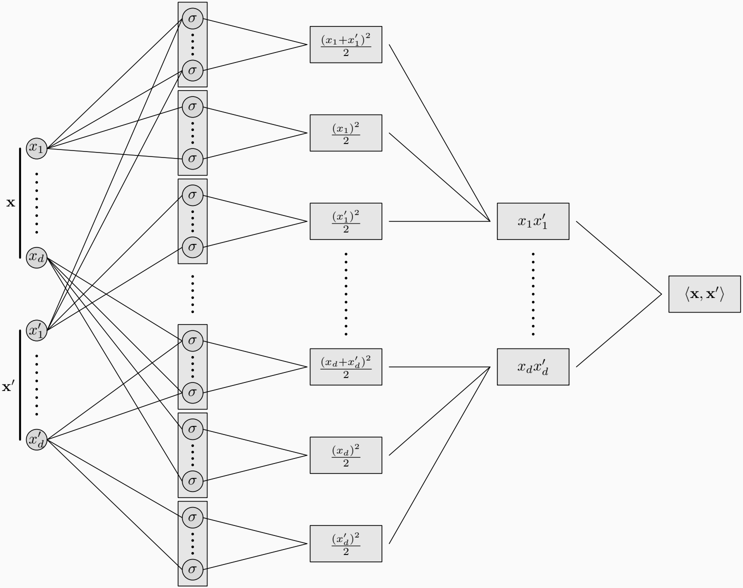
Therefore we have

$$\left.\begin{array}{l} |N_2(\frac{i\varepsilon}{L}) - f(x)| \leq \varepsilon \\ |N_2(\frac{(i+1)\varepsilon}{L}) - f(x)| \leq \varepsilon \end{array}\right\} |N_2(x) - f(x)| \leq \varepsilon$$

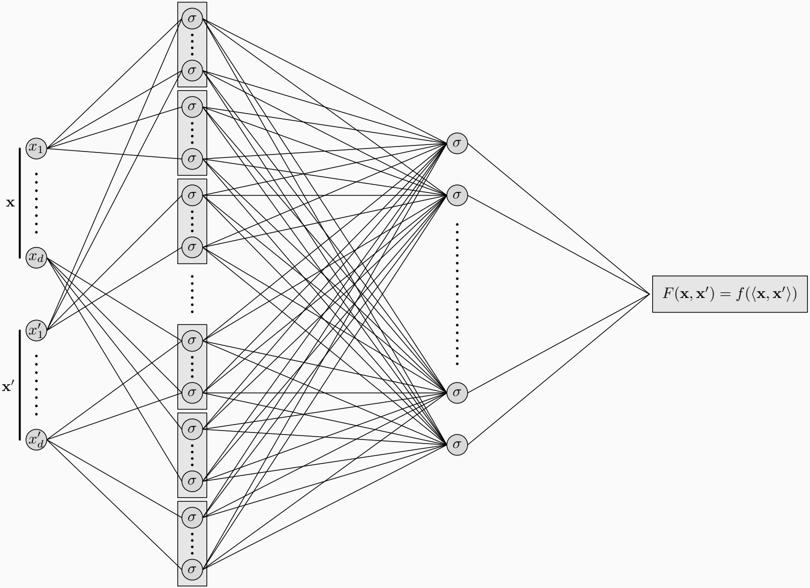
And $m \leq \frac{2RL}{\varepsilon}$



Inner product approximation



Inner product function approximation



Inner product function approximation

Inner product approximated by N_i

- Approximation precision: $\frac{\varepsilon}{2L}$
- Width of approximation N_i: $\frac{16d^2L}{\varepsilon}$

L-Lipschitz function f approximated by N_f

- Approximation precision: $\frac{\varepsilon}{2}$
- Width of approximation N_f: $\frac{4L}{\varepsilon}$

Inner product function approximated by $N_F = N_f \circ N_i$.

- Width of approximation N_F: $\frac{16d^2L}{\varepsilon}$
- Approximation precision:

$$\begin{aligned} |N_F(\mathbf{x}, \mathbf{x}') - F(\mathbf{x}, \mathbf{x}')| &= |N_f(N_i(\mathbf{x}, \mathbf{x}')) - f(\langle \mathbf{x}, \mathbf{x}' \rangle)| \\ &\leq |N_f(N_i(\mathbf{x}, \mathbf{x}')) - N_f(\langle \mathbf{x}, \mathbf{x}' \rangle)| + |N_f(\langle \mathbf{x}, \mathbf{x}' \rangle) - f(\langle \mathbf{x}, \mathbf{x}' \rangle)| \\ &\leq L|N_i(\mathbf{x}, \mathbf{x}') - \langle \mathbf{x}, \mathbf{x}' \rangle| + \frac{\varepsilon}{2} \leq L\frac{\varepsilon}{2L} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

Example

Highly oscillating inner product function:

$$\boxed{F(\mathbf{x}, \mathbf{x}') = f(\langle \mathbf{x}, \mathbf{x}' \rangle) = \sin(\pi d^3 \langle \mathbf{x}, \mathbf{x}' \rangle)}$$

$$\sin(x) \text{ is 1-Lipschitz} \quad \Longrightarrow \quad \sin(\pi d^3 x) \text{ is } (\pi d^3)\text{-Lipschitz}$$

We can ε-approximate F by a depth-3 neural network of width at most

$$\frac{16 d^2 L}{\varepsilon} = \frac{16 \pi d^5}{\varepsilon}$$

Depth-2 neural network

Slightly more technical part

Legendre polynomials

$P_0(x) = 1,\ P_1(x) = x$

$$\boxed{P_n(x) = \frac{2n+d-4}{n+d-3} x P_{n-1}(x) - \frac{n-1}{n+d-3} P_{n-2}(x)}$$

Sequence $\{\sqrt{N_{d,n}} P_n\}_{n\geq 0}$ is **orthonormal basis** of $L^2(\mu_d)$ where

$$N_{n,d} = \binom{d+n-1}{d-1} - \binom{d+n-3}{d-1}$$

and μ_d is defined by pushing forward the uniform measure on $\mathbb{S}^{d-1}$ using function $\mathbf{x} \to x_1$

$$d\mu_d(x) = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}(1-x^2)^{\frac{d-3}{2}} dx$$

Legendre polynomials

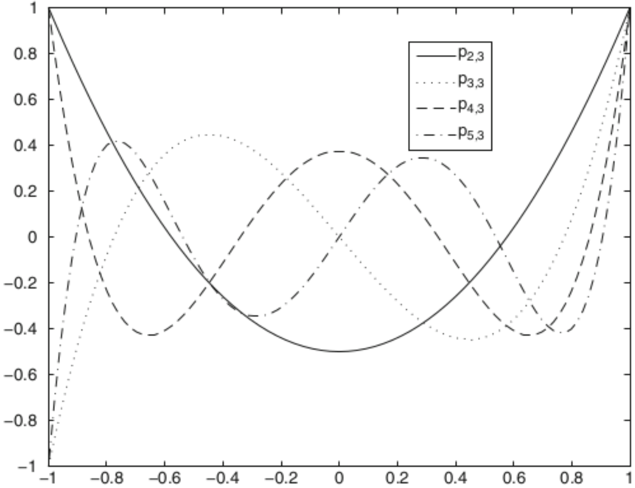


Fig. 2.2 Legendre polynomials for dimension 3

Inner product functions

Denote

$$h_n(\mathbf{x}, \mathbf{x}') = \sqrt{N_{d,n}} P_n(\langle \mathbf{x}, \mathbf{x}' \rangle)$$

Then

$\{h_n\}_{n \geq 0}$ is a **basis** of the space of inner product functions

Let $F(\mathbf{x}, \mathbf{x}') = f(\langle \mathbf{x}, \mathbf{x}' \rangle)$ be any inner product function. Then

$$F(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} \alpha_i h_i(\mathbf{x}, \mathbf{x}')$$

Separable functions

Function $g(\mathbf{x}, \mathbf{x}')$ is **(v, v')-separable function** if

$$\boxed{g(\mathbf{x}, \mathbf{x}') = \psi(\langle \mathbf{v}, \mathbf{x} \rangle, \langle \mathbf{v}', \mathbf{x}' \rangle)}$$

Denote

$$L_n^{\mathbf{x}}(\mathbf{x}') = h_n(\mathbf{x}, \mathbf{x}') = \sqrt{N_{d,n}} P_n(\langle \mathbf{x}, \mathbf{x}' \rangle)$$

$$\boxed{\{L_i^{\mathbf{v}}(\mathbf{x}) L_j^{\mathbf{v}'}(\mathbf{x}')\}_{i,j \geq 0} \text{ – } \textbf{basis} \text{ of } (\mathbf{v}, \mathbf{v}')\text{–separable functions}}$$

Any $(\mathbf{v}, \mathbf{v}')$–separable function $g(\mathbf{x}, \mathbf{x}')$ can be written as

$$g(\mathbf{x}, \mathbf{x}') = \sum_{i,j \geq 0} \beta_{i,j} L_i^{\mathbf{v}}(\mathbf{x}) L_j^{\mathbf{v}'}(\mathbf{x}')$$

Note: neuron $\sigma(\langle \mathbf{v}, \mathbf{x} \rangle + \langle \mathbf{v}', \mathbf{x}' \rangle + \mathbf{b})$ is a separable function

Main result

Theorem

Let $F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ *be an inner product function and let* $g_1, g_2, \ldots, g_r$ *be separable functions. Then we have*

$$\left\| F - \sum_{j=1}^{r} g_j \right\|^2 = \|\mathcal{P}_n F\| \left(\|\mathcal{P}_n F\| - \frac{2 \sum_{j=1}^{r} \|g_j\|}{\sqrt{N_{d,\,n}}} \right).$$

where $\mathcal{P}_n$ is a projection operator such that

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

Note: whenever F has heavy Legendre tail, N_2 needs to be wide

Main result - proof

$$
\begin{aligned}
\|F - N_2\|^2 &= \sum_{i=0}^{\infty} \left\| \alpha_i h_i - \sum_{j=1}^{r} \beta_i^j L_i^{\mathbf{v}_j} \otimes L_i^{\mathbf{v}'_j} \right\|^2 \geq \sum_{i=n}^{\infty} \left\| \alpha_i h_i - \sum_{j=1}^{r} \beta_i^j L_i^{\mathbf{v}_j} \otimes L_i^{\mathbf{v}'_j} \right\|^2 \\
&\geq \sum_{i=n}^{\infty} \alpha_i^2 - 2 \sum_{i=n}^{\infty} \sum_{j=1}^{r} \langle \alpha_i h_i,\, \beta_i^j L_i^{\mathbf{v}_j} \otimes L_i^{\mathbf{v}'_j} \rangle \\
&= \|\mathcal{P}_n F\|^2 - 2 \sum_{i=n}^{\infty} \sum_{j=1}^{r} \frac{\beta_i^j \alpha_i P_i(\langle \mathbf{v}_j, \mathbf{v}'_j \rangle)}{\sqrt{N_{d,i}}} \\
&\geq \|\mathcal{P}_n F\|^2 - 2 \sum_{j=1}^{r} \sum_{i=n}^{\infty} \frac{|\beta_i^j||\alpha_i|}{\sqrt{N_{d,n}}} \\
&\geq \|\mathcal{P}_n F\|^2 - 2 \sum_{j=1}^{r} \frac{1}{\sqrt{N_{d,n}}} \sqrt{\sum_{i=n}^{\infty} |\alpha_i|^2} \sqrt{\sum_{i=n}^{\infty} |\beta_i^j|^2} \\
&\geq \|\mathcal{P}_n F\|^2 - \frac{2 \|\mathcal{P}_n F\| \sum_{j=1}^{r} \|g_j\|}{\sqrt{N_{d,n}}}
\end{aligned}
$$

Example

We are looking for a function that can not be well approximated by a low degree polynomial. For example:

$$\sin(\pi\sqrt{d}mx)$$

Lemma

Let $s_{d,m}(x) = \sin(\pi\sqrt{d}mx)$. Then for any $d > d_0$ and for any degree k polynomial p we have

$$\int_{-1}^{1}(s_{d,m}(x) - p(x))^2 d\mu(x) \geq \frac{m-k}{4e\pi m}$$

Example

Proof of the Lemma

For large enough d and $|x| \leq \frac{1}{\sqrt{d}}$ we have

$$1 - x^2 \geq e^{-2x^2} \geq e^{-\frac{2}{d}} \quad \Longrightarrow \quad (1-x^2)^{\frac{d-3}{2}} \geq e^{-\frac{d-3}{d}} \geq e^{-1}$$

This, together with the fact that $\Gamma(\frac{d}{2})/\Gamma(\frac{d-1}{2}) \approx \sqrt{\frac{d}{2}}$, gives us

$$d\mu(x) = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}(1-x^2)^{\frac{d-3}{2}} dx \geq \frac{\sqrt{d}}{2e\pi} dx$$

and therefore,

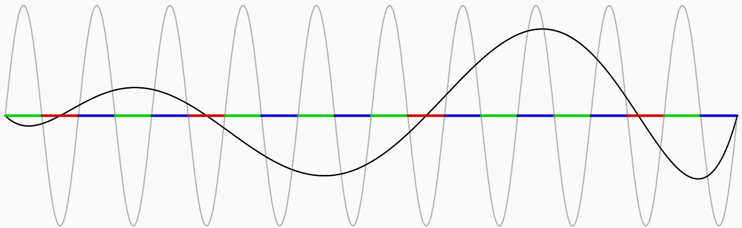
$$\int_{-1}^{1} f(x) d\mu_d(x) \geq \frac{\sqrt{d}}{2e\pi} \int_{-d^{-\frac{1}{2}}}^{d^{-\frac{1}{2}}} f(x) dx \geq \frac{1}{2e\pi} \int_{-1}^{1} f\left(\frac{t}{\sqrt{d}}\right) dt$$

Example

$$\int_{-1}^{1} f(x) d\mu_d(x) \geq \frac{1}{2e\pi} \int_{-1}^{1} f\left(\frac{t}{\sqrt{d}}\right) dt$$

Setting $f(x) = (\sin(\pi\sqrt{d}mx) - p(x))^2$ we obtain

$$\int_{-1}^{1} \left(\sin(\pi\sqrt{d}mx) - p\left(x\right)\right)^2 d\mu(x) \geq \frac{1}{2e\pi} \int_{-1}^{1} \left(\sin(\pi mx) - p\left(\frac{x}{\sqrt{d}}\right)\right)^2 dx$$



$$\frac{1}{2e\pi} \int_{-1}^{1} \left(\sin(\pi mx) - p\left(\frac{x}{\sqrt{d}}\right)\right)^2 dx \geq \frac{1}{2e\pi} \frac{m-k}{2m}$$

Example - conclusion

Setting

- $f(x) = \sin(\pi d^3 x)$
- $n = d^2$
- $B = 2^d$

and using our main theorem, we get

- $\mathcal{P}_n F \geq \frac{1}{5e\pi}$
- To get $\frac{1}{50e^2\pi^2}$–approximatopn of F, the width of NN should be

$$\frac{\sqrt{N_{d,d^2}}}{20e\pi 2^{2d}(1+\sqrt{4d}) + 2^{d+1}} = 2^{\Omega(d\log(d))}$$

Open questions

- Separation result for other classes of functions
- Depth-3 and depth-4 separation
- General depth-i and depth-$(i+1)$ separation
- ...

Thank you for attention!

Questions?